PRODUCTION AND OPERATIONS MANAGEMENT

PRODUCTION AND OPERATIONS MANAGEMENT

POMS

Vol. 28, No. 7, July 2019, pp. 1735–1756
ISSN 1059-1478 | EISSN 1937-5956 | 19 | 2807 | 1735

DOI 10.1111/poms.13012

Integrated Scheduling and Capacity Planning with Considerations for Patients' Length-of-Stays

Nan Liu*
Operations Management Department, Carroll School of Management, Boston College, 140 Commonwealth Avenue, Chestnut Hill, Massachusetts 02467, USA, nan.liu@bc.edu

Van-Anh Truong
Department of Industrial Engineering and Operations Research, Columbia University, 500 West 120th Street, New York City, New York 10027, USA, vatruong@ieor.columbia.edu

Xinshang Wang
Antai College of Economics and Management, Shanghai Jiao Tong University, 1954 Huashan Road, Xuhui District, Shanghai 200030, China, xs.wang@sjtu.edu.cn

Brett R. Anderson
Columbia University Irving Medical Center, 3959 Broadway, CH-2N, New York City, New York 10032, USA, bra2113@cumc.columbia.edu

Despite the fact that hospital care is often delivered in successive stages, current healthcare scheduling and capacity planning methods usually treat different hospital units in isolation. To address such a shortcoming, we introduce the first Markov decision process model for scheduling surgical patients on a daily basis, explicitly taking into account patient length-of-stay in hospital after surgeries and inpatient census. By way of a simple and yet innovative variable transformation, we reveal the hidden submodularity structure in our model. This transformation, in particular, allows us to show that the optimal number of patients to admit increases when the waitlist of surgical patients is longer, given the number of patients recovering downstream is fixed. We conduct extensive simulation experiments to study the applicability of our theoretical model in various settings. Our simulations based on real data demonstrate substantial values in making *integrated* scheduling decisions that simultaneously consider capacity usage at all locations in a hospital, especially when demand and system capacities are balanced or more elective patients present in the patient mix. The traditional scheduling policy, which is solely driven by operating room usage, however, can lead to significantly suboptimal use of downstream capacity and, as our numerical experiments show, may result in up to a three-fold increase in total expenses. In contrast, a scheduling policy based on downstream capacity usage often performs relatively close to the integrated scheduling policy, and therefore may serve as a simple, effective scheduling heuristic for hospital managers—especially when the downstream capacity is costly and less flexible.

Key words: surgical scheduling; hospital capacity; healthcare operations management; Markov decision process; analytics
History: Received: January 2018; Accepted: January 2019 by Michael Pinedo, after 2 revisions.

1. Introduction

In many health care settings, patient care is delivered in several successive stages at different locations. Among these stages, there are often *two* key stages marking the transition between high and low-intensity care. For instance, in maternity wards, women go through labor rooms to maternity units (before discharged home). For surgical patients, they receive surgeries in operating rooms (ORs) and then recover in inpatient units. In certain geriatric hospitals, patients pass from "acute-care" facilities to "long-stay" facilities. These two-stage systems share a common feature that patients are often scheduled to arrive at the upstream stage, flow spontaneously downstream, and then spend several days recovering there. If the downstream stage becomes fully occupied, access to the upstream and thus the whole health care system is likely to be *blocked*.

Indeed, upstream scheduling that fails to account for downstream patient *length-of-stay* (LOS) often leads to blocking, inefficient use of capacity, and consequently, high cost and reduced quality of care. Robb et al. (2004) report that "no bed" was the reason for

cancellation in general OR procedures for up to 62.5% of all canceled cases in a large university teaching hospital. Cochran and Bharti (2006) study an obstetrics hospital and find that when postpartum beds are full, patients are blocked in the upstream labor and delivery areas, preventing new admissions and leading to delays for scheduled inductions. In the critical care setting, a shortage of Intensive Care Unit (ICU) beds often forces surgeons to cancel or reschedule elective patients who might need ICU beds post surgery (Kim and Horowitz 2002).

In reality, the effective operation of these two-stage healthcare systems should depend on the "balanced" use of capacity at both stages to ensure a smooth patient flow. For instance, Griffin et al. (2012) show that blocking can be prevented by balancing capacities at different stages in an obstetric department, via the use of "swing" beds. In addition, making scheduling decisions for the upstream stage that account for patient LOS in the downstream stage may bring significant benefits. Using simulations, Robinson et al. (1968) test several heuristic scheduling rules and demonstrate an 8% to 17% cost reduction in a 100-bed hospital by taking into account patient LOS and downstream census when scheduling patient admissions.

Despite the potential value of integrated decision making that simultaneously considers capacity usage in multiple hospital locations, many hospital units still operate in isolation, without considering other connecting units. In particular, capacity planning for surgeries usually focuses on the use of ORs only, "with beds being considered as a secondary resource requirement that seldom constrains the overall capacity" (Bowers 2013); but this is often *not* true as discussed above.

With more hospitals adopting electronic health record systems, hospitals are acquiring the necessary information-technology support to coordinate capacity usage among different units and stages. However, the development of upstream scheduling methods accounting for downstream patient LOS has remained a challenge and an open research area. As Gupta (2007) points out, the random nature of LOS makes it difficult to formulate a tractable model. In such a model, the state space needs to be very large to capture the number of patients at each stage as well as their partially experienced LOS. The high dimensionality of the state space prevents tractable analyses.

In this paper, we study a surgical unit as a canonical example of a two-stage system in which patients receive surgeries at the upstream stage (ORs) and may spend multiple days at the downstream stage (inpatient wards). Both stages have finite capacity. There are two classes of patients who arrive randomly on each day. *Emergency* patients are assumed to enter the upstream stage on the day that they become emergent, whereas *elective* patients are initially added to a waitlist (the use of a surgical waitlist is common in single-payer health systems, such as UK and Canada; more on this below). From this waitlist, a certain number of elective patients are chosen to be admitted each day. On the day of admission, patients receive care first at the upstream stage, and then move to the downstream stage, where they stay multiple days before discharge. Idling and overtime costs occur at both stages for under and over consumption of available capacity at these stages, respectively. There is also a waiting cost for delaying surgeries on elective patients. The goal of the present research is to determine a dynamic admission policy for elective patients to minimize the total expected discounted cost in the system over a finite or infinite horizon. Such a policy would schedule patients to account for the linked usage of both stages of service, as well as patient downstream LOS.

Our model is an aggregate planning model similar to those studied in Gerchak et al. (1996), Ayvaz and Huh (2010) and Huh et al. (2013). These models are used in the first step of a typical two-step planning process. In this first step, the number of elective patients to be served on a given day is determined to balance the cost of capacity overuse with the cost of making patients wait and that of capacity under-utilization. In the second step, the sequence and timing of individual surgeries on a given day are determined to minimize intra-day wait time of patients and idle time of providers. The second step is usually *not* explicitly considered in an aggregate-planning model.

As noted above, the use of a surgical waitlist is common in countries like Canada and the UK. According to the Canadian Department of Health and Community Services (2016), "patients are selected from wait list and scheduled for surgery on the basis on urgency, best use of operating room time, and availability of hospital resources and staff." That is, providers have the right to choose patients from the waitlist, and patients will receive procedures on relatively short notice. This practice is different from that in the US, where patients are often provided a specific date for surgery in advance. Though our model cannot be directly applied to the advance scheduling context, it can be considered as an easier intermediate model that provides insights into the tradeoff between overtime capacity usage and patient waiting time, as well as insights on how to manage upstream capacity taking into account downstream resource usage. Indeed, a scheduling model that uses a waitlist is more tractable than an advance scheduling model; the latter is very challenging for analysis (even in a single-stage service setting) and has limited, provable structural insights (Liu et al. 2010, Patrick et al. 2008).

In certain settings, patients may receive care in more than two stages, e.g., from ORs to post-anesthesia-care (PACU) facilities and then to recovery units. In these settings, our two-stage model may be used to approximate the system, either by focusing on two main stages (usually the beginning stages), or by grouping the stages into two main groups. For instance, Bowers (2013) studies a Scotland-based center that primarily provides cardiothoracic surgical service. The majority of patients there are elective patients. Patients receive scheduled operating theater procedures, after which they transfer to ICUs, then to a High Dependency Unit (HDU), and finally into a conventional ward before they are discharged from the center. The HDU and ward capacities are high enough that they do not constrain the throughput of the center. Thus, the ORs and the ICUs are the main stages that must be considered in determining daily elective admission.

We have been working with the Congenital Heart Center at Columbia University Medical Center (CUMC). The care paths in this center also occur in two main stages. Take the pediatric population as an example. Patients $\leq$21 years of age receive cardiac surgery in one of the two pediatric ORs and recover in the pediatric cardiac intensive care unit (CICU). Many then move to a step-down unit, after the CICU, before discharged from the hospital. In this setting, hospital discharges to home are dependent on patients' clinical readiness, but discharges from the CICU to the step-down unit is often dependent on hospital operations unrelated to patients' health. Thus, OR is the upstream stage and the entire inpatient postoperative stay can be viewed as the downstream stage. Clearly, the linked capacities of the ORs and the downstream stage (i.e., combined step-down units and CICU) at this center impact the optimal number of elective patients to admit for surgeries on any given day.

Our contributions in this work can be summarized as follows. We formulate and analyze the first dynamic multi-day scheduling model that integrates information about capacity usage at two linked stages (see more discussions of the relevant literature in section 2). In particular, our model accounts for patient LOS and downstream census in scheduling decisions. We demonstrate that a formulation that uses the "natural" definition of decision variables does not generate clear structural results or insights. We are able to exploit a simple and yet innovative variable transformation to reveal a hidden submodularity structure in the formulation. This transformation allows us to show that more patients should be admitted when the waitlist is longer, given that the number of patients recovering downstream is fixed. When the waitlist is fixed but the number of patients recovering downstream increases, however, the optimal number of patients to admit may increase or decrease, depending on the current utilization level of downstream capacity.

We also study joint scheduling and capacity decisions in such a two-stage system. We find that adding more capacity to the upstream stage does *not* necessarily call for admitting more patients from the waitlist, and only an increase in the downstream capacity does. The more straightforward relationship between optimal scheduling decisions and downstream capacity suggests that (utilization of) downstream capacity may serve as a good guide for making scheduling decisions, which is indeed verified by our numerical experiments.

Via extensive numerical experiments based on data collected from CUMC, we show substantial values of *integrated* scheduling, that is, making scheduling decisions while taking into account capacity use of both service stages as opposed to only one stage. Integrated scheduling is particularly beneficial when the average demand and capacities in the system are relatively balanced, or when more elective patients present in the patient mix. We also show that our integrated scheduling model is robust and still generates scheduling policies that perform well in complex environments beyond the model setting.

Traditional scheduling policy used by practice is solely driven by OR capacity usage, and can lead to significantly suboptimal use of downstream capacity. In particular, we observe that such localized decision-rules may result in up to a three-fold increase in total expenses. In contrast, we find that a scheduling policy based on downstream capacity usage consistently outperforms the traditional scheduling policy, likely because that the overall downstream costs are of a larger magnitude than those of the upstream (and thus focusing on the downstream stage tends to perform better). Indeed, a scheduling policy based on the downstream stage often performs relatively close to integrated scheduling, especially when the downstream capacity is costly and less flexible. Thus, it may serve as an effective scheduling heuristic for hospital managers. This insight is likely to hold in a wide range of hospital environments because (downstream) hospital beds are known to represent the largest cost incurred by hospitals in general (Roberts et al. 1999).

The remainder of this study is organized as follows. Section 2 reviews the relevant literature. Section 3 describes our model and its natural formulation. Section 4 introduces the variable transformation for the model. Section 5 discusses the structural properties of our transformed formulation and its optimal

decisions. Section 6 treats capacities at both stages as new decision variables and investigates their relationship to the optimal cost and the optimal scheduling decisions. Section 7 presents our numerical study. Finally, section 8 summarizes our work and discusses potential future research directions. All technical proofs are shown in the Online Appendix.

2. Related Literature

Our work draws upon several streams of literature in the area of healthcare operations management: surgical scheduling, advance (appointment) scheduling and hospital patient flow modeling. From a mathematical modeling point of view, our work is related to the literature on admission control for tandem queues. We review each stream of the literature below.

Surgical scheduling has been an active research area for decades; see, e.g., Gupta (2007) and May et al. (2011), and Guerriero and Guido (2011) for in-depth reviews. This body of work has examined a variety of decision problems, including how to distribute the OR time among different surgeons, how many ORs to be open and when to open them, and how to sequence different procedures in a day. Among this literature, the most relevant work to ours includes Gerchak et al. (1996), Ayvaz and Huh (2010) and Huh et al. (2013). These three papers consider the allocation of elective patients to surgery days, using models similar to ours. However, they only model a single stage of service and do not take into account the usage of downstream capacity. In contrast, our work features a two-stage service system and develops integrated scheduling methods that explicitly consider capacity usage in both stages.

Our work is also related to the literature on advance (appointment) scheduling. This literature considers the problem of assigning patients to future days on their arrivals, with the objective of optimizing daily capacity utilization. Patrick et al. (2008) and Liu et al. (2010) are among the first to study such problems and to develop relevant dynamic optimization models. Some recent development has incorporated more sophisticated appointment requirements and details in the optimization model, including heterogeneous patient classes, treatment due dates and multiple appointments (Diamant et al. 2018, Saure et al. 2012). Our model fundamentally differs from this literature in the following aspects. First, our model considers allocation scheduling (i.e., how many patients to "pull" from the wailist for service in each day), but not advance scheduling (which directly assigns patients into future days on their arrivals). Second, we explicitly consider two linked stages of services, while the advance scheduling literature often considers a single stage of service.

Recently, scheduling models that explicitly consider patient flow through a hospital have received growing attention. The idea is to take a holistic system point of view, and to optimize operations by explicitly modeling patient utilization of multiple resources (instead of a single resource) in a hospital. Our research contributes to this emerging literature. One stream of this work has mainly focused on simulation models due to the complexity of the problem studied; see, e.g., Kim and Horowitz (2002) and White et al. (2011). A second stream develops static optimization models to determine the cyclic schedules of elective hospital admissions, taking into account their impact on the use of various hospital resources. Some recent examples include Beliën and Demeulemeester (2007), Adan et al. (2009), Bekker and Koeleman (2011), Price et al. (2011), Chow et al. (2011), Hulshof et al. (2013), Helm and Van Oyen (2014) and Gartner and Kolisch (2014). While these optimization models focus on different settings, they share a similar objective which is to smooth demand (resulting from upstream surgical scheduling) to downstream beds. They all assume that patients will be scheduled according to the cyclic schedules that have been set up, and then "passively" flow through the system. In contrast to these two streams of work, we develop an *analytic*, *dynamic* model of scheduling. We show that actively managed scheduling systems, where the decision to admit a patient accounts for resource usage in a downstream stage, can significantly improve operations by smoothing out potential imbalances between the stages.

Among the literature that develops scheduling models explicitly taking into account patient flow through a hospital, some consider dynamic decision making. In this literature, the most relevant study to ours is Helm et al. (2011) and deserves a detailed comparison. Helm et al. (2011) consider three patient arrival streams: emergency, scheduled elective and expedited. (Expedited patients are those with medical conditions that are less acute than emergency patients and whose admission can be delayed 1–3 days). All these patients are competing for a limited number of inpatient beds. At each decision epoch, the scheduler decides whether or not to cancel an elective surgery or call in an expedited patient who is currently waiting, in order to optimize the use of inpatient beds. Our work shares a similar feature with Helm et al. (2011) in that we also consider a two-stage hospital service system. However, our work differs in several important ways. First, our model aims to control the admission scheduling of elective patients, whereas they assume an exogenous stream of elective arrivals and are "not addressing elective admission

scheduling optimization." Second, our model explicitly captures the usage of capacity in both stages of service, while their model only considers the usage of inpatient beds. Therefore, though the expedited patients (who may be called-in to fill inpatient beds) in their model seem to play a similar role as elective patients (who may be admitted and then stay downstream) in our model, they are actually quite different: expedited patients are directly called-in to inpatient beds while elective patients in our model still consume resource in the first stage and then flow into the downstage stage. This distinction leads to different system dynamics. Third, due to the differences above, the decision making in our model depends on three quantities: the queue length of elective patients and the resources usage in both stages; the optimal scheduling policy thus has different structures and implications. In addition to Helm et al. (2011), Nunes et al. (2009) propose a dynamic decision model for elective patient admission control for distinct specialties; Samiedaluie et al. (2017) consider patient admission policy in a neurology ward, where there are two stages of service, the emergency department and the neurology ward. However, due to the complexity of their models, the last two studies do not identify the structure of an optimal scheduling policy.

Finally, from the mathematical modeling perspective, our work is related to the literature on admission control for tandem queues. See Zhang and Ayhan (2013) for a brief review. These studies are concerned with whether to accept or reject an arriving customer. One key feature of these models is the "hard" capacity constraint imposed on the buffers, such that customers will balk (i.e., leave the system) whenever buffers are full. That is, customers may leave before getting to the last stage of service. While these models are perfectly applicable for communication networks such as the Internet, they may not be appropriate to some of the healthcare contexts we consider because capacity in hospital sometimes can be flexible rather than fixed. The special features of healthcare delivery lead us to develop the new model with "soft" capacity constraints presented in this work.

3. Model

In this section, we describe our modeling framework. By convention, we will use Greek letters to denote random variables, upper-case letters to denote constants, and lower-case letters to denote variables. We will consider all subscripted or superscripted quantities as vectors when we omit their subscripts or superscripts, respectively.

Consider a planning horizon of T days, numbered $t = 1, 2, \ldots, T$. We allow $T = \infty$. Demand for elective and emergency surgeries that arise over each day t are non-negative integer-valued random variables denoted by δ_t and ϵ_t, respectively. We assume that δ_t and ϵ_t are independent and identically distributed (i.i.d.) for $t = 1, 2, \ldots, T$, and bounded. Emergency surgeries must be performed on the same day in which they arise, whereas elective surgeries can be waitlisted and performed in the future. Each elective case that is waitlisted incurs a waiting cost of W per day. The waiting cost captures the inconvenience and loss of goodwill in patients due to waiting. It can also capture loss in productivity to the patient and to society that is caused by delays in treatment. This model of waiting costs follows Gerchak et al. (1996) and Ayvaz and Huh (2010).

A patient undergoing surgery always proceeds through two main stages in the hospital. Stage 0, called the *entry stage* or *upstream stage*, takes place on the day when the patient is admitted into the hospital. In this stage, surgery is performed. The patient stays in the entry stage for no more than a fraction of a day. After receiving surgery in the entry stage, the patient will move to stage 1, called the *downstream stage*, for recovery and observation. The downstream stage may represent an intensive care unit (ICU), a step-down unit, or other inpatient ward. In addition to surgical patients, there may be other patients who need to be directly admitted to the downstream stage without going through surgeries, e.g., those directly admitted from emergency department. We let θ_t denote the number of this external stream of patients in period t, and assume that θ_t's are i.i.d. non-negative integer-valued and bounded random variables for $t = 1, 2, \ldots, T$. All these admitted patients stay in the downstream stage for a random number of days before they are finally discharged. Specifically, a random fraction $1 - \xi_t$, $\xi_t \in (0, 1)$, of patients at stage 1 exit the system at the end of period t. We assume that ξ_t's are i.i.d. for $t = 1, 2, \ldots, T$. This is in spirit similar to assuming that patient LOS in stage 1 is exponentially (geometrically) distributed. Litvak et al. (2008) have shown that patient LOS in ICUs can be modeled as an exponential (geometric) random variable. Bowers (2013) has also noted that the exponential (geometric) distribution provides an "approximate" fit to the LOS data at the cardiothoracic center he studies. Assuming a geometrically distributed LOS leads to a Markovian system which is amenable for analysis. In our numerical study later, we will test our model in general settings where patient LOS does *not* follow the geometric distribution.

We assume that there is a single resource consumed by patients in each stage $i \in \{0, 1\}$. We call this resource *stage-i capacity* and denote it by C_i. For example, capacity can be measured in surgeon time in the

entry stage or in number of beds in the downstream stage. Each patient consumes a random amount v^0 of capacity in stage 0, and v^1 of capacity in each day that she remains in stage 1. We assume that v^i is i.i.d. over time and over patients for $i \in \{0, 1\}$.

Since our model is an aggregate planning model that determines the total number of elective patients to be treated each day, we estimate the total amount of stage-0 and stage-1 capacity used by any k patients on any given day by the convolutions $S^0(k)$ and $S^1(k)$ of k i.i.d. random variables, distributed as v^0 and v^1, respectively. Conceptually, our model assumes that one procedure/treatment begins when the previous one ends. We do not explicitly model intra-day wait time by patients, idle time by doctors or preparation time between procedures that depend on the sequencing and timing of procedures within a day. As discussed in Gerchak et al. (1996) and Choi and Wilhelm (2014), such approximation of the workload within a day is reasonable for aggregate-planning models. Readers who are interested in intra-day dynamics of scheduling may refer to papers such as Denton and Gupta (2003) and Zacharias and Pinedo (2014).

Another important (implicit) assumption made in our model is that patient delays in the waitlist (before receiving surgeries) and patient capacity usage in both stages are independent. Some recent empirical studies find that patient resource use may be endogenized to the system state, e.g., bed occupancy level; see, e.g., Kc and Terwiesch (2009, 2012) and Chan et al. (2016). However, to keep our analysis tractable, we assume that patient resource use is exogenous, and leave it for future research to consider state-dependent decision-making models with endogenous patient LOS.

On any given day, if more capacity is required than is available at stage i, then surge capacity will be used, incurring an *overtime cost* of $O_i \geq 0$ per unit. Conversely, if less capacity is required than is available at stage i, an *idling cost* of $L_i \geq 0$ is incurred per unit. Specifically, the expected overtime and idling costs at stage i, given that k patients are served, is $O_i\mathbf{E}[S^i(k) - C_i]^+ + L_i\mathbf{E}[S^i(k) - C_i]^-$ where $(\cdot)^+ = \max(\cdot, 0)$ and $(\cdot)^- = -\min(\cdot, 0)$. We note that our capacity utilization model either incurs overtime cost or idling cost, but not both. Our cost structure is in line with those in the earlier literature, which charges costs for deviations from target capacity levels in order to stabilize average hospital resource utilization to desired levels (Adan and Vissers 2002, Nunes et al. 2009).

We acknowledge that sometimes when the downstream stage provides more intensive or specialized care (e.g., ICU), surge capacity might not be available, meaning that the constraint on downstream capacity is *strict*. While our model cannot enforce an absolutely strict constraint on downstream capacity, it can provide a good approximation by setting the downstream overtime cost sufficiently high (see more discussion in section 7.2.1). To incorporate a hard downstream capacity limit requires a different model, and we leave it for future research.

To avoid unnecessary complications imposed by the integral requirement, we will allow the number of patients admitted to be non-integral. Accordingly, we extend the stage i-cost above to be defined on real-valued k by piecewise-linear extension. We note that a fully discrete model would be a more natural one to assume, but using a discrete state space makes problems "intractable to analyze" (Kang et al. 2016). The continuous approximations have been widely used because they are found to be suitable in the service applications and to provide technical tractability; see, e.g., Talluri and Van Ryzin (2005), Chao et al. (2009), Kang et al. (2016). We follow this wisdom from the earlier literature.

To sum up, the events in each day occur in the following sequence.

1. At the beginning of day t, there are w_t elective patients on the waitlist. There is no patient upstream (i.e., at stage 0) because all patients admitted on day $t-1$ have completed their service at stage 0 on the same day. There are n_t patients downstream (i.e., at stage 1). Waiting costs are incurred for each of the w_t patients on the waitlist. We allow w_t and n_t to be non-negative real numbers.
2. A random number δ_t of new elective surgery requests arises, bringing the total number of patients in the waitlist to $\bar{w}_t = w_t + \delta_t$, and the total number of patients in the system to $w_t + \delta_t + n_t$, which include patients waiting for surgery as well as those in the downstream stage.
3. The scheduling manager decides, out of the $w_t + \delta_t$ outstanding elective cases, the number q_t of elective surgeries to fulfill on day t. Immediately after the decision, the number of patients at stage 0 increases to q_t. Again, we allow q_t to be a non-negative real number.
4. An additional random number ϵ_t emergency patients arrive and are served at stage 0. Idling or overtime costs are incurred at stage 0 for the service of $q_t + \epsilon_t$ patients.
5. Each patient in stage 0 moves to stage 1. A random number θ_t patients arrive and are directly admitted to stage 1. Idling or overtime costs are incurred at stage 1.
6. A random fraction $1 - \xi_t$, $\xi_t \in (0, 1)$, of patients at stage 1 exit the system.

The objective of the problem is to find a scheduling policy that minimizes the total expected discounted costs of the system over the planning horizon, assuming a discount factor of $\gamma \in (0, 1)$.

3.1. Dynamic-Programming Formulation

The decision problem introduced above can be formulated as a Markov decision process (MDP). The system has the following tradeoff. If it schedules too many elective surgeries in a day, the waiting cost is reduced but overtime cost might be high in both stages. In contrast, if it schedules too few elective surgeries, it risks incurring high waiting costs for elective patients and high idling costs in both stages. Very importantly, the scheduling decision needs to consider the use of capacity system-wide, as the decision that optimizes the cost in one stage may not be optimal for the other stage, nor for the system as a whole.

Recall that the decision to make in each day is the number of elective patients q_t to serve. The state of the system just before decision q_t is made is represented by a triplet (w_t, n_t, δ_t), where $w_t + \delta_t = \bar{w}_t$ and n_t represent the total number of patients on the waitlist and downstream, respectively. The decision q_t is constrained by $0 \le q_t \le \bar{w}_t$, since the number to be scheduled cannot exceed the number currently on the waitlist.

The system evolves as follows:

$$w_{t+1} = w_t + \delta_t - q_t, \tag{1}$$

$$n_{t+1} = \xi_t(n_t + q_t + \epsilon_t + \theta_t). \tag{2}$$

To see the second equation, note that a fraction $1 - \xi_t$ of the $n_t + q_t + \epsilon_t + \theta_t$ patients who stay in stage 1 on day t exit the system.

The single-day cost function can be written as

$$\begin{aligned}\tilde{F}(w_t, n_t, \delta_t, q_t) = {} & Ww_t + O_0\mathbf{E}[S^0(q_t + \epsilon_t) - C_0]^+ \\ & + L_0\mathbf{E}[S^0(q_t + \epsilon_t) - C_0]^- \\ & + O_1\mathbf{E}[S^1(n_t + q_t + \epsilon_t + \theta_t) - C_1]^+ \\ & + L_1\mathbf{E}[S^1(n_t + q_t + \epsilon_t + \theta_t) - C_1]^-.\end{aligned} \tag{3}$$

Above, the first term captures the waiting cost for elective patients who are waitlisted on day t; the second and third terms evaluate the overtime and idling costs for stage 0 on day t, respectively; and the last two terms compute these costs for stage 1 on day t, respectively. Note that our method of charging the waiting cost is equivalent to charging a unit waiting cost for each customer in the waitlist for the next period, since $w_t + \delta_t - q_t$ is just another way to write w_{t+1}.

Let $\tilde{V}_t(w_t, n_t, \delta_t)$ denote the optimal total discounted cost incurred from days t to T when the state just before the decision is made on day t is given by (w_t, n_t, δ_t). The Bellman equation can be written as follows:

$$\begin{aligned}\tilde{V}_t(w_t, n_t, \delta_t) &= \min_{0 \le q_t \le w_t + \delta_t} \tilde{G}_t(w_t, n_t, \delta_t, q_t), \text{ where} \\ \tilde{G}_t(w_t, n_t, \delta_t, q_t) &= \tilde{F}(w_t, n_t, \delta_t, q_t) \\ &\quad + \gamma\mathbf{E}\big[\tilde{V}_{t+1}(w_{t+1}, n_{t+1}, \delta_{t+1}) | \\ &\quad\quad (w_t, n_t, q_t, \delta_t)\big], \\ &= \tilde{F}(w_t, n_t, \delta_t, q_t) \\ &\quad + \gamma\mathbf{E}\big[\tilde{V}_{t+1}(w_t + \delta_t - q_t, \\ &\quad\quad \xi_t(n_t + q_t + \epsilon_t + \theta_t), \delta_{t+1})\big].\end{aligned} \tag{4}$$

For convenience, we take the termination function when $T < \infty$ to be $\tilde{V}_{T+1}(\cdot, \cdot, \cdot) = 0$, but any linear function would be acceptable. We suppress the capacity vector C except when it is explicitly required by the discussion. In the infinite-horizon case, since the demands are bounded and all costs are non-negative, the time index can be dropped from the optimality equation (4).

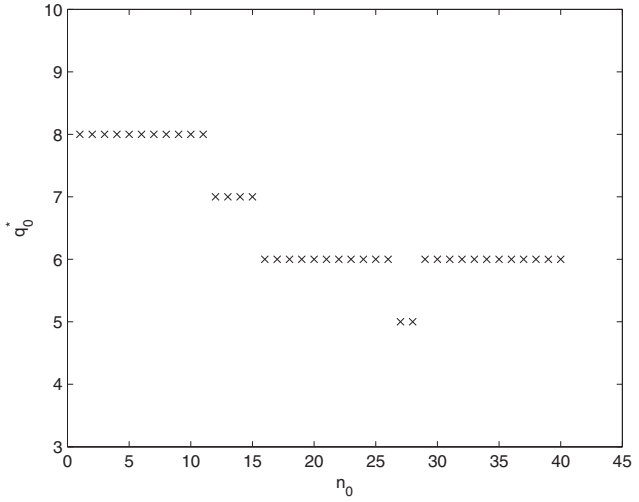
We call formulation (4) the *natural formulation* of the problem because it uses the variables that one would usually use to formulate such a problem. These variables are also used in early classic work on single-stage surgical scheduling (Gerchak et al. 1996). In the next section, we shall find it necessary to transform this natural formulation into one that is more analytically tractable.

4. Transformation of Variables

The natural formulation of the previous section turns out to be rather difficult to analyze. It does not yield a clear and intuitive relationship between the system state and decision variables, nor does it generate very useful managerial insights. For example, recall that the state variable n_t tracks the total number of patients downstream, and the decision variable q_t controls the daily rate at which regular patients are admitted. As the total number of patients downstream n_t increases, intuition seems to suggest that fewer patients should be admitted to the upstream, that is, q_t should be smaller, to avoid overtime downstream. However, as it turns out, q_t might increase or decrease when n_t grows, depending on the relative value of n_t compared to the downstream capacity.

A numerical example is shown in Figure 1, in which the optimal decision q_t^* initially decreases in n_t but then increases, when the length of the waitlist w_t is fixed. This suggests hospital managers that when inpatient capacity becomes constrained, few surgeries should be scheduled due to downstream blocking. However, when inpatient unit becomes overly congested, the focus should return to optimizing the use of (very) expensive OR capacity by increasing surgical admissions. In practice, ORs are frequently operated beyond

Figure 1 For a Fixed $w_0 = 8$, the Optimal Decision q_0^* is Not Monotone in the State Variable n_0. $C_0 = 7$, $C_1 = 20$, $W = 2$, $O_0 = 6$, $L_0 = 6$, $O_1 = 5$, $L_1 = 5$, $\mathbf{E}[v^0] = 1$, $\mathbf{E}[v^1] = 1$, $\gamma = 0.9$, $\mathbf{E}[\delta_t] = 6$, $\mathbf{E}[\epsilon_t] = 0.5$, $\mathbf{E}[\theta_t] = 0$, $T = 50$, $\mathbf{E}[\xi_t] = 0.7$ and ξ_t is Uniformly Distributed Over [0.6, 0.8]



what the inpatient capacity allows. This is usually driven by the desire to use OR capacity to the greatest extent possible, and is often due to lack of intra department collaboration (Jweinat et al. 2013, Kosnik 2006). A sole focus on OR utilization is surely not optimal when considering the use of inpatient capacity together; but, as this example shows, it may still be a good choice when the inpatient unit is already (very) crowded.

To "quantitatively" explain the observations above, note that when the number n_t of patients downstream is relatively small compared to the downstream capacity, it is crucial to reduce the idling cost downstream as much as possible, by pulling patients from the waitlist. Thus, we see that when n_t increases from a relatively small number (e.g., 10), fewer and fewer patients are admitted; this strategy maintains an optimal level of downstream occupancy. A decreased number of patients admitted from the waitlist may incur more idling costs upstream, but when n_t is relatively small compared to the downstream capacity, ensuring a good use of downstream capacity dominates the admission decision.

When n_t is sufficiently large (e.g., 27), the overtime cost downstream is almost the same for any newly admitted patient because to serve each one of them will most likely require the use of surge capacity. In this case, balancing idling cost and overtime cost upstream becomes more relevant, and thus more patients should be admitted to the upstream stage when n_t increases from a relatively high level.

The example above shows that the relationship among the original model variables does *not* provide a clear direction on how to adjust decisions as the system state changes. Next, we perform a simple transformation of variables that places them in approximately the same "space," thereby helping to reveal the relationship among them. Let a_t denote the total number of patients in the system at the beginning of day t, including those in the waitlist and those in stage 1. That is, $a_t = w_t + n_t$. Let m_t denote the number of patients in *both* stages (excluding those on the waitlist) immediately after the decision q_t. In other words, $m_t = n_t + q_t$. We reformulate problem (4) with variables (a_t, m_t) replacing (w_t, q_t). See Figure 2 for our model schematic with different variables.

Observe that with the above transformation, the decision variable becomes the total number m_t of patients to be in the hospital[1] by the end of period t. It is more compatible with the state variables a_t and n_t than the original use of q_t as decision variable in the sense that, rather than specifying daily admission counts, it also accounts for the total number of patients in the system at and beyond a point, in this case at stage 0 and beyond. In comparison, a_t accounts for the total number of patients on the waitlist and beyond, and n_t accounts for the number of patients at stage 1. In short, a_t, m_t, and n_t correspond to the size of three nested sets of patients.

From day t to $t + 1$, the system evolves as follows,

$$\begin{aligned} a_{t+1} &= a_t + \delta_t + \epsilon_t + \theta_t - (1 - \xi_t)(m_t + \epsilon_t + \theta_t) \\ &= a_t + \delta_t - m_t + \xi_t(m_t + \epsilon_t + \theta_t), \end{aligned} \tag{5}$$

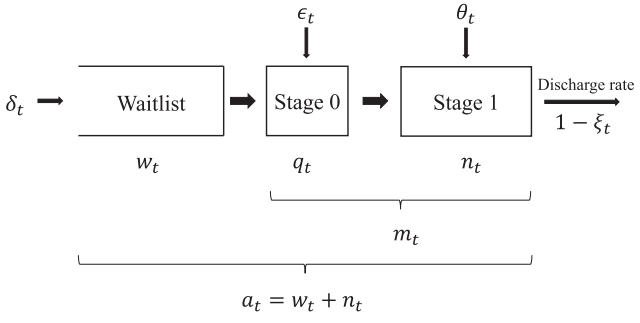
$$n_{t+1} = \xi_t(m_t + \epsilon_t + \theta_t). \tag{6}$$

The single-day cost function can be written as

$$\begin{aligned} F(m_t, a_t, n_t, \delta_t) = {} & W(a_t - n_t) \\ & + O_0\mathbf{E}[S^0(m_t - n_t + \epsilon_t) - C_0]^+ \\ & + L_0\mathbf{E}[S^0(m_t - n_t + \epsilon_t) - C_0]^- \\ & + O_1\mathbf{E}[S^1(m_t + \epsilon_t + \theta_t) - C_1]^+ \\ & + L_1\mathbf{E}[S^1(m_t + \epsilon_t + \theta_t) - C_1]^-. \end{aligned} \tag{7}$$

Let $V_t(a_t, n_t, \delta_t)$ denote the optimal total discounted cost incurred from day t to T when the system state just before the decision m_t is made on day t is (a_t, n_t, δ_t).

Figure 2 Model Schematic



The Bellman equation can be written as follows:

$$V_t(a_t, n_t, \delta_t) = \min_{n_t \le m_t \le a_t + \delta_t} G_t(m_t, a_t, n_t, \delta_t), \text{ where} \quad (8)$$

$$\begin{aligned} G_t(m_t, a_t, n_t, \delta_t) &= F(m_t, a_t, n_t, \delta_t) \\ &\quad + \gamma \mathbf{E}[V_{t+1}(a_{t+1}, n_{t+1}, \delta_{t+1}) | (a_t, n_t, m_t, \delta_t)], \\ &= F(m_t, a_t, n_t, \delta_t) \\ &\quad + \gamma \mathbf{E}[V_{t+1}(a_t + \delta_t - m_t + \xi_t(m_t + \epsilon_t + \theta_t), \\ &\quad \xi_t(m_t + \epsilon_t + \theta_t), \delta_{t+1})], \end{aligned}$$

and the termination function is given by $V_{T+1}(\cdot, \cdot, \cdot) = 0$ or any linear function. Note that the feasible region for m_t is $[n_t, a_t + \delta_t]$, ranging from the total number n_t of patients downstream to the total number $a_t + \delta_t$ patients in the system. Again, in the infinite-horizon case, since the demands are bounded and all costs are non-negative, the time index can be dropped from the optimality equation.

We call (8) the *transformed formulation*. Via the variable transformation, the decision variables in our model (i.e., a_t, m_t and n_t) are the number of three nested sets of patients. While they seem to play a similar role of inventory positions in a multi-echelon inventory system, it is not clear that one can draw a direct parallel between our model and an inventory system. An inventory position accounts for inventory that is present and in transit. All such units are the same. In contrast, patients upstream are different from those waiting downstream because the latter group has been partially treated. In the inventory system, the decision maker decides on the inventory position. In our system, the decision maker decides on the internal movements of patients in between the two stages. The costs in the inventory position are holding costs for all units and shortage costs on units ordered late. The costs in our system are waiting costs for patients upstream and overtime costs in both stages, which are not the same.

In the following section, we show that the transformed formulation yields a well-structured relationship between the optimal decision and the state variables. In particular, the transformed formulation exhibits submodularity whereas the natural formulation does not.

5. Structure of Optimal Scheduling Policies

We now investigate the transformed formulation (8). We derive the structural properties that will shed light on the characteristics of the optimal policies, thus providing decision makers with helpful guidance on these policies. We first note that the convexity of the formulation follows from the convexity of the single-period cost function and linear state transitions over time.

PROPOSITION 1. *For every* $t = 1, 2, \ldots, T$, $F(\cdot, \cdot, \cdot, \cdot)$, $G_t(\cdot, \cdot, \cdot, \cdot)$, *and* $V_t(\cdot, \cdot, \cdot)$ *are jointly convex in their arguments, respectively.*

Following Topkis (1998), we say that a function $g: \mathbb{R}^n \to \mathbb{R}$ is submodular if

$$g(x) + g(y) \ge g(x \wedge y) + g(x \vee y),$$

for all $x, y \in \mathbb{R}^n$, where $x \vee y$ denotes the component-wise maximum and $x \wedge y$ the component-wise minimum of x and y. We can prove that the transformed formulation is submodular using the submodularity of the one-period cost function, the joint convexity of the value function shown above, the linear state transitions, and the lattice structure of the feasible region.

THEOREM 1. *For every t, the following properties hold:*

1. $F(m_t, a_t, n_t, \delta_t)$ *is submodular in* (m_t, a_t, n_t);
2. $G_t(m_t, a_t, n_t, \delta_t)$ *is submodular in* (m_t, a_t, n_t); *and*
3. $V_t(a_t, n_t, \delta_t)$ *is submodular in* (a_t, n_t).

The submodularity results established above are crucial in establishing the monotonicity of the optimal decisions in the state variables. These monotonicity properties provide decision makers with easy directions for policy adjustment.

As the optimal decision might not be unique, we define the minimum and maximum optimal decision m_t in day t to be, respectively,

$$m_t^{\min}(a_t, n_t, \delta_t) = \min \arg \min_{n_t \le m_t \le a_t + \delta_t} G_t(m_t, a_t, n_t, \delta_t), \text{ and} \quad (9)$$

$$m_t^{\max}(a_t, n_t, \delta_t) = \max \arg \min_{n_t \le m_t \le a_t + \delta_t} G_t(m_t, a_t, n_t, \delta_t). \quad (10)$$

Then we have the following result. (In this study, we use the terms "increasing" and "decreasing" to mean "non-decreasing" and "non-increasing," respectively, unless otherwise specified.)

COROLLARY 1. *For every period t and demand instance* δ_t, *the maximum and minimum optimal numbers of elective patients in stage* 0 *and beyond, namely* $m_t^{\max}(a_t, n_t, \delta_t)$ *and* $m_t^{\min}(a_t, n_t, \delta_t)$, *respectively, are both increasing in the system state* (a_t, n_t).

The monotonicity of the optimal decision in the system state is quite intuitive after the variable transformation. As noted above, a_t, m_t, and n_t account for the total number of elective patients on the waitlist and beyond, at stage 0 and beyond, and at stage 1, respectively. They correspond to the sizes of three nested sets of patients (see Figure 2). These sets have a correspondence in size under an optimal policy.

Put into managerial language, Corollary 1 suggests that, everything else being equal, the inpatient census—after the optimal surgical admission decision is made—will be higher, if the inpatient unit is more congested to start with or the whole system contains more patients. This interpretation provides hospital managers with two general, intuitive pointers on when to expect (extended) congestion in the downstream unit: observing an already crowded downstream unit or having a heavy workload in the system as a whole. Next, we elaborate how hospital managers may control surgical admissions, that is, q_t, to manage congestion in the downstream unit as well as the whole system.

With the size of the smallest set containing n_t recovering patients being fixed, as the largest set containing a_t patients in the whole system increases in size, the middle set containing m_t admitted and recovering patients also enlarges. This suggests that, *when* n_t *is fixed* one should admit more patients when the waitlist is longer (i.e., $q_t = m_t - n_t$ does not decrease when $w_t = a_t - n_t$ becomes larger)—this monotonicity result is independent of the upstream capacity. Simply put, hospital managers should consider increasing surgical admissions when the surgical waitlist grows. This implication is consistent with the classic result in the previous literature on single-stage surgical scheduling (Gerchak et al. 1996), but the important condition here when we account for the use of the downstream stage is that the downstream inpatient census n_t is fixed.

When n_t increases and w_t is fixed (or w_t becomes larger), however, we only know that a_t is larger and thus the optimal m_t increases—we do *not* have a definite answer on whether q_t, the number to admit for surgeries from the waitlist, becomes larger or not. Figure 1 shows a concrete example, in which q_t first decreases and then increases when n_t increases. Recall the driver of admission decision making there is to ensure a good tradeoff between the use of OR and inpatient capacity: one should decrease surgical admissions when the inpatient unit becomes more congested to avoid overburdening the inpatient unit; but when the inpatient unit has already become overcrowded, one should admit more surgeries to ensure a good use of the (very) expensive OR capacity.

When a_t is fixed and n_t increases (i.e., when w_t decreases at the same rate as n_t increases), the middle set containing m_t admitted and recovering patients increases. To explain, consider two systems with the same, total number of patients (who either wait for surgeries or recover in the hospital). One system has a more congested inpatient unit (and thus a shorter waitlist); otherwise the two systems are identical. Then, the system with a more congested inpatient unit, after the optimal surgical admission decision is made (in both systems), remains more congested in its inpatient unit.[2] This suggests that congestion in a inpatient unit, once formed, may not be dissolved by arbitrarily decreasing surgical admissions; hospital managers need to have a contingency plan in place (e.g., use flexible capacity such as swing beds and get help from providers on-call).

In sum, the relationships among the original model variables (w_t, n_t and q_t) under an optimal policy are in general non-monotonic, except for one case in which q_t increases when w_t increases and n_t is fixed. The transformed variables (a_t, n_t and m_t) allow us to uncover structural results, which may (partially) explain the non-monotonicity observed in the original variables. More importantly, they provide a better lens to glean managerial insights as discussed above.

Finally, while our model requires patients to receive surgeries on relatively short notice by determining the number of surgical patients to admit from the waitlist on each day, it may be applied to advance scheduling contexts in the following heuristic way, in the spirit of Gerchak et al. (1996). When elective patients arrive, their surgical dates will be planned for some future day, but may be moved earlier or later by a few days. Our model can be used to calculate the optimal number of patients that will be served on the current day. If the number of scheduled patients on the current day exceeds the optimal number, some of them may be postponed; otherwise, some patients who are scheduled later may be asked to receive surgeries earlier. To entirely adopt the optimal solution suggested by our model requires that patients are fully flexible in terms of their surgical dates, and this is likely to be impossible in practice. However, if the provider can establish good communication with patients and even if some patients are flexible, our model may still provide useful information to adjust the surgical schedule towards a more efficient one.

6. Relationship to Capacity

So far we have assumed that the capacity is fixed at both stages. These capacities in our model are *not* the "physical" capacity of a hospital, changes to which would require building new space and making capital investments. The capacities in our model represent

the *ideal* number of patients that each stage aims to serve in a day given the current physical capacity. At the upstream stage, the capacity C_0 is determined by the regular number of health care providers and resources put to work in a day; at the downstream, the capacity C_1 is determined by the number staffed beds. These capacities in our model can be adjusted by tactical decisions (e.g., by the use of overtime and flexible beds), and therefore are interesting subjects of research.

In this section, we treat the daily capacity C_i's as additional decision variables and study how the optimal cost function V_t and the optimal scheduling decisions $m_t^{\min}$ and $m_t^{\max}$ change with respect to changes in C_i's. This analysis provides useful information for hospital managers to make joint decisions on scheduling and capacity planning.

6.1. Impact of Capacity Changes on the Optimal Cost

Including the capacity vector C into analysis makes it difficult to investigate the convexity and other structural properties of V_t. To see why, consider the fourth term in the single-day cost function (7) which has the following functional form $O_1\mathbf{E}[S^1(m) - C_1]^+$. It is not clear how to define the joint convexity of this term in $\{(m, C_1) : m \in \mathbb{Z}_+, C_1 \in \mathbb{R}_+\}$, where $\mathbb{Z}_+$ and $\mathbb{R}_+$ represent the set of non-negative integers and that of non-negative real numbers, respectively. For technical tractability, we make the following simplification to the model. Instead of assuming the capacity used by each patient is i.i.d., we assume that the total amount of stage-i capacity used by any k patients on any given day is given by $v^i k$, where v^i is a non-negative random variable representing the average capacity used by a patient for stage i, $i = 0, 1$. That is, each of these k patients uses the same amount of capacity, and this capacity is randomly distributed as v^i for stage i. We rely on this simplification to derive structural insights, but we will conduct numerical analysis to verify if these insights would continue to hold in our original model setting, which assumes i.i.d. demand usage by each patients.

For the cost of capacity, we assume that at any stage each additional unit of capacity incurs a daily cost. This daily cost can be thought of as daily depreciation of the infrastructure investment plus the staffing cost associated with the capacity. Once the capacity is determined, it cannot be changed and thus its cost can be viewed as a sunk cost. Without loss of generality, we can assume zero capacity cost because adding a linear capacity cost to our formulation will not change the structural insights on how the capacity at the two stages affects the optimal policies.

With these modifications, the single-day cost function becomes

$$\begin{aligned} F(m_t,a_t,n_t,\delta_t,C) = W(a_t-n_t) &+ O_0\mathbf{E}[v^0(m_t-n_t+\epsilon_t)-C_0]^+ \\ &+ L_0\mathbf{E}[v^0(m_t-n_t+\epsilon_t)-C_0]^- \\ &+ O_1\mathbf{E}[v^1(m_t+\epsilon_t+\theta_t)-C_1]^+ \\ &+ L_1\mathbf{E}[v^1(m_t+\epsilon_t+\theta_t)-C_1]^-, \end{aligned} \tag{11}$$

and the value function V_t remains the same as defined in Equation (8) except that C_i's are now included as new decision variables. Then, we can show the following results similar to Proposition 1.

THEOREM 2. *For each t, $F(C, m_t, a_t, n_t)$, $G_t(C, m_t, a_t, n_t)$, and $V_t(C, a_t, n_t)$ are jointly convex in their arguments, respectively.*

The theorem above suggests a diminishing return as the capacity upstream or downstream increases. This trend has been described previously in a simulation study by Bowers (2013). With greater investments in capacity, we eventually experience lower marginal returns when patient demand remains the same. These convexity results are derived by relaxing the assumption that capacity use of different patients are i.i.d. A natural question is whether such convexity with respect to capacity C would still hold with the i.i.d. assumption. Our intuition suggests the answer to be yes, and our numerical experiments below indeed confirm our intuition. In particular, Figures 3 and 4 illustrate the joint convexity of $V_0(C, a_0, n_0)$ as a

Figure 3 **Total Expected Cost $V_0(C, a_0, n_0)$ as a Convex Function of (C_0, C_1), in the Scenario with Higher Upstream Costs. $O_0 = 10$, $L_0 = 10$, $O_1 = 0.5$, $L_1 = 0.2$, $a_0 = 18$, $n_0 = 13$, $W = 1$, $\mathbf{E}[v^0] = 1$, $\mathbf{E}[v^1] = 1$, $\gamma = 0.8$, $\mathbf{E}[\delta_t] = 8$, $\mathbf{E}[\epsilon_t] = 1.5$, $\mathbf{E}[\theta_t] = 0$, $T = 90$, $\mathbf{E}[\xi_t] = 0.85$ and ξ_t is Uniformly Distributed Over [0.7, 1] [Color figure can be viewed at wileyonlinelibrary.com]**

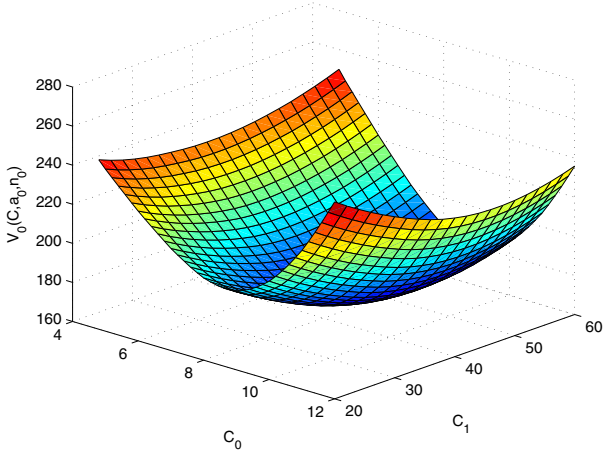


Figure 4 **Total Expected Cost $V_0(C, a_0, n_0)$ as a Convex Function of (C_0, C_1), in the Scenario with Higher Downstream Costs. $O_0 = 1$, $L_0 = 1$, $O_1 = 3$, $L_1 = 1.2$, $a_0 = 18$, $n_0 = 13$, $W = 1$, $\mathbf{E}[v^0] = 1$, $\mathbf{E}[v^1] = 1$, $\gamma = 0.8$, $\mathbf{E}[\delta_t] = 8$, $\mathbf{E}[\epsilon_t] = 1.5$, $\mathbf{E}[\theta_t] = 0$, $T = 90$, $\mathbf{E}[\xi_t] = 0.85$ and ξ_t is Uniformly Distributed Over [0.7, 1] [Color figure can be viewed at wile yonlinelibrary.com]**

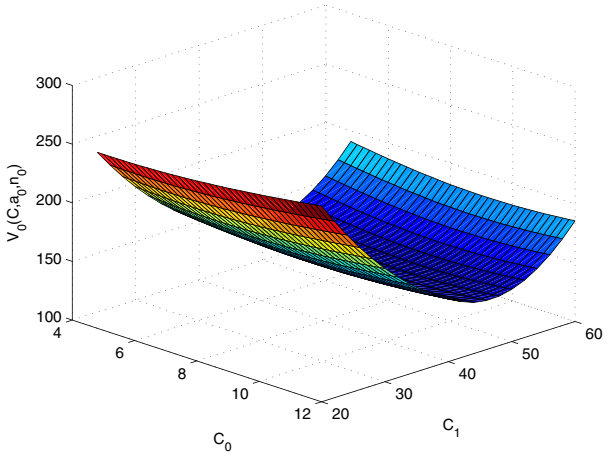


Figure 5 **Total Expected Cost $V_0(C, a_0, n_0)$ as a Convex Function of C_0. $a_0 = 18$, $n_0 = 13$, $W = 1$, $C_1 = 30$, $\mathbf{E}[v^0] = 1$, $\mathbf{E}[v^1] = 1$, $\gamma = 0.8$, $\mathbf{E}[\delta_t] = 8$, $\mathbf{E}[\epsilon_t] = 1.5$, $\mathbf{E}[\theta_t] = 0$, $T = 90$, $\mathbf{E}[\xi_t] = 0.85$ and ξ_t is Uniformly Distributed Over [0.7, 1]**

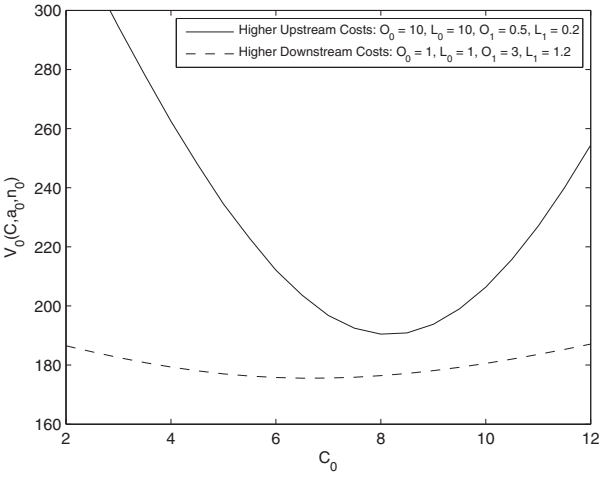
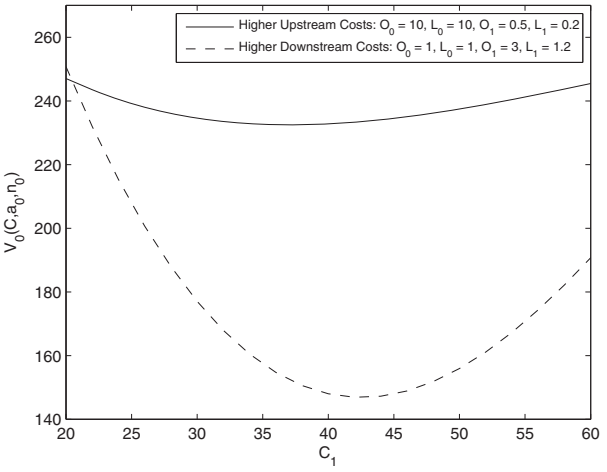


Figure 6 **Total Expected Cost $V_0(C, a_0, n_0)$ as a Convex Function of C_1. $a_0 = 18$, $n_0 = 13$, $W = 1$, $C_0 = 5$, $\mathbf{E}[v^0] = 1$, $\mathbf{E}[v^1] = 1$, $\gamma = 0.8$, $\mathbf{E}[\delta_t] = 8$, $\mathbf{E}[\epsilon_t] = 1.5$, $\mathbf{E}[\theta_t] = 0$, $T = 90$, $\mathbf{E}[\xi_t] = 0.85$ and ξ_t is Uniformly Distributed Over [0.7, 1]**



function of (C_0, C_1) for two different sets of parameters, when capacity use of different patients are i.i.d. That is, these figures are plotted for the original model presented in Section 4.

To make it easier to see, we further plot Figures 5 and 6 which show the convexity of the value function $V_t(C, a_t, n_t)$ in the capacities C_0 and C_1, respectively, when capacity use of different patients are i.i.d. These two figures use the same set of parameters, except that Figure 5 shows how V_t changes in upstream capacity C_0 with downstream capacity C_1 fixed, while Figure 6 presents how V_t changes by varying C_1 but fixing C_0. These trends imply that expanding the capacity of a single resource provides diminishing marginal returns. In addition, the optimal capacity of one stage depends on the relative weight of costs in both stages, and the total system cost is more sensitive to the capacity change in a stage that carries higher costs. Note that the cost function $V_0(C, a_0, n_0)$ also depends on the initial state (a_0, n_0). For different initial states, the optimal capacity level is different, but the impact of the initial state vanishes when the discount factor γ approaches 1 and the planning horizon increases.

In Figure 5, the downstream capacity is $C_1 = 30$. Each patient is expected to stay about $1/(1 - \mathbf{E}[\xi_t]) = 6.67$ days in the system, and therefore, on average, the downstream stage can discharge $C_1(1 - \mathbf{E}[\xi_t]) = 4.5$ patients each day, which is much smaller than the expected daily arrival rate of $\mathbf{E}[\delta_t] + \mathbf{E}[\epsilon_t] = 9.5$. As a result of the lack of downstream capacity, the optimal value of C_0 is sensitive to the relative weight between surgery costs (O_0, L_0) and bed-stay costs (O_1, L_1). When the surgery costs are relatively higher $(O_0 = 10, L_0 = 10, O_1 = 0.5, L_0 = 0.2)$, the system throughput is primarily driven by the need to use the upstream capacity efficiently, making the optimal value of C_0 closer to the external arrival rate 9.5. On the other hand, when the surgery costs are relatively lower $(O_0 = 1, L_0 = 1, O_1 = 3, L_1 = 1.2)$, the optimal value of C_0 is closer to $C_1(1 - \mathbf{E}[\xi_t]) = 4.5$, as the throughput is primarily determined by the downstream capacity.

In Figure 6, the capacity upstream is $C_0 = 5$, which is smaller than the average daily demand of

$\mathbf{E}[\delta_t] + \mathbf{E}[\epsilon_t] = 9.5$. The optimal downstream capacity C_1 again is sensitive to whether upstream or downstream costs dominate. With higher downstream costs, the total cost V_t is more sensitive to the choice of C_1 (see dotted lines). In this case, the optimal C_1 is driven by the urgency to satisfy patient demand from upstream, and also the need to consider the tradeoff between under-utilization and over-utilization downstream. Note that $3 = O_1 > L_1 = 1.2$, and thus we expect the optimal C_1 to be close to but smaller than $9.5/(1 - \mathbf{E}[\xi_t]) = 63$. When downstream costs are lower, the total cost V_t is less sensitive to the choice of C_1 (see solid lines). The optimal C_1 should match the capacity upstream, and is close to $C_0/(1 - \mathbf{E}[\xi_t]) = 33$.

6.2. Impact of Capacity Changes on the Optimal Decisions

In this section, we study how the optimal scheduling decisions change with varying levels of capacity. For the same technical reason above, we still assume that the total amount of stage-i capacity used by any k patients on any given day is given by the following: $v^i k$ for $i = 0, 1$, where v^i is a non-negative random variable. Then, we prove that the formulation is submodular with respect to the capacity at stage 1, that is, the downstream stage. The proof is similar to that of Theorem 1.

COROLLARY 2. *For every t and every fixed C_0,*

1. *$F(\cdot, \cdot, \cdot, \cdot)$ is submodular in (C_1, m_t, a_t, n_t);*
2. *$G_t(\cdot, \cdot, \cdot, \cdot)$ is submodular in (C_1, m_t, a_t, n_t); and*
3. *$V_t(\cdot, \cdot, \cdot)$ is submodular in (C_1, a_t, n_t).*

Submodularity implies, as before, monotonicity of the decisions in the downstream capacity, which is formalized in the corollary below.

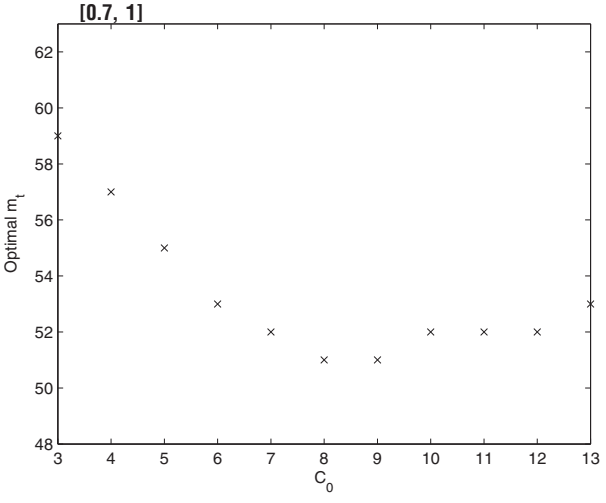
COROLLARY 3. *For every t and every fixed C_0, the optimal decisions $m_t^{\max}(C, a_t, n_t)$ and $m_t^{\min}(C, a_t, n_t)$ are increasing in the downstream capacity C_1.*

The monotonicity property stated in the corollary is easy to see. Any increase in capacity downstream enables more patients to be accommodated in the system regardless of the current level of capacity available upstream. Intuitively, a higher level of capacity downstream "pulls" more patients through the system.

In a sharp contrast, a similar result does not necessarily hold for the capacity upstream. That is, if we fix the capacity level C_1 downstream and increase the capacity level C_0 upstream, the optimal decisions $m_t^{\max}(C, a_t, n_t)$ and $m_t^{\min}(C, a_t, n_t)$ might *not* increase. To see, note that upstream capacity affects the rate at which patients may be admitted into the hospital, whereas the decision m_t controls the total number of patients in the hospital at t. An increase in upstream capacity has a countervailing effect on the total number of patients in the hospital at t. On the one hand, more patients might be admitted without incurring high upstream overtime costs. This effect tends to increase the total number of patients in the hospital at t. On the other hand, with an increase in upstream capacity, more patients can be admitted in the future who will be sharing the limited amount of available downstream capacity with current patients. This effect may induce the system manager to decrease the total number of patients in the hospital at t by reducing daily admits to avoid incurring high downstream overtime cost immediately. Instead, the manager knows that she could admit more patients in the future, she may choose to leave more patients on the waitlist but admit them later.

The numerical example in Figure 7 illustrates the points above. In this case, the waiting cost is $W = 1$, while the overtime cost in upstream stage, $O_0 = 3$, is three times that. Thus, if the system could discharge a patient within 3 days of his arrival without using surge capacity, the scheduler would choose to keep a patient on the waitlist to avoid incurring overtime costs right away. When the upstream capacity is relatively small compared to the length of the waitlist ($w_t = a_t - n_t = 30$ patients), the optimal number of patients admitted is at a level beyond which admitting any additional patient would almost surely incur overtime cost in the upstream stage. (For example, when $C_0 = 3$, the optimal number of admissions $q_t = m_t - n_t = 19$, and these patients would collectively consume 19 units of resource upstream on

Figure 7 The Optimal Decisions $m_t^{\max}(C, a_t, n_t)$ and $m_t^{\min}(C, a_t, n_t)$ are Not Monotonic in C_0. $a_t = 70$, $n_t = 40$, $W = 1$, $O_0 = 3$, $L_0 = 3$, $O_1 = 1$, $L_1 = 1$, $C_1 = 48$, $\mathbf{E}[v^0] = 1$, $\mathbf{E}[v^1] = 1$, $\gamma = 0.9$, $\mathbf{E}[\delta_t] = 6$, $\mathbf{E}[\epsilon_t] = 0.5$, $\mathbf{E}[\theta_t] = 0$, $T = 50$, $\mathbf{E}[\xi_t] = 0.85$ and ξ_t is Uniformly Distributed Over [0.7, 1]



average.) At such a low capacity level $C_0 = 3$, additional upstream capacity allows the system to process more patients on the waitlist in the future, and thus leads the optimal decision m_t to be *smaller* in order to save upstream overtime costs now and also to alleviate congestion downstream. However, as C_0 increases sufficiently, the emphasis of the optimal scheduling policy changes from preventing overtime costs to reducing idling costs in the upstream stage, and as a result it may admit more patients from the waitlist.

In summary, increasing the upstream capacity allows a manager to be more reactive to fluctuations in downstream congestion; the manager might not use the full upstream capacity as the hospital might prefer to ramp up or ramp down admissions in order to "chase" inpatient occupancy. This is fundamentally different from a single-stage model that only focuses on optimizing the upstream operations. In a single-stage model, the classic result is that more patients should be admitted from the waitlist when the (surgical) capacity increases (Ayvaz and Huh 2010, Huh et al. 2013). However, if the manager takes a global perspective and looks at two stages of service as a whole, then more capacity in the upstream stage does *not* necessarily call for admitting more patients from the waitlist, and only an increase in the downstream capacity does. Indeed, the more straightforward relationship between optimal scheduling decisions and downstream capacity seems to suggest that (utilization of) downstream capacity may serve as a good guide for scheduling decision making, which we will discuss in our numerical studies next.

7. Numerical Studies

As discussed earlier, previous literature on healthcare operations management has either focused on dynamic scheduling decisions for a single stage (e.g., Gerchak et al. 1996), or considered systems design under static scheduling rules in a facility with multiple units (e.g., Helm and Van Oyen 2014). Only a few recent studies have investigated how to dynamically schedule patients taking into account downstream capacity and patient census information (Helm et al. 2011, Samiedaluie et al. 2017). Our paper is among the first to analytically study such integrated decision making. Our numerical experiments in this section follow up on the theoretical work above to investigate the performance of our proposed scheduling method and compare it with traditional scheduling policies that make decisions independently of operations in other units. It is expected to see that integrated scheduling policies perform better than traditional ones. Our numerical study goes beyond this and aims to provide deeper insights to the following questions: (i) how much improvement integrated scheduling can make; (ii) under what conditions the gains tend to be significant; and (iii) how robust is the improvement made by integrated scheduling.

To study integrated scheduling in a realistic setting, we populate our experiments with data collected from the Congenital Heart Center at CUMC. These data include information on all children ≤21 years of age undergoing cardiac surgery in the pediatric operating rooms (OR) in 2014 and recovering in the pediatric cardiac intensive care unit (CICU). Some patients move to a step-down unit, after CICU, before discharge from the hospital. Patients recovering in the neonating CICU and adults operated on in the adult OR and recovering in the adult CICU are excluded, as these patients move along a different care pathway. In total, there were 572 surgery cases performed in 2014. The dataset records the date of the surgery, whether it was an elective one or emergency one, and the length of stay of the patient after surgery.

This clinical setting fits our model well, as discussed earlier. Specifically, we treat OR as the upstream stage and the entire inpatient postoperative stay as the downstream stage. In this setting, hospital discharges to home are dependent on patient readiness, but discharges from the CICU to the step-down unit are often dependent on hospital factors unrelated to patients' health. One potential deviation from the data and our modeling context is that our model is mainly motivated by the Canadian healthcare system (where a waitlist is used to manage elective surgeries), while our data comes from a US healthcare institution (in which advance scheduling is used in elective surgical admission). Although the scheduling approaches are different in these two systems, the data is still very valuable because it allows us to populate a more realistic clinical environment (than otherwise a fictitious setting created by assumed data) to test our model.

We test the empirical performances of the following scheduling policies.

- OPT: An optimal scheduling policy where variables a_t, m_t, n_t only take discrete values. In this policy, the number n_t after patients have left the downstream stage is rounded to the nearest integer.
- OPT-F: A fractional optimal scheduling policy that we have analyzed in this study, where a_t, m_t, n_t take non-negative real values. When used in practice, the decision variable m_t can be rounded to the nearest integer.
- SINGLE0: A single-stage "optimal" scheduling policy that treats the downstream capacity as infinity. This mimics what the current practice usually follows—a patient scheduling policy solely based on OR usage.

- SINGLE1: A single-stage "optimal" scheduling policy that treats the upstream capacity as infinity. This is a downstream capacity-driven policy.

We compare the performance of the optimal policies (OPT and OPT-F) against two single-stage policies (SINGLE0 and SINGLE1). Each single-stage policy treats one of the stages, either stage 0 or stage 1, as the only bottleneck stage and assumes infinite capacity in the other stage. A single-stage policy can be thought of as being used by a manager operating her own unit in isolation and making decisions without regards to their impact on other units. A large performance gap between the optimal policy and the single-stage policy indicates a higher *value of integrated scheduling*.

We conduct two extensive sets of numerical experiments. First, we consider the model setting, that is, the environment assumed by our model. This setting allows us to focus on the potential impact of factors other than environmental complexities (e.g., non-stationary capacity changes) on the value of integrated scheduling, for example, patient mixes, provider idling and overtime costs.

In addition, we study the performance of scheduling policies suggested by our model in a variety of more complex environments, including those with a strict capacity constraint on downstream beds, non-stationary demand and capacity changes, patient abandonment and correlated capacity usage in two stages. These more realistic environments are populated by the actual data and allow us to investigate the robustness of the performance of integrated scheduling. We benchmark the performances of integrated scheduling policy with those of the actual strategy currently used by the hospital. This comparison also shows how integrated scheduling could help improve practice.

7.1. Model Setting

We use the following parameters in our model setting.

Arrival Rates. The average daily number of admissions is 1.23 for elective patients and 0.98 for emergency patients, averaged over all weekdays in 2014. Accordingly, we set the elective arrival rate $\mathbf{E}[\delta_t] = 1.23$ and the emergency arrival rate $\mathbf{E}[\epsilon_t] = 0.98$ in our base setting. We also consider other mixes of patient arrivals such that $(\mathbf{E}[\delta_t], \mathbf{E}[\epsilon_t]) \in \{(0.63, 1.58), (1.83, 0.38), (1.48, 0.98)\}$. In the first two cases, the total arrival rate of elective and emergency patients is fixed at 2.21, the same as the base setting. In the third case, the emergency arrival rate is fixed, but the elective arrival rate is assumed to be 20% higher than in the base setting. The number of arrivals on each day is assumed to be a Poisson random variable. Since in our collaboration organization no patients are directly admitted to the downstream stage, we assume $\theta_t = 0$ in our numerical study.

Upstream and Downstream Capacities. The upstream capacity is fixed at $C_0 = 3$, and the downstream capacity C_1 is varied in the numerical experiments.

Cost Rates. We set the waiting cost $W = 1$ without loss of generality. One classic way to measure the value of time is its opportunity cost via the wage rate (Becker 1965). The median personal income in the US is $24,062, and the idling cost (e.g., due to clinician's income) may be multiple times of that. Thus, we experiment with different idling costs in two stages by choosing $(L_0, L_1) \in \{(10, 10), (5, 15), (15, 5)\}$. To set the overtime cost rate, we note that the US federal government mandates overtime salary rate to be at least 1.5 times of the regular time salary rate. Using this as a benchmark, we choose overtime costs to be $O_0 = 15, 30$, and $O_1 = 40$. Our choice on the ratios between idling cost (overtime cost) and patient waiting cost is aligned with those in the prior literature. For instance, Gerchak et al. (1996) assume the overtime cost rate to be $15 and the wait cost to be 0 to $2.99, that is, the overtime cost would be in the range of $(5, \infty)$ if waiting cost is normalized to 1.

Discharge Rate. We assume that a random fraction $1 - \xi_t$ of patients are discharged from the downstream stage on each day t. We use the empirical distribution to generate ξ_t, that is, ξ_t is the fraction of patients who continue to stay in downstream on a random day in 2014. Specifically, we uniformly sample days in 2014, and set $1 - \xi_t$ to be the fraction of patients leaving on those days. The average daily discharge rate is 6.8%, that is, $1 - \mathbf{E}[\xi_t] = 6.8\%$.

Planning Horizon and Other Parameters. We simulate the total costs of the optimal policies and single-stage policies over an infinite horizon. We randomly sample patients from the pool of all patients admitted in 2014. We initialize the system with a non-empty inpatient unit. We also sample the patients who are initially staying in the downstream stage from the pool of real patients in 2014. We set the resource usage to be $v^0 = v^1 = 1$. That is, each patient consumes one unit of capacity. The discount factor is $\gamma = 0.95$.

For each combination of the chosen parameters, we calculate the total discounted costs under each of the scheduling policies discussed above. In all test cases the difference between OPT and OPT-F is within 1%,

so we do not report the performance of OPT-F here. For easy comparison, we present all results as the performance ratio with respect to the total discounted cost under OPT. We summarize the ratios SINGLE0/OPT, SINGLE1/OPT, and the minimum of these two ratios in Tables 1–3; to save space, additional results are shown in Tables A1–A2 in the Online Appendix D. These cost ratios indicate the value of integrated scheduling. Higher ratios correspond to more cost saving when integrated scheduling is used. Note that Tables 1 and 2 have the same patient arrival rates as in the base setting, but Table 2 has a higher upstream overtime cost rate O_0. Table 3 has the same parameter setting as in Table 2 except for the patient mix.

We see that when we misidentify the bottleneck stage, that is, when we use the worse of SINGLE0 and SINGLE1, the performance gap between the integrated scheduling policy and policies based on single-stage optimization can be quite significant. In some cases, integrated scheduling can reduce costs by more than 50% (i.e., the cost ratio is larger than 2). However, even when we correctly identify the bottleneck stage—that is, when we select the better policy between SINGLE0 and SINGLE1—integrated scheduling may still reduce costs by up to 11%.

More specifically, the cost ratio SINGLE0/OPT represents the performance of a system admitting patients according to a surgical scheduler who ignores the operations in the downstream inpatient stage. This is indeed what the current practice usually follows—admitting patients only based on the OR capacity. We see that the largest ratio of SINGLE0/OPT is 310% (Table A2 in the Online Appendix D), suggesting that *the system overspends three-fold when not taking the downstream stage into account*. More importantly, in more than 60% of the scenarios tested, the system overspends by more than 20% when overlooking downstream.

As the downstream capacity C_1 increases, we observe that this cost ratio SINGLE0/OPT consistently decreases across the three tables. This is because that stage 0 becomes the bottleneck stage as C_1 gets larger, making SINGLE0 to behave similarly to OPT and closing the performance gap of these two policies. Comparing Tables 1 and 2, we also see that SINGLE0 in general performs better when the overtime cost upstream O_0 is larger. The explanation is

Table 1 Performance of Different Scheduling Policies When $E[\delta_t] = 1.23$, $E[\epsilon_t] = 0.98$ and $O_0 = 15$

	SINGLE0/OPT			SINGLE1/OPT			min(SINGLE0,SINGLE1)/OPT		
C_1	$L_0 = 10$ $L_1 = 10$	$L_0 = 5$ $L_1 = 15$	$L_0 = 15$ $L_1 = 5$	$L_0 = 10$ $L_1 = 10$	$L_0 = 5$ $L_1 = 15$	$L_0 = 15$ $L_1 = 5$	$L_0 = 10$ $L_1 = 10$	$L_0 = 5$ $L_1 = 15$	$L_0 = 15$ $L_1 = 5$
30	2.230	2.163	2.336	1.014	1.003	1.031	1.014	1.003	1.031
31	1.976	1.895	2.112	1.013	1.003	1.035	1.013	1.003	1.035
32	1.742	1.656	1.887	1.015	1.003	1.036	1.015	1.003	1.036
33	1.545	1.468	1.681	1.014	1.002	1.036	1.014	1.002	1.036
34	1.385	1.321	1.508	1.013	1.001	1.036	1.013	1.001	1.036
35	1.265	1.217	1.362	1.011	1.002	1.034	1.011	1.002	1.034
36	1.177	1.143	1.251	1.011	1.001	1.032	1.011	1.001	1.032
37	1.117	1.093	1.170	1.010	1.001	1.029	1.010	1.001	1.029
38	1.075	1.060	1.111	1.009	1.001	1.026	1.009	1.001	1.026
39	1.048	1.038	1.072	1.007	1.001	1.024	1.007	1.001	1.024
40	1.030	1.023	1.045	1.006	1.001	1.022	1.006	1.001	1.022

Table 2 Performance of Different Scheduling Policies When $E[\delta_t] = 1.23$, $E[\epsilon_t] = 0.98$ and $O_0 = 30$

	SINGLE0/OPT			SINGLE1/OPT			min(SINGLE0,SINGLE1)/OPT		
C_1	$L_0 = 10$ $L_1 = 10$	$L_0 = 5$ $L_1 = 15$	$L_0 = 15$ $L_1 = 5$	$L_0 = 10$ $L_1 = 10$	$L_0 = 5$ $L_1 = 15$	$L_0 = 15$ $L_1 = 5$	$L_0 = 10$ $L_1 = 10$	$L_0 = 5$ $L_1 = 15$	$L_0 = 15$ $L_1 = 5$
30	2.163	2.081	2.285	1.028	1.012	1.054	1.028	1.012	1.054
31	1.921	1.825	2.065	1.029	1.012	1.060	1.029	1.012	1.060
32	1.698	1.606	1.846	1.030	1.012	1.062	1.030	1.012	1.062
33	1.510	1.431	1.647	1.029	1.011	1.063	1.029	1.011	1.063
34	1.361	1.296	1.476	1.027	1.010	1.061	1.027	1.010	1.061
35	1.248	1.200	1.339	1.025	1.009	1.059	1.025	1.009	1.059
36	1.166	1.131	1.234	1.024	1.008	1.055	1.024	1.008	1.055
37	1.108	1.085	1.157	1.022	1.008	1.052	1.022	1.008	1.052
38	1.070	1.054	1.103	1.020	1.007	1.048	1.020	1.007	1.048
39	1.044	1.034	1.066	1.018	1.006	1.045	1.018	1.006	1.045
40	1.027	1.021	1.042	1.016	1.006	1.041	1.016	1.006	1.041

Table 3 Performance of Different Scheduling Policies When $E[\delta_t] = 1.83$, $E[\epsilon_t] = 0.38$, and $O_0 = 30$

	SINGLE0/OPT			SINGLE1/OPT			min(SINGLE0,SINGLE1)/OPT		
C_1	$L_0 = 10$ $L_1 = 10$	$L_0 = 5$ $L_1 = 15$	$L_0 = 15$ $L_1 = 5$	$L_0 = 10$ $L_1 = 10$	$L_0 = 5$ $L_1 = 15$	$L_0 = 15$ $L_1 = 5$	$L_0 = 10$ $L_1 = 10$	$L_0 = 5$ $L_1 = 15$	$L_0 = 15$ $L_1 = 5$
30	2.565	2.442	2.662	1.052	1.032	1.090	1.052	1.032	1.090
31	2.210	2.074	2.352	1.055	1.033	1.098	1.055	1.033	1.098
32	1.899	1.761	2.055	1.056	1.032	1.104	1.056	1.032	1.104
33	1.642	1.533	1.796	1.055	1.030	1.107	1.055	1.030	1.107
34	1.446	1.359	1.575	1.052	1.028	1.107	1.052	1.028	1.107
35	1.301	1.241	1.402	1.050	1.026	1.105	1.050	1.026	1.105
36	1.201	1.153	1.275	1.047	1.024	1.102	1.047	1.024	1.102
37	1.131	1.099	1.182	1.044	1.022	1.097	1.044	1.022	1.097
38	1.083	1.063	1.117	1.040	1.020	1.092	1.040	1.020	1.092
39	1.052	1.039	1.075	1.038	1.019	1.087	1.038	1.019	1.075
40	1.032	1.024	1.046	1.035	1.017	1.082	1.032	1.017	1.046

that when upstream costs become more significant, a policy that aims to minimize costs upstream is likely to perform well.

The cost ratio of SINGLE1/OPT indicates the system performance following decisions of an inpatient unit manager who does not pay attention to the OR usage. Our experiments show that the system could overspend up to 11% of the optimal cost without considering the upstream stage. The most significant improvement of OPT over SINGLE1 occurs in the neighborhood of $C_1 = 32$ to 36, where the downstream capacity is balanced with the external demand (for instance, C_1 needs to be roughly 2.21/6.8% = 32.5 in Tables 1–3 to match demand). We also note that SINGLE1 consistently outperforms SINGLE0. This is likely because that the daily cost (both idling and overuse) incurred in downstream is of a larger magnitude than upstream. As a result, a scheduling policy that focuses on downstream tends to perform better than one that only looks at upstream.

In summary, single-stage policies SINGLE0 or SINGLE1, which make scheduling decisions based on capacity usage at only one stage in the system, can lead to significant inefficiency and financial loss. Integrated decision making, conversely, can bring significant values to the system as a whole. We remark the following two conditions under which integrated scheduling is particularly beneficial.

- First, integrated scheduling is quite beneficial when average demand and capacities in the system are relatively balanced. This is likely due to the fact that when one of the stages is clearly short on capacity, there is not much room for integrated scheduling to make a difference. By identifying the bottleneck stage correctly, simple policies like SINGLE0 or SINGLE1 may already perform reasonably well.
- Second, integrated scheduling can be more valuable when there are more elective patients in the patient mix. In particular, comparing Tables 2 and 3 (and Table A1 in the Online Appendix D) reveals that improvement due to adopting integrated scheduling instead of single-stage policies is consistently higher when the system sees more elective patients. This is likely because a higher volume of elective patients gives more room for integrated scheduling to control and optimize the system, and at the same time fewer emergency patients lead to lower uncertainty in the system, both resulting in lower costs.

Before closing this section, we note that while our current numerical study focuses on somewhat small systems motivated by our dataset ($C_0 = 3$, $C_1 \approx 35$), our MDP model can handle relatively large systems (say, ten times as large as the base setting) in a reasonable amount of time. See Table A3 in the Online Appendix D for some computation time results with different system scales.

7.2. Complex Environments

In this section, we describe our numerical study on the applicability and performance of our model in a variety of more complex environments. We first investigate a situation in which the downstream beds have a strict capacity constraint, that is, no more than C_1 patients can be admitted downstream. We compare the performances of integrated scheduling and single-stage policies. Then, we use our dataset to populate a non-stationary environment in which OR capacity and patient demand are changing over time. In such a non-stationary environment, we further consider the impact of potential patient abandonment and correlated capacity usage in two stages of services. We compare policies suggested by our

integrated scheduling model and the *actual* policies used in practice.

7.2.1. Strict Capacity Constraint in Downstream. When the downstream stage is not a regular recovery unit but provides more intensive or specialized care (e.g., PACU or ICU), then it is often difficult to find alternative beds to place additional patients if the downstream unit is at capacity. That is, the capacity constraint in such downstream units is *strict*. Indeed, surgeries can be canceled due to insufficient ICU beds (Kahn 2012). Our model can provide an approximate way to incorporate such a strict constraint on downstream capacity, by setting the overtime cost sufficiently large once the downstream census is beyond its capacity C_1. Next, we demonstrate such an application using a set of numerical experiments.

We adopt the same parameter setting described in section 7.1, except that we assume the downstream overtime cost now takes the form $O_1[(n - C_1)^+]^\kappa$, where n is the downstream census and $\kappa \geq 1$ is a parameter of choice. A larger κ suggests more difficulty to find surge capacity in downstream. In our numerical experiments, we let $\kappa \in \{1.2, 2.0\}$ to study how different levels of constraint strictness on downstream capacity affect system performances. Table 4 presents the detailed results when $\kappa = 2.0$, and additional results can be found in the Online Appendix (Table A4). In Table 4, we also show the average daily downstream overflows (i.e., those admitted beyond the downstream capacity C_1) in the last three columns under different policies, respectively.

In Table 4, we observe that integrated scheduling makes a sizable improvement over the single-stage policies (manifested by the performance ratios); the daily downstream overflow is almost negligible when the downstream overtime cost takes a quadratic form ($\kappa = 2.0$). Comparing to the results with $\kappa = 1.2$ shown in the Online Appendix, we see the downstream overflow decreases as the value of κ increases. Therefore, in practice managers can choose a proper value of κ to enforce a desirable level of strictness on the downstream capacity constraint.

We also find that SINGLE1 performs much better than SINGLE0. This is not surprising, because when surge capacity downstream becomes much costly, scheduling decisions driven by downstream capacity use ought to outperform those solely depending on upstream capacity (and ignoring downstream). This suggests that managers should *not* admit patients only based on the OR capacity, especially when downstream capacity is inflexible. In this case, a policy driven by downstream capacity use, that is, SINGLE1, is likely a good heuristic to use in practice.

Table 4 Performance of Different Scheduling Policies When $E[\delta_t] = 1.23$, $E[\epsilon_t] = 0.98$, $\kappa = 2.0$, $O_0 = 30$, $O_1 = 40$ and $L_0 = L_1 = 10$

	Performance ratio			Average downstream overflow		
C_1	SINGLE0/ OPT	SINGLE1/ OPT	min(SINGLE0, SINGLE1)/OPT	SINGLE0	SINGLE1	OPT
30	7.162	1.028	1.028	2.73	0.09	0.11
31	6.171	1.031	1.031	2.23	0.08	0.09
32	5.196	1.032	1.032	1.77	0.08	0.07
33	4.094	1.031	1.031	1.37	0.07	0.07
34	3.327	1.032	1.032	1.08	0.06	0.05
35	2.588	1.031	1.031	0.80	0.04	0.04
36	2.090	1.028	1.028	0.60	0.03	0.03
37	1.672	1.025	1.025	0.44	0.02	0.02
38	1.421	1.024	1.024	0.31	0.02	0.02
39	1.271	1.021	1.021	0.22	0.01	0.01
40	1.170	1.019	1.019	0.15	0.01	0.01

7.2.2. Non-Stationary Environment. In this section, we consider an environment in which OR capacity and patient demand are changing over time. We populate such a non-stationary environment using data observed in a three-month period from April 2014 to June 2014. In this period, the upstream capacity largely depends on the number of ORs available to the division. Table 5 summarizes the number of ORs available in each day. The upstream capacity C_0 is measured by the number of surgeries that can be performed on each day. Given the duration of these particular types of operations and associated room turnover times, usually two cases can be performed in an OR per day. That is, C_0 takes a value that is twice of the number of ORs available.

For the downstream stage, we analyze the whole year's data to get a sense of the daily census. There is no significant long-term trends in the data. The average daily number of patients staying in the downstream stage is 23 over the year 2014. Instead of studying a fixed downstream capacity, we test various downstream capacity values in the range [15, 29], as they cover most of the range of actual values (see Figure A2 in the Online Appendix C for a histogram of the downstream census). Indeed, in our clinical setting, the exact number of downstream beds available

Table 5 Upstream Capacities Used in the Tests of Non-Stationary Environments

	1st and 3rd Monday	2nd and 4th Monday	Tue	Wed	Thur	Fri	Sat	Sun
Number of available ORs	1	2	2	2	1	1	0	0
Upstream capacity C_0	2	4	4	4	2	2	0	0

for postoperative patients may not be constant as some of the chronic patients might hold beds there for a long time, e.g., a child awaiting a cardiac transplant might occupy a bed for months before receiving a heart.

Other parameters in our experiments are set as follows.

Arrival Rates. We use the long-term admission rate as the arrival rate during weekdays just like in the base model setting above (i.e., $\mathbf{E}[\epsilon_t] = 0.98$ and $\mathbf{E}[\delta_t] = 1.23$). The arrival rates during weekends, however, are zero for both elective and emergency patients because no elective cases are scheduled in weekends and very few emergency patients arrived during weekends either (only 1 case every two to three months according to our data).

Cost Rates. We normalize $W = 1$, set $(L_0, L_1) = \{(5, 15), (15, 5)\}$, $O_0 = 20$, and $O_1 = 40$. We choose a higher downstream overage cost rate O_1 than upstream overage cost rate O_0, because in our setting the downstream is ICU and overuse of ICU capacity may block critical patients, leading to catastrophic (health) outcomes.

Planning Horizon and Other Parameters. We simulate the costs of different scheduling policies using real data on patient surgery dates and LOS within the three-month period from April 2014 to June 2014. Specifically, we apply different scheduling policies to the sample arrivals over this period, and then sum up the costs incurred on each day. We set v^0, v^1 and γ to be the same as in the model setting above. Because we do not have data on the actual dates when patients or referring physicians request surgeries, we use the hospital admission date as the arrival date. This last assumption has two important implications.

1. Waiting costs computed in our simulations are smaller than real cost (incurred when policies were implemented in reality) by the same amount, for all policies we consider.
2. The optimal policies and single-stage policies have less flexibility to assign patients to the future, because they are required to make patients wait for at least the same number of days as actual practice. Therefore, the performance we report for these policies is worse than if implemented in reality.

We evaluate the cost ratios between the optimal policies and the actual strategy, as well as those of the single-stage policies and the actual strategy (see Table 6). Note that costs used in Table 6 are not expected future costs but based on sample arrivals; therefore it is possible that SINGLE1 outperforms OPT marginally in the table. Due to the two implications mentioned above, the cost ratios shown in Table 6 are more *conservative* (i.e., larger) than they would be were the optimal policy and single-stage policies implemented in reality. We also note that the actual patient LOS in our data does *not* follow a geometric distribution; see Figure A1 in the Online Appendix C and its notes.

We observe that in this non-stationary environment we tested, integrated scheduling still performs the best and leads to significant improvement over simpler policies and over actual practice, even when the underlying LOS is not geometrically distributed as assumed in the model. These results suggest that *the model, although it is stylized in some of its assumptions, still has value in a range of settings that fall outside of the modeled setting*.

We also observe that the actual strategy performs quite close to SINGLE0. This conforms to the

Table 6 Comparison Against Actual Practice in a Non-Stationary Environment

	$L_0 = 5,\ L_1 = 15$				$L_0 = 15,\ L_1 = 5$			
C_1	$\frac{\text{OPT}}{\text{ACTUAL}}$	$\frac{\text{SINGLE0}}{\text{ACTUAL}}$	$\frac{\text{SINGLE1}}{\text{ACTUAL}}$	$\frac{\min(\text{SINGLE0}, \text{SINGLE1})}{\text{ACTUAL}}$	$\frac{\text{OPT}}{\text{ACTUAL}}$	$\frac{\text{SINGLE0}}{\text{ACTUAL}}$	$\frac{\text{SINGLE1}}{\text{ACTUAL}}$	$\frac{\min(\text{SINGLE0}, \text{SINGLE1})}{\text{ACTUAL}}$
15	0.51	0.99	0.51	0.51	0.53	0.97	0.53	0.53
16	0.46	0.99	0.46	0.46	0.50	0.97	0.49	0.49
17	0.45	0.99	0.45	0.45	0.48	0.96	0.44	0.44
18	0.44	0.99	0.44	0.44	0.47	0.94	0.46	0.46
19	0.44	0.99	0.44	0.44	0.48	0.92	0.50	0.50
20	0.42	0.98	0.44	0.44	0.45	0.87	0.49	0.49
21	0.37	0.98	0.38	0.38	0.43	0.83	0.44	0.44
22	0.49	0.98	0.47	0.47	0.42	0.84	0.43	0.43
23	0.60	0.97	0.61	0.61	0.47	0.85	0.54	0.54
24	0.74	0.97	0.75	0.75	0.58	0.89	0.67	0.67
25	0.84	0.97	0.85	0.85	0.67	0.91	0.78	0.78
26	0.88	0.98	0.90	0.90	0.75	0.90	0.83	0.83
27	0.93	0.98	0.93	0.93	0.82	0.93	0.89	0.89
28	0.97	1.00	0.97	0.97	0.89	0.96	0.96	0.96
29	0.97	1.00	0.97	0.97	0.92	0.95	0.96	0.95

anecdotal note we received from this hospital that patient admission decisions in practice are almost uniformly made based on OR usage without considering the downstream stage. The other noteworthy finding is that the performance of the optimal policy is relatively similar to that of SINGLE1 in most cases we tested. One important reason, as discussed earlier, is that the downstream stage incurs higher overall costs than the upstream. The other contributing factor here is that the downstream stage is actually the bottleneck in this system (we would need $C_1 = (1.23 + 0.98)/6.8\% = 32.5$ beds to balance the upstream demand). These observations echo our earlier findings that misclassifying the bottleneck in the system can lead to a significant loss. In addition, we note that the manager can benefit significantly from an admission policy based on the usage of downstream capacity, rather than using the traditional OR capacity-driven admission.

7.2.3. Patient Abandonment. In practice, patients may leave the waitlist (i.e., abandon the queue) when wait times are too long. In this section, we test how such abandonment behavior, if it exists, affects the applicability of our model. Note that in our theoretical model we do not make assumptions about patient abandonment for tractability reasons. Indeed, in certain settings patients do not abandon. For instance, literally no patients ever abandon the surgical service in the hospital we collaborate with. According to our communication with clinicians, this is because that congenital heart surgery is a highly specialized field and that there are not many alternatives for care. The market competition for the hospital is primarily with providers in other states, so patients make decisions on which institution to receive care more on the basis of outcomes (rather than wait times). Nevertheless, it is of general interest to study the impact of potential patient abandonment, which is the focus of this section.

Using the non-stationary setting described in section 7.2.2, we test various scheduling policies, which are unaware of abandonment, in simulations where each patient has the same probability to leave the waitlist after each period. We consider several choices for this abandonment probability: 0.01, 0.1, 0.3 and 0.5. Detailed results can be found in Tables A5–A8 in the Online Appendix D.

We note that with patient abandonment, same insights obtained above remain valid. Integrated scheduling may save up to half of the total system costs compared to policies currently in use, and misclassifying bottleneck can lead to significant loss. In the practice environment we tested, the performance of a single stage policy based on downstream capacity use is close to that of integrated scheduling in most cases (because that the downstream incurs higher total costs than the upstream and it is also the bottleneck here).

7.2.4. Correlated Capacity Usage. In our theoretical model, we assume that capacity usage in two stages is independent, that is, the downstream LOS does *not* depend on the upstream surgical time. This assumption is valid in certain contexts (Wexner and Cera 2005); however, it is possible that capacity usage in two stages is correlated. In this section, we experiment with a setting in which the amounts of resource consumed in the upstream and downstream stages are correlated.

The evidence on the relationship of operative duration with hospital LOS is relatively limited and mixed as well (Tan et al. 2012). Thus, instead of using a hypothetical construct, we use exact observations in our data to capture such a correlation in simulation. Specifically, in our simulation the upstream (downstream) resource consumption is the *actual* surgery duration (hospital LOS). According to our data, the correlation coefficient between patient surgery duration and downstream LOS is 0.32, that is, patients with longer surgeries tend to stay in the downstream stage for more days.

As in earlier sections, we test various scheduling policies, which are generated based on our simple theoretical model by assuming i.i.d. capacity usage across different patients and no correlation between capacity usage in two stages, in a non-stationary environment where capacity usage in two stages is correlated in a way described by actual data. We compare the performance of these model-based scheduling policies with that under the actual scheduling policy in use; see Table A9 of the Online Appendix D for detailed results.

The performance of all scheduling policies is similar to that in the non-stationary setting. This is likely because, when a scheduling policy is unaware of patients' heterogeneous length-of-stay distributions, adding correlation between the upstream and downstream resource consumption does not affect the total cost of the policy too much in the long run. We note that the improvement made by integrated scheduling over the actual policy in use remains significant, even when the integrated policy is not aware of the correlation in capacity usage. As the downstream stage bears higher costs than the upstream (and it is also the bottleneck stage), a scheduling policy solely based on the downstream capacity usage works fairly well and has similar performances as integrated scheduling.

8. Conclusion

This study introduces a centralized scheduling decision model based on capacity usage in different units of a hospital. In particular, we analyze the first dynamic model that accounts for patient LOS and downstream census in daily scheduling decisions. We develop effective scheduling methods and provide useful insights for practitioners. Through extensive numerical experiments based on practical data, we demonstrate that using our model-based integrated scheduling policy enables hospitals to significantly improve their operational efficiency compared to following the policies currently in use, and that such superior performances are robust in a variety of complex environments. In the clinical environment we test (congenital heart surgery), we find that the traditional scheduling policy, which is solely driven by OR usage, may result in up to a three-fold increase in total expenses. In contrast, a single-stage policy based on the downstream capacity usage consistently outperforms the traditional scheduling policy. Indeed, such a scheduling policy based on downstream often performs relatively close to integrated scheduling, and therefore it may serve as a simple, effective scheduling heuristic for hospital managers. This insight is likely to extend to other hospital environments, because (downstream) hospital beds are known to represent the largest cost incurred by hospitals (Roberts et al. 1999).

Our theoretical model makes a few key assumptions for tractability, and these assumptions are (i) patients are willing to wait for service and to receive surgeries on relatively short notice; (ii) surge capacity in the downstream stage is available at extra costs; (iii) no one abandons during waiting; (iv) patient capacity utilizations in each stage are i.i.d. and capacity usages in two stages are independent; and (v) the discrete system can be well approximated by a continuous system. Our numerical experiments are designed to study the applicability and performance of our model with some of these assumptions relaxed. Our work provides a stepping stone to studying scheduling decisions in more complex settings. Future research may focus on systems with more than two stages and multiple patient classes with different urgency for care or expected LOS, and settings with strict downstream capacity constraints. Advance scheduling decisions that assign patients directly to future days are also an important future research topic. Along this line, one may also consider a "batch decision" setting, e.g., making elective admission decisions for the work days at the beginning of each week. In addition, recent advances in empirical studies support the endogeneity of patient LOS in certain hospital units. For instance, Kc and Terwiesch (2009, 2012) and Anderson et al. (2011) find that high occupancy levels can result in shorter patient LOS (due to the needs of accommodating new patients), while Chan et al. (2016) show that increased emergency room boarding times are associated with longer ICU LOS. In practice, hospitals may often have a projection of how many patients will leave today/tomorrow. Our theoretical models assume that patient resource utilization is exogenous and independent of system state; it would be an interesting and fruitful direction to incorporate endogeneity of patient resource usage, hospital projections for patient discharges, or both in surgical admission decisions.

Acknowledgments

The authors thank the department editor, the senior editor, and all the referees for their constructive comments throughout the review process.

Notes

[1]For convenience, we call stage 0 and stage 1 altogether as the hospital.
[2]This does *not* imply that the hospital with a more congested inpatient unit would admit more surgical patients; it only means that the resulting inpatient census in a congested unit remains higher under optimal control.

References

Adan, I. J. B. F., J. M. H. Vissers. 2002. Patient mix optimisation in hospital admission planning: A case study. *Int. J. Oper. Prod. Manage.* **22**(4): 445–461.

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

Anderson, D., C. Price, B. Golden, W. Jank, E. Wasil. 2011. Examining the discharge practices of surgeons at a large medical center. *Health Care Manage. Sci.* **14**(4): 338–347.

Ayvaz, N., W. T. Huh. 2010. Allocation of hospital capacity to multiple types of patients. *J. Rev. Pric. Manag.* **9**(5): 386–398.

Becker, G. S. 1965. A theory of the allocation of time. *Econ. J.* **75** (299): 493–517.

Bekker, R.P. M. Koeleman 2011. Scheduling admissions and reducing variability in bed demand. *Health Care Manage. Sci.* **14**(3): 237.

Beliën, J., E. Demeulemeester. 2007. Building cyclic master surgery schedules with leveled resulting bed occupancy. *Eur. J. Oper. Res.* **176**(2): 1185–1204.

Bowers, J. 2013. Balancing operating theatre and bed capacity in a cardiothoracic centre. *Health Care Manage. Sci.* **16**: 236–244.

Canadian Department of Health and Community Services. 2016. Wait Times: FAQs. Available at http://www.health.gov.nl.ca/health/wait_times/faq.html (accessed date October 27, 2016).

Chan, C. W., V. F. Farias, G. J. Escobar. 2016. The impact of delays on service times in the intensive care unit. *Management Sci.* **63**(7): 2049–2072.

Chao, X., H. Chen, S. Zheng. 2009. Dynamic capacity expansion for a service firm with capacity deterioration and supply uncertainty. *Oper. Res.* **57**(1): 82–93.

Choi, S., W. E. Wilhelm. 2014. An approach to optimize block surgical schedules. *Eur. J. Oper. Res.* **235**(1): 138–148.

Chow, V. S., M. L. Puterman, N. Salehirad, W. Huang, D. Atkins. 2011. Reducing surgical ward congestion through improved surgical scheduling and uncapacitated simulation. *Prod. Oper. Manag.* **20**(3): 418–430.

Cochran, J. K, A. Bharti. 2006. Stochastic bed balancing of an obstetrics hospital. *Health Care Manage. Sci.* **9**(1): 31–45.

Denton, B., D. Gupta. 2003. A sequential bounding approach for optimal appointment scheduling. *IIE Trans.* **35**(11): 1003–1016.

Diamant, A., J. Milner, F. Quereshy. 2018. Dynamic patient scheduling for multiappointment health care programs. *Prod. Oper. Manag.* **27**(1): 58–79.

Gartner, D., R. Kolisch. 2014. Scheduling the hospital-wide flow of elective patients. *Eur. J. Oper. Res.* **233**(3): 689–699.

Gerchak, Y., D. Gupta, M. Henig. 1996. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Sci.* **42**(3): 321–334.

Griffin, J., S. Xia, S. Peng, P. Keskinocak. 2012. Improving patient flow in an obstetric unit. *Health Care Manage. Sci.* **15**(1): 1–14.

Guerriero, F., R. Guido. 2011. Operational research in the management of the operating theatre: A survey. *Health Care Manage. Sci.* **14**(1): 89–114.

Gupta, D. 2007. Surgical suites' operations management. *Prod. Oper. Manag.* **16**(6): 689–700.

Helm, J. E., M. P. Van Oyen. 2014. Design and optimization methods for elective hospital admissions. *Oper. Res.* **62**(6): 1265–1282.

Helm, J. E, S. AhmadBeygi, M. P. Van Oyen. 2011. Design and analysis of hospital admission control for operational effectiveness. *Prod. Oper. Manag.* **20**(3): 359–374.

Huh, W. T., N. Liu, V.-A. Truong. 2013. Multiresource allocation scheduling in dynamic environments. *Manuf. Serv. Oper. Manag.* **15**(2): 280–291.

Hulshof, P. J. H., R. J. Boucherie, E. W. Hans, J. L. Hurink. 2013. Tactical resource allocation and elective patient admission planning in care processes. *Health Care Manage. Sci.* **16**(2): 152–166.

Jweinat, J., P. Damore, V. Morris, R. DAquila, S. Bacon, T. J. Balcezak. 2013. The safe patient ow initiative: A collaborative quality improvement journey at yalenew haven hospital. *Jt Comm. J. Qual. Patient Saf.* **39**(10): 447–AP9.

Kahn, J. M. 2012. The risks and rewards of expanding icu capacity. *Crit. Care* **16**(5): 156.

Kang, Keumseok, J George Shanthikumar, Kemal Altinkemer. 2016. Postponable acceptance and assignment: A stochastic dynamic programming approach. *Manuf. Serv. Oper. Manag.* **18**: 493–508.

Kc, D. S., C. Terwiesch. 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Sci.* **55**(9): 1486–1498.

Kc, D. S., C. Terwiesch. 2012. An econometric analysis of patient flows in the cardiac intensive care unit. *Manuf. Serv. Oper. Manag.* **14**(1): 50–65.

Kim, S.-C., I. Horowitz. 2002. Scheduling hospital services: The efficacy of elective-surgery quotas. *Omega* **30**(5): 335–346.

Kosnik, L. 2006. Breakthrough demand-capacity management strategies to improve hospital flow, safety and satisfaction. R. W. Hall, *Patient Flow: Reducing Delay in Healthcare Delivery* Springer, Boston, MA, 101–122.

Litvak, N., M. van Rijsbergen, R. J. Boucherie, M. van Houdenhoven. 2008. Managing the over flow of intensive care patients. *Eur. J. Oper. Res.* **185**(3): 998–1010.

Liu, N., S. Ziya, V. G. Kulkarni. 2010. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manuf. Serv. Oper. Manag.* **12**(2): 347–364.

May, J. H., W. E. Spangler, D. P. Strum, L. G. Vargas. 2011. The surgical scheduling problem: Current research and future opportunities. *Prod. Oper. Manag.* **20**(3): 392–405.

Nunes, L. G. N., S. V. de Carvalho, R. de Cássia Meneses Rodrigues. 2009. Markov decision process applied to the control of hospital elective admissions. *Artif. Intell. Med.* **47**(2): 159–171.

Patrick, J., M. L. Puterman, M. Queyranne. 2008. Dynamic multipriority patient scheduling for a diagnostic resource. *Oper. Res.* **56**(6): 1507–1525.

Price, C., B. Golden, M. Harrington, R. Konewko, E. Wasil, W. Herring. 2011. Reducing boarding in a post-anesthesia care unit. *Prod. Oper. Manag.* **20**(3): 431–441.

Robb, W. B., M. J. Osullivan, A. E. Brannigan, D. J. Bouchier-Hayes. 2004. Are elective surgical operations cancelled due to increasing medical admissions? *Ir. J. Med. Sci.* **173**(3): 129–132.

Roberts, R. R., P. W. Frutos, G. G. Ciavarella, L. M. Gussow, E. K. Mensah, L. M. Kampe, H. E. Straus, G. Joseph, R. J. Rydman. 1999. Distribution of variable vs. fixed costs of hospital care. *J. Am. Med. Assoc.* **281**(7): 644–649.

Robinson, G. H., P. Wing, L. E. Davis. 1968. Computer simulation of hospital patient scheduling systems. *Health Serv. Res.* **3**(2): 130–141.

Samiedaluie, S., B. Kucukyazici, V. Verter, D. Zhang. 2017. Managing patient admissions in a neurology ward. *Oper. Res.* **65**(3): 635–656.

Saure, A., J. Patrick, S. Tyldesley, M. L. Puterman. 2012. Dynamic multiappointment patient scheduling for radiation therapy. *Eur. J. Oper. Res.* **223**(2): 573–584.

Talluri, K. T., G. Van Ryzin. 2005. *The Theory and Practice of Revenue Management*, vol. 68. Springer Verlag, New York.

Tan, T.-W., J. A. Kalish, N. M. Hamburg, D. Rybin, G. Doros, R. T. Eberhardt, A. Farber. 2012. Shorter duration of femoral-popliteal bypass is associated with decreased surgical site infection and shorter hospital length of stay. *J. Am. Coll. Surg.* **215**(4): 512–518.

Topkis, D. M. 1998. *Supermodularity and Complementarity*. Princeton University Press, Princeton, NJ.

Wexner, S. D., S. M. Cera. 2005. Laparoscopic surgery for ulcerative colitis. *Surg. Clin.* **85**(1): 35–47.

White, D. L., C. M. Froehle, K. J. Klassen. 2011. The effect of integrated scheduling and capacity policies on clinical efficiency. *Prod. Oper. Manag.* **20**(3): 442–455.

Zacharias, C., M. Pinedo. 2014. Appointment scheduling with no-shows and overbooking. *Prod. Oper. Manag.* **23**(5): 788–801.

Zhang, B., H. Ayhan. 2013. Optimal admission control for tandem queues with loss. *IEEE Trans. Autom. Contr.* **58**(1): 163–167.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A. Ancillary Results.
Appendix B. Proofs of the Results.
Appendix C. Additional Figures.
Appendix D. Additional Numerical Results.