Submitted to *Operations Research*
manuscript (Please, provide the mansucript number!)

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

Van-Anh Truong, Xinshang Wang
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY, USA,
vatruong@ieor.columbia.edu

Nan Liu
Department of Health Policy and Management, Mailman School of Public Health, Columbia University, New York, NY, USA,
nl2320@columbia.edu

Motivated by the shortcoming of current hospital scheduling and capacity planning methods which often model different units in isolation, we introduce the first dynamic multi-day scheduling model that integrates information about capacity usage at more than one location in a hospital. In particular, we analyze the first dynamic model that accounts for patients' length-of-stay and downstream census in scheduling decisions. Via a simple and innovative variable transformation, we show that the optimal number of patients to be allowed in the system is increasing in the state of the system and in the downstream capacity. Moreover, the total system cost exhibits decreasing marginal returns as the capacity increases at any location independently of another location. Through numerical experiments on realistic data, we show that there is substantial value in making integrated scheduling decisions. In contrast, localized decision rules that only focus on a single location of a hospital can result in up to 60% higher expenses.

History:

1. Introduction

In many healthcare systems, patient care is delivered in several successive stages at different locations. For example, in maternity wards, women go through "labor rooms, delivery rooms, postpartum beds, [and] maternity beds." In coronary-care facilities, patients flow through "coronary care, post-coronary care, intensive care, medical care, surgical, and ambulatory-care [units]" (Cohen, Hershey, and Weiss 1980). In certain geriatric hospitals, patients pass from "acute-care" facilities to "rehabilitation" and "long-stay" facilities (El-Darzi, Vasilakis, Chaussalet and Millard 1998). In many ambulatory surgical centers, patients move from operating rooms (ORs) to post-anesthesia-care (PACU) facilities (Hsu, Ning, de Matta, and Lee 2003). In these multi-stage systems, patients' *length-of-stay* (LOS) at a particular stage, frequently a downstream stage, can span several days.

While staying at a downstream stage, patients consume the resource and service capacity there. If the downstream stage becomes fully occupied, access to it might be *blocked* for other patients upstream (Koizumi, Kuno, and Smith 2005).

Failure to account for patient LOS and potential blocking from downstream often leads to poor scheduling, inefficient use of capacity, and consequently, high cost and reduced quality of care. Robb, Osullivan, Brannigan, and Bouchier-Hayes (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, Intensive Care Units (ICUs) are often operated at or above nominal capacity (Chan, Yom-Tov, and Escobar 2011), and a shortage of such beds often forces surgeons to cancel or reschedule elective patients who might need ICU beds post surgery (Kim and Horowitz 2002). Gupta (2007) finds that "when surgeons book OR time for various procedures, quite often there is no mechanism in place to ensure that there are adequate downstream resources (specifically PACU, ICU, and bed capacity) for postoperative care." The resulting blocking can lead to "increased patient waits, excessive use of overtime, inability to handle emergency cases, and ambulance diversion."

In reality, the effective operation of a facility should depend on the "balanced" use of capacity at different stages to ensure smooth patient flow through the facility. Indeed, evidence indicates that accounting for patient LOS in downstream stages can confer significant benefits. Robinson, Wing, and Davis (1968) demonstrate an 8 to 17 percent cost reduction in a 100-bed hospital by taking into account patient LOS and downstream census when scheduling admissions. Griffin, Xia, Peng, and Keskinocak (2012) show that blocking can be prevented by balancing capacities at different stages in an obstetric department. Despite the evidence, many hospitals units still operate in isolation, without considering other 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).

With more hospitals adopting electronic health record systems following the Federal Government's "meaningful use" initiative (Blumenthal 2009), hospitals are acquiring the necessary information-technology support to coordinate capacity usage among different units and stages. However, the development of scheduling methods for multi-stage systems accounting for 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 such a model from being easily analyzed or solved.

In this paper, we consider a two-stage system in which patients receive surgeries at the upstream stage and may spend multiple days at the downstream stage. Both stages have finite capacity. There are two classes of patients who arrive randomly on each day. *Emergency* patients are always admitted on the day of arrival, 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). From this waitlist, a certain number of elective patients are chosen to be admitted each day. Once admitted, patients receive care at the upstream stage on the day of admission, and then move to the downstream stage, where they stay over multiple days before getting discharged. There are idling and overtime costs at both stages for under and over consumption of available capacity at these stages. There is also a waiting cost for making elective patients wait. Our goal is to determine a dynamic admission policy for elective patients to minimize the total discounted expected cost of 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 LOS in the downstream stage.

Our model is an aggregate planning model similar to those studied in Gerchak, Gupta and Henig (1996) and Ayvaz and Huh (2010). 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 overtime capacity usage with the cost of making patients wait and that of capacity under-utilization. In the second step, the sequence and timing of individual

procedures on a given day is determined to minimize within-day wait time of patients and idle time of providers. The second step is usually not explicitly considered in an aggregate-planning model.

Clearly, our model captures a relatively simple service system with two sequential stages and two customer classes. A hospital may have more complex network structure and patient priority rules. However our model is, to the best of our knowledge, the first attempt in the literature to explore dynamic scheduling decisions in a multi-stage system. More importantly, our simple model can be used to approximate more complex multi-stage healthcare systems, especially when there are two obvious bottleneck stages. In this paper, we will specifically focus on a two-stage cardiothoracic surgery centre described by Bowers (2013) as a canonical example. This Scotland-based centre primarily provides cardiothoracic surgical service. The majority of patients are elective patients. Patients receive scheduled operating theatre procedures, after which they transfer to ICUs, then to a High Dependency Unit (HDU), and finally on to a conventional ward before they are discharged from the centre. Patients' LOS in the ICUs ranges from 1 to 50 days. The HDU and ward capacities are high enough that they do not constrain the throughput of the centre. Thus, the ORs and the ICUs are the main stages that must be considered in determining daily elective admission. Clearly, the linked capacity of the ORs and the ICUs impacts the optimal number of elective patients to receive surgery 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 more than one service stage. In particular, ours is the first dynamic model that accounts for patient LOS in scheduling decisions (see discussions below on the relevant literature). We demonstrate that a formulation that uses the "natural" definition of decision variables does not generate helpful structural results or insights. But we are able to exploit a simple and yet innovative variable transformation to reveal a hidden submodularity structure in the formulation. Building upon this transformation, we prove that the number of patients allowed in the system in each period is monotone in the state variables and in the downstream capacity, thereby generating useful guidelines for adjusting scheduling decisions in practice. In addition, we show that the total expected cost of the

system exhibits decreasing marginal returns as the capacity in each stage increases independently of the the other stage, a result that has been confirmed earlier by simulation studies (Bowers 2013). In an infinite-horizon setting, we examine conditions under which the patient waitlist may grow without bounds. In particular, we show that the number of admissions might be uniformly bounded even as the number of patients waiting approaches infinity, as long as the waiting cost rate is low enough. Though this result seems intuitive, it has seldom been studied in the literature, and is particularly useful for informing the choice of model parameters to ensure that everyone on the waitlist is served. Finally, through numerical experiments on realistic data, we show that there are substantial values to make *integrated* scheduling decisions, that is, making scheduling decisions while taking into account patient LOS and census information downstream. We find that ignoring this information can lead to significant efficiency and financial loss to the whole system.

The remainder of this paper 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 studies the stability of the waitlist. Section 8 presents numerical results on the value of integrated scheduling. Finally, Section 9 discusses extensions of our model and concluding remarks.

2. Related Literature

Our work draws upon several streams of literature. The first stream is the literature on surgical scheduling. See Gupta (2007), Cardoen, Brecht, Demeulemeester, Belien (2010), May, Spangler, Strum, and Vargas (2010), 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 operating room time among different surgeons, how many operating rooms to be open and when to open them, and how to sequence different procedures on the day of operation. In this literature, the papers that are most relevant to our work include Gerchak, Gupta, and Henig (1996), Ayvaz and Huh (2010) and

Huh, Liu and Truong (2012). These papers investigate the allocation of elective patients to surgery days, using models similar to ours. They identify similar structure for the optimal scheduling policy. However, their work only models a single stage of service and does 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.

The second stream of research relevant to ours is the work on hospital capacity planning. This body of work recognizes different care paths for patients and develops either simulation or analytical models to study congestion in hospitals. Though this literature confirms the impact of downstream beds on overall system performance by modeling interactions among multiple units, they have seldom incorporated scheduling decisions into the models. Simulation models have been developed in a variety of settings, including geriatric departments (El-Darzi, Vasilakis, Chaussalet, and Millard 1998), obstetrics units (Griffin, Xia, Peng, and Keskinocak 2012), acute care services (Wang, Hare, Vertesi and Rutherford 2010) and special cardiac and thoracic surgery centers (Bowers 2013). More generally, Harper (2002) describes a framework for developing simulation models for hospital operations.

Queuing models featuring blocking have been studied extensively. See Bretthauer, Heese, Pun, and Coe (2011) for a brief review. The papers focusing specifically on hospital capacity planning include Hershey, Weiss, and Cohen (1981), Weiss, Cohen, and Hershey (1982), Koizumi, Kuno, and Smith (2005) and Bretthauer, Heese, Pun, and Coe (2011). These models make the assumption that systems are passive in their admission decisions; that is, patients are admitted in a greedy manner into a stage whenever there is sufficient capacity at the stage. As we will demonstrate in this paper, actively managed scheduling systems, where the decision to admit a patient accounts for resource usage in all succeeding stages, can significantly improve usage of capacity by smoothing out potential imbalances between the stages.

Recently, joint scheduling and capacity planning decisions have received growing attention. Our work contributes to this emerging literature. Our paper is distinct from this literature in the use of an analytical, dynamic model of scheduling. Previous work has mainly focused on simulation or

static optimization models. Ridge, Jones, Nielsen, and Shahani (1998) develop a simulation model for bed capacity planning in intensive care. Kim and Horowitz (2002) use simulation to study the impact of an *ad hoc* scheduling policy for elective surgeries on the capacity use of a downstream ICU. Hsu, Ning, de Matta, and Lee (2003) develop a deterministic model to schedule patients in an ambulatory surgical center to minimize the number of PACU nurses used. Helm and Van Oyen (2012) develop a static optimization model to determine the weekly schedule of elective hospital admission. Samiedaluie, Kucukyazici, Verter, and Zhang (2013) also consider patient admission policy in a neurology ward, where there are two stages of service, the ED and the neurology ward. Due to the complex actions allowed by their model, they do not identify the structure of an optimal scheduling policy.

Our work is also related to the literature on the classical flow shop scheduling. In these problems, a set of jobs flow sequentially through multiple stages, each stage consisting of one machine (Johnson 1954). The hybrid flow shop problem extends the classical flow shop problem by considering multiple machines in each stage, job- and machine-dependent processing times, and general precedence constraints for job processing. See Ruiz and Vázquez-Rodríguez (2010) and Ribas, Leisten, and Framinan (2010) for in-depth reviews of this literature. The key decisions in these problems include the allocation of jobs to machines as well as the sequencing of jobs through the shop. The objective is to minimize the completion time of the jobs, the delays of the jobs, or a combination of both. Our model differs in several respects. First, we do not explicitly model the assignment of patients to resources. Second, we only consider a simple and fixed sequence of procedures for patients. Third, while the flow shop problem focuses on a single period, our model makes dynamic, multi-period decisions. Finally, we allow patient demand and resource utilization to be random, whereas in the flow shop problem, job processing times are known deterministically.

3. Model

Although our model is more generally applicable, for concreteness, our terminology in the rest of the paper will be adapted to the canonical example of the cardiothoracic surgery centre described

by Bowers (2013). By convention and with a few exceptions, 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.) over time, and bounded. Emergency surgeries must be performed in 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 cost follows Gerchak, Gupta, and Henig (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 be moved to stage 1, called the *downstream stage*, for recovery and observation. The downstream stage may represent an intensive care unit (ICU), a progressive care unit (PCU), or an inpatient ward. The patient stays in the downstream stage for a random number of days before she is finally discharged.

We assume that there is a single bottleneck resource that is consumed by patients in each stage $i \in \{0, 1\}$. We called this resource *stage-i capacity* and denote it by C_i. For example, capacity might be measured in surgeon time in the entry stage or in number of ICU beds in the downstream stage. Each patient consumes a random amount υ^0 of capacity in stage 0, and υ^1 of capacity in each day that she remains in stage 1. For each $i \in \{0, 1\}$, we assume that υ^i is i.i.d. over the patient population and over time.

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. That is, we ignore any within day wait time by patients and idle time by doctors that depend on the sequencing and timing of procedures within a day. As discussed earlier, such approximation of the workload within a day is often used in aggregate-planning models; see, e.g., Gerchak, Gupta, and Henig (1996) and Ayvaz and Huh (2010).

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. Our overtime and idling costs are motivated as follows. We take the hospital's perspective and assume that there is a sunk fixed cost, for example facility maintenance cost, at each stage. There are also variable costs, such as staffing costs, that depend on the capacity installed at each stage. Take stage 0 as an example. Suppose that each scheduled patient brings in a revenue R per hour. Suppose also that the hospital maintains a fixed daily capacity level C every day, each unit of this capacity incurs a variable unit cost of A (e.g., salary rate), and overusing this capacity via overtime results in an overtime unit cost of P. If the number of patients scheduled for stage 0 is k, then the total cost incurred at stage 0 is $AC + P(S^0(k) - C)^+ - RS^0(k)$, which can be rewritten as $(A-R)C + (P-R)(S^0(k) - C)^+ + R(S^0(k) - C)^-$. Without loss of generality, we can drop the first term $(A-R)C$ as it is a constant with C fixed, and focus on the last two terms which depend on the scheduling decision k. In particular, we can think of $P - R$ as the unit overtime cost and R as the unit idling cost. Our implicit assumption is that $P \geq R$, so that the unit overtime cost is non-negative. If this condition is not satisfied, then we can directly model the costs P and R, the capacity level C and the revenue A. None of our structural results will change. We have chosen to assume $P \geq R$ to reduce the number of parameters for the problem. Following a similar argument, this cost structure also applies to stage 1.

To give a sense the magnitude of these costs, we note that in the US, the average OR charge per hour (i.e., the constant R above) is around \$3600, the OR staffing cost per hour (i.e., the constant A above) excluding physician costs is about \$450 to \$600 and this cost can be much higher when it includes physician costs (Macario, Vitz, Dunn, and McDonald 1995). In addition, overtime cost per hour (i.e., the constant P above) for staff are often 1.5 to 1.75 times the regular per hour cost due to the federally mandated overtime rate of 1.5 (Dexter Epstein, and Marsh 2001).

The expected overtime and idling costs at stage 0, given that k patients are served, is $O_0\mathbf{E}[S^0(k) - C_0]^+ + L_0\mathbf{E}[S^0(k) - C_0]^-$ for any integer k. 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 0-cost above to be defined on real-valued k by piecewise-linear extension. Similarly, we extend the overtime and idling costs at stage 1.

The events in each day occur in the following sequence:

1. At the beginning of day t, there are w_t elective patients 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. The 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 $w_t + \delta_t + n_t$.
3. The manager decides, out of the $w_t + \delta_t$ outstanding elective cases, the number q_t of elective surgeries to fulfill in day t. Immediately after the decision, the number of patients at stages 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 and 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. Idling and 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. We assume that the sequence $\{\xi_t\}_t$ is i.i.d. over time. This is similar to assuming that patient LOS in stage 1 is geometrically distributed. Litvak, van Rijsbergen, Boucherie, and van Houdenhoven (2008) have shown that patient LOS in ICUs can be modeled as an exponential random variable. Bowers (2013) has also noted that the exponential distribution provides an "approximate" fit to the LOS data at the cardiothoracic center he studies. In our discrete-time system, we use a geometric random variable, which is essentially the discrete time version of an exponential variable, to model patient LOS.

The objective of the problem is to determine a scheduling policy which minimizes the total discounted cost 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 $0 \leq q_t \leq \bar{w}_t$, since the number to be scheduled cannot exceed the number currently on the waitlist.

The system evolves as follows:

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

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

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

The single-day cost function can be written as

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

where $(\cdot)^+ = \max(\cdot, 0)$, and $(\cdot)^- = -\min(\cdot, 0)$. 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.

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:

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

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 will 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 the variables used are natural quantities to define. 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, i.e. q_t should be smaller, to avoid overtime in 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 capacity. A numerical example is shown in Figure 1, in which the optimal decision q_t^* initially decreases in n_t but then increases.

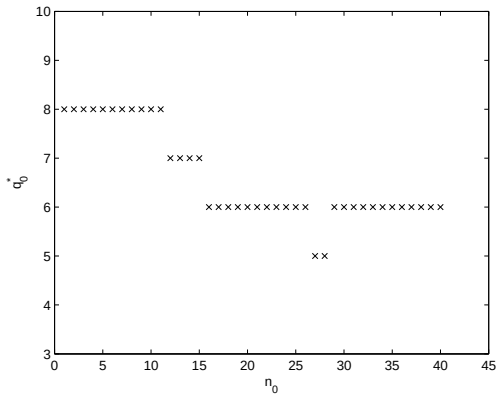

Figure 1 For a fixed waitlist size $w_0 = 8$, the optimal decision q_0^* is not monotone in the state variable n_0. The number of patients is rounded to the nearest integer. $C_0 = 7$, $C_1 = 20$, $W = 2$, $O_0 = 6$, $L_0 = 6$, $O_1 = 5$, $L_1 = 5$, $\mathbf{E}[v^0] = 1$, $\mathbf{E}[v^1] = 1$, $\gamma = 0.9$, $\mathbf{E}[\delta] = 6$, $\mathbf{E}[\epsilon] = 0.5$, T=50, $\mathbf{E}[\xi_t] = 0.7$ and ξ_t is uniformly distributed over $[0.6, 0.8]$.

To give an explanation, when the number n_t of patients downstream is small compared to the downstream capacity, it is crucial to reduce the idling cost downstream as much as possible. Thus it is optimal to pull more patients from the waitlist even if it leads to overtime costs upstream. As n_t increases, it becomes more important to reduce the overtime cost downstream by admitting fewer patients. In order reduce overtime costs downstream, it is possible that the optimal decision

will incur more idling cost upstream. However, when n_t is sufficiently large, 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 therefore more patients should be admitted to the upstream stage.

The example above shows that the relationship among the original model variables does not provide a clear direction on how to adjust decisions if the system state changes. Next, we will perform a transformation of variables that will place 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. Note that $a_t = w_t + n_t$. 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 will reformulate problem (4) with variables (a_t, m_t) replacing (w_t, q_t).

Observe that with the above transformation, the decision variable becomes the total number m_t of patients to be in the hospital by the end of period t. It is more compatible with the state variables a_t and n_t in the sense that, rather than being a rate, 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 and beyond. 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,

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

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

The single-day cost function can be written as

$$F(m_t, a_t, n_t, \delta_t) = W(a_t - n_t) + O_0\mathbf{E}[S^0(m_t - n_t + \epsilon_t) - C_0]^+ + L_0\mathbf{E}[S^0(m_t - n_t + \epsilon_t) - C_0]^-$$

$$+O_1\mathbf{E}[S^1(m_t+\epsilon_t)-C_1]^+ + L_1\mathbf{E}[S^1(m_t+\epsilon_t)-C_1]^-. \tag{7}$$

Let $V_t(a_t, n_t, \delta_t)$ denote the optimal total discounted cost incurred from day t to T when the state just before the decision m_t is made on day t is (a_t, n_t, δ_t). The Bellman equation can be written as follows:

$$\begin{aligned} V_t(a_t, n_t, \delta_t) &= \min_{n_t \le m_t \le a_t+\delta_t} G_t(m_t, a_t, n_t, \delta_t) \ , \text{ where} \\ G_t(m_t, a_t, n_t, \delta_t) &= F(m_t, a_t, n_t, \delta_t) + \gamma \mathbf{E}\left[V_{t+1}(a_{t+1}, n_{t+1}, \delta_{t+1}) | (a_t, n_t, m_t, \delta_t)\right], \\ &= F(m_t, a_t, n_t, \delta_t) + \gamma \mathbf{E}\left[V_{t+1}(a_t+\delta_t-m_t+\xi_t(m_t+\epsilon_t),\ \xi_t(m_t+\epsilon_t),\ \delta_{t+1})\right], \end{aligned} \tag{8}$$

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*. In the following section we will 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 Solutions

We now investigate the transformed formulation (8). We will 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.*

Proof. We first show the joint convexity of $F(\cdot,\cdot,\cdot,\cdot)$. The first component $W(a_t - n_t)$ is clearly convex. Because $S^i(k)$ is a sum of k i.i.d non-negative random variables, $\{S^i(k), k = 0,1,2,\dots\}$ is stochastic increasing and linear in sample path sense; or in short, $\{S^i(k), k = 0,1,2,\dots\} \in \text{SIL(sp)}$ (see Example 4.3 in Shaked and Shanthikumar (1988)). Thus $\mathbf{E}f(S^i(k))$ is convex in k if f is a convex function (see Proposition 3.7 in Shaked and Shanthikumar (1988)). Since $O_1\mathbf{E}(x + S_t^1(\epsilon_t) - C_1)^+$ is convex in x, we have that $O_1\mathbf{E}(S^1(m_t) + S_t^1(\epsilon_t) - C_0)^+$ is convex for $m_t \in \{0,1,2,\dots\}$. Thus, its piecewise-linear extension is also convex. Due to the fact that the convexity of a real function is preserved under linear transformation of its arguments (see Theorem 5.7 in Rockafellar (1972)), we know that $O_0\mathbf{E}[S^0(m_t - n_t + \epsilon_t) - C_0]^+$ is jointly convex in (m_t, n_t). The convexity of the other terms in $F(\cdot,\cdot,\cdot,\cdot)$ follow a similar argument. Since F is obtained by adding up convex terms, it must be convex.

Assuming that $V_{t+1}(\cdot,\cdot,\cdot)$ is jointly convex in its arguments (which is true for $t = T$), we know that

$$\mathbf{E}\left[V_{t+1}(a_t + \delta_t - m_t + \xi(m_t + \epsilon_t), \xi(m_t + \epsilon_t)), \delta_{t+1}\right],$$

is also convex (see Theorem 5.7 in Rockafellar (1972)). Therefore, $G_t(\cdot,\cdot,\cdot,\cdot)$ is jointly convex. Because $V_t(\cdot,\cdot,\cdot)$ is obtained by minimizing a convex function $G_t(\cdot,\cdot,\cdot,\cdot)$ in a convex feasible region, it follows that $V_t(\cdot,\cdot,\cdot)$ is also convex (see Theorem A.4. in Porteus (2002)). By induction, the theorem holds for all t. □

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

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

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 costs, 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).*

Proof. We start by proving (1). The first term $W(a_t - n_t)$ in $F(\cdot,\cdot,\cdot,\cdot)$ is linear and thus submodular. If a function $h(x)$ is convex in x, then $h(x-y)$ is submodular in x and y. It follows that the second and the third terms of $F(\cdot,\cdot,\cdot,\cdot)$ are submodular. The last two terms are trivially submodular as they are both single variable functions.

We will prove (2) and (3) by induction. Assume that $V_{t+1}(a_{t+1}, n_{t+1}, \delta_{t+1})$ is submodular in (a_{t+1}, n_{t+1}). Because of the submodularity of $F(\cdot,\cdot,,\cdot,\cdot)$ shown above, to prove (2), it remains to check the submodularity of its second term

$$V_{t+1}(a_t + \delta_t - m_t + \xi_t(m_t + \epsilon_t), \xi_t(m_t + \epsilon_t)) = V_{t+1}(a_t + \delta_t - (1-\xi_t)m_t + \xi_t\epsilon_t, \xi_t m_t + \xi_t\epsilon_t))$$

for any realization of δ_t, ξ_t and ϵ_t. We drop δ_{t+1} from the expression above for notational convenience. For any $a_t^+ \geq a_t^-$ and $m_t^+ \geq m_t^-$, we have

$$\begin{aligned}
& V_{t+1}(a_t^+ + \delta_t - (1-\xi_t)m_t^+ + \xi_t\epsilon_t, \xi_t m_t^+ + \xi_t\epsilon_t) \\
& -V_{t+1}(a_t^+ + \delta_t - (1-\xi_t)m_t^- + \xi_t\epsilon_t, \xi_t m_t^- + \xi_t\epsilon_t) \\
& -V_{t+1}(a_t^- + \delta_t - (1-\xi_t)m_t^+ + \xi_t\epsilon_t, \xi_t m_t^+ + \xi_t\epsilon_t) \\
& +V_{t+1}(a_t^- + \delta_t - (1-\xi_t)m_t^- + \xi_t\epsilon_t, \xi_t m_t^- + \xi_t\epsilon_t) \\
\leq\ & V_{t+1}(a_t^+ + \delta_t - (1-\xi_t)m_t^+ + \xi_t\epsilon_t, \xi_t m_t^- + \xi_t\epsilon_t) \\
& -V_{t+1}(a_t^+ + \delta_t - (1-\xi_t)m_t^- + \xi_t\epsilon_t, \xi_t m_t^- + \xi_t\epsilon_t) \\
& -V_{t+1}(a_t^- + \delta_t - (1-\xi_t)m_t^+ + \xi_t\epsilon_t, \xi_t m_t^- + \xi_t\epsilon_t) \\
& +V_{t+1}(a_t^- + \delta_t - (1-\xi_t)m_t^- + \xi_t\epsilon_t, \xi_t m_t^- + \xi_t\epsilon_t) \\
\leq\ & V_{t+1}(a_t^- + \delta_t - (1-\xi_t)m_t^+ + \xi_t\epsilon_t, \xi_t m_t^- + \xi_t\epsilon_t) \\
& -V_{t+1}(a_t^- + \delta_t - (1-\xi_t)m_t^- + \xi_t\epsilon_t, \xi_t m_t^- + \xi_t\epsilon_t) \\
& -V_{t+1}(a_t^- + \delta_t - (1-\xi_t)m_t^+ + \xi_t\epsilon_t, \xi_t m_t^- + \xi_t\epsilon_t)
\end{aligned}$$

$$+V_{t+1}(a_t^- + \delta_t - (1-\xi_t)m_t^- + \xi_t\epsilon_t, \xi_t m_t^- + \xi_t\epsilon_t)$$
$$= 0,$$

where the first inequality follows from the submodularity, the second from the joint convexity of $V_{t+1}(\cdot,\cdot,\cdot)$. Thus,

$$\mathbf{E}\{V_{t+1}(a_t + \delta_t - m_t + \xi_t(m_t+\epsilon_t), \xi_t(m_t+\epsilon_t), \delta_{t+1})\}$$

is submodular in (a_t, m_t), and $G_t(m_t, a_t, n_t, \delta_t)$ is submodular in (m_t, a_t, n_t).

Lastly, note that for fixed δ_t, the set $\{(m_t, a_t, n_t) \in \mathbb{R}^3_+ \mid n_t \le m_t \le a_t + \delta_t\}$ is a lattice, and $V_t(a_t, n_t, \delta_t)$ is obtained by minimizing $G_t(m_t, a_t, n_t, \delta_t)$ over all m_t such that (m_t, a_t, n_t) belongs to this set. Thus, we have the submodularity of $V_t(a_t, n_t, \delta_t)$ in (a_t, n_t), proving (3) (see Theorem 2.7.6 in Topkis (1998)). □

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.

Define the minimum and maximum optimal decision m_t in day t to be, respectively,

$$m_t^{min}(a_t, n_t, \delta_t) = \min\arg \min_{n_t \le m_t \le a_t+\delta_t} G_t(m_t, a_t, n_t, \delta_t), \text{ and} \tag{9}$$
$$m_t^{max}(a_t, n_t, \delta_t) = \max\arg \min_{n_t \le m_t \le a_t+\delta_t} G_t(m_t, a_t, n_t, \delta_t). \tag{10}$$

Then we have the following result. (In this paper, 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 number 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 state (a_t, n_t).*

Proof. This directly follows from Theorem 1 and the fact that the feasible region $\{(m_t, a_t, n_t) \in \mathbb{R}^3_+ \mid n_t \le m_t \le a_t + \delta_t\}$ is a lattice (see Theorem 2.8.2 in Topkis (1998)). □

The monotonicity of the optimal decision in the 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 size of three nested sets of patients. These sets have a correspondence in size under an optimal policy. As the smallest set containing n_t recovering patients increases in size, the middle set containing m_t admitted and recovering patients increases. Similarly, 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.

6. Relationship to Capacity

So far we have assumed that the capacity vector is fixed at both stages. 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.

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, $i = 1, 2$.

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. However, as our goal is to obtain structural insights on how the capacity at the two stages affects overall costs and affects the optimal policies,

we can assume zero capacity expansion cost because adding a linear capacity cost to our formulation will not change its structural properties.

With these modifications, the single-day cost function becomes

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

and the value function V_t remains the same as defined in (8) except that C_i's are now included as new 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.*

Proof. It is easy to check the single-day function defined in (11) is jointly convex in its arguments, and the rest of the proof is similar to that of Proposition 1. □

The theorem above suggests a diminishing return as the capacity upstream or downstream increases, a trend that has been confirmed by simulation in Bowers (2013). With greater investments in capacity, we eventually experience lower marginal returns because the population of patients as reflected in the demand distribution is fixed. 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, however, does not seem to suggest otherwise, and our numerical experiments below indeed confirm our intuition.

Figures 2 and 3 illustrate the convexity of the value function $V_i(C, a_t, n_t)$ in the capacities C_0 and C_1, respectively, when capacity use of different patients are i.i.d. That is, these figures are plotted for the original model presented in Section 4. These two figures use the same set of parameters, except that Figure 2 shows how V_t changes in upstream capacity C_0 with downstream capacity C_1 fixed, while Figure 3 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 or the planning horizon increases.

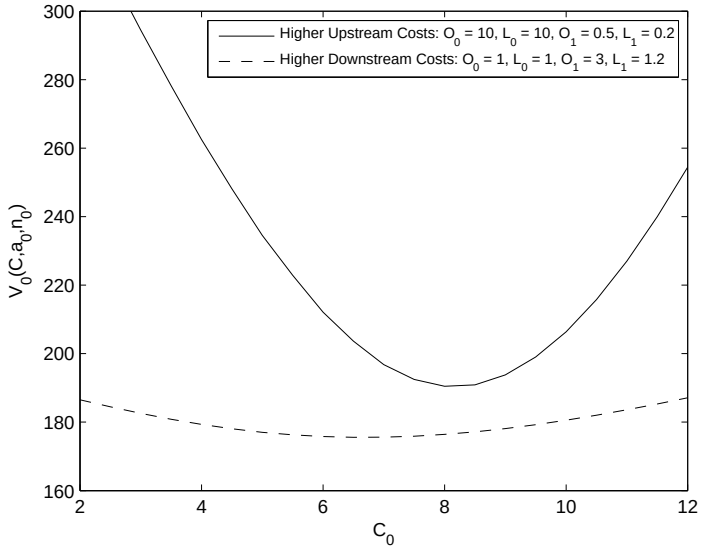

Figure 2 Total expected cost $V_0(C, a_0, n_0)$ as a convex function of C_0. $a_0 = 18$, $n_0 = 13$, $W = 1$, $C_1 = 30$, $\mathbf{E}[v^0] = 1$, $\mathbf{E}[v^1] = 1$, $\gamma = 0.8$, $\mathbf{E}[\delta] = 8$, $\mathbf{E}[\epsilon] = 1.5$, T=90, $\mathbf{E}[\xi_t] = 0.85$ and ξ_t is uniformly distributed over $[0.7, 1]$.

In Figure 2, the downstream capacity is $C_1 = 30$. Each patient is expected to stay $1/(1-\mathbf{E}[\xi_t]) = 6.67$ days in the system, and therefore, on average, the downstream stage can process $C_1(1-\mathbf{E}[\xi]) = 4.5$ patients each day, which is much smaller than the expected daily arrival rate of $\mathbf{E}[\delta]+\mathbf{E}[\epsilon] = 9.5$. 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 $C_1(1-\mathbf{E}[\xi]) = 4.5$, as the throughput

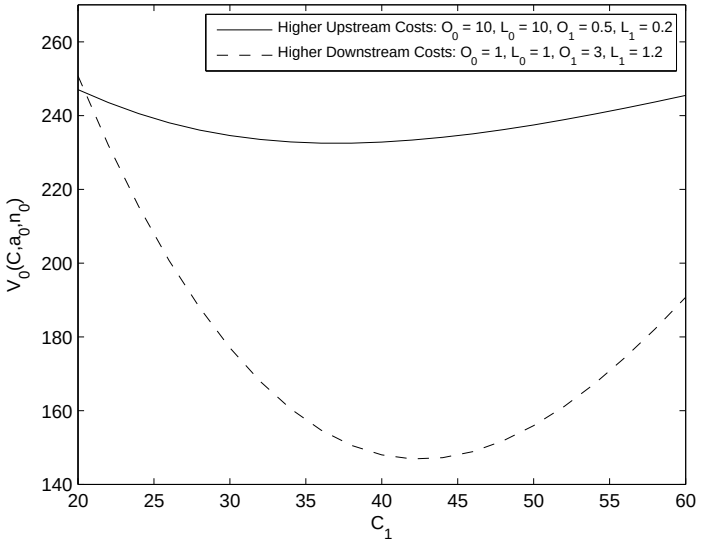

Figure 3 Total expected cost $V_0(C, a_0, n_0)$ as a convex function of C_1. $a_0 = 18$, $n_0 = 13$, $W = 1$, $C_0 = 5$, $\mathbf{E}[v^0] = 1$, $\mathbf{E}[v^1] = 1$, $\gamma = 0.8$, $\mathbf{E}[\delta] = 8$, $\mathbf{E}[\epsilon] = 1.5$, T=90, $\mathbf{E}[\xi_t] = 0.85$ and ξ_t is uniformly distributed over $[0.7, 1]$.

is primarily determined by the downstream capacity.

In Figure 3, the capacity upstream is $C_0 = 5$, which is smaller than the average daily patient demand rate of $\mathbf{E}[\delta] + \mathbf{E}[\epsilon] = 9.5$. The optimal downstream capacity C_1 again is sensitive to whether upstream or downstream cost dominates. 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 $v^i k$ for $i = 1, 2$, where v^i is a non-negative random

variable. Then, we can prove that the formulation is submodular with respect to the capacity at stage 1, i.e., the downstream stage. The proof is similar to that of Theorem 1, and thus we omit it here.

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 of 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.

A similar result, however, 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 dual effect on the total number of patients in the hospital at t. On the one hand, more patient may 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 tends to decrease the total number of patients in the hospital at t.

The numerical example in Figure 4 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 knows that it could process a patient within 3 days of his arrival without using surge capacity, it would choose to keep the patient on waitlist to avoid incurring overtime cost 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 to admit 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 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 overtime cost now. However, as C_0 increases, the emphasis of the optimal scheduling policy changes from preventing overtime cost to reducing idling cost in the upstream stage, and as a result it may admit more patients from the waitlist.

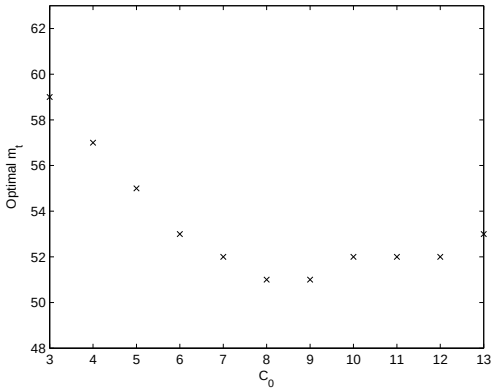

Figure 4 The optimal decisions $m_t^{max}(C, a_t, n_t)$ and $m_t^{min}(C, a_t, n_t)$ are not monotonic in C_0. $a_t = 70$, $n_t = 40$, $W = 1$, $O_0 = 3$, $L_0 = 3$, $O_1 = 1$, $L_1 = 1$, $C_1 = 48$, $\mathbf{E}[v^0] = 1$, $\mathbf{E}[v^1] = 1$, $\gamma = 0.9$, $\mathbf{E}[\delta] = 6$, $\mathbf{E}[\epsilon] = 0.5$, T=50, $\mathbf{E}[\xi_t] = 0.85$ and ξ_t is uniformly distributed over $[0.7, 1]$.

7. Growth of the Waitlist

When we consider an infinite horizon, the system may grow out of control (i.e., the waitlist will increase without bounds) under certain conditions, although the total cost remains bounded because of the discounting effect. From classical queuing results, we know that when the average patient demand exceeds the daily processing capacities, the size of the waitlist will grow over time and the system will blow up eventually. However, stability in our system requires more subtle considerations. The optimal number of patients admitted everyday, and thus the system throughput rate are endogenous in our case. In particular, the optimal scheduling policy balances patient waiting costs and service costs at both stages. Therefore, if patient waiting costs are relatively low but service costs are high, the system would admit fewer patients per day, keeping more of them in the waitlist and in this way, become more likely to explode.

In this section, we will study the growth of waitlist and the stability of the system. We will show that when the waiting cost is sufficiently small and the patient demand significantly high, the waitlist is expected to grow without bounds. We will characterize sufficient conditions under which the system size is stable, i.e., $\{a_t\}$ being finite in an infinite horizon. Though we rarely see an exploding system in reality, our analysis here generates useful insights on system dynamics. In addition, it can inform the choice of model parameters for decisions makers so that they can be sure that the waitlist will be cleared eventually.

Define

$$W^* = \frac{1-\gamma}{\gamma}\left[\mathbf{E}[v^0]O_0 + \frac{\mathbf{E}[v^1]O_1}{1-\mathbf{E}[\xi_t]\gamma}\right] \tag{12}$$

to be our *critical waiting cost*. Note that W^* is a function of the capacity required by patients, their tendency to persist in the system as reflected in the distribution of ξ, the overtime costs, and the discount factor γ. We first show that when the waiting cost is below W^*, the optimal maximum number of regular admissions q_t^{max} in each period is uniformly bounded above for all system states. Therefore the optimal decisions $m_t^{min}(a_t, n_t, \delta_t)$ and $m_t^{max}(a_t, n_t, \delta_t)$ approach finite limits as the number of patients in the system, a_t, approaches infinity.

PROPOSITION 2. *For fixed C, $T=\infty$ and $W<W^*$,*

1. *The optimal number of admissions $q_t^{max}=q_t^{max}(n_t,w_t,\delta_t)$ is uniformly bounded above for any (n_t,w_t,δ_t);*
2. *For each fixed n_t, there are finite limiting functions for $\tilde{m}^{max}(n_t)$ and $\tilde{m}^{min}(n_t)$ such that*

$$\lim_{a_t\to+\infty} m^{min}(a_t,n_t,\delta_t)=\tilde{m}^{min}(n_t), \quad \textit{and}$$

$$\lim_{a_t\to+\infty} m^{max}(a_t,n_t,\delta_t)=\tilde{m}^{max}(n_t).$$

Proof. The total cost of having a new patient stay in the waitlist forever (discounted from the current period) is

$$c_1=\gamma W+\gamma^2 W+\cdots=\frac{\gamma W}{1-\gamma}. \tag{13}$$

After we schedule a patient to surgery, the expected downstream resource consumed by this patient on the k-th day after the admission is

$$\begin{aligned}&\mathbf{E}[v^1\mathbf{1}\{\text{The patient has a length-of-stay of at least } k \text{ days}\}]\\ =\ &\mathbf{E}[v^1]P(\text{The patient has a length-of-stay of at least } k \text{ days})\\ =\ &\mathbf{E}[v^1]\mathbf{E}[\xi_t]^k.\end{aligned} \tag{14}$$

The worst case of admitting an additional patient happens when the patient consumes only overtime resources. This patient incurs a discounted total cost of

$$c_2=O_0\mathbf{E}[v^0]+\sum_{k=0}^{\infty}\gamma^k O_1\mathbf{E}[v^1]\mathbf{E}[\xi_t]^k=O_0\mathbf{E}[v^0]+O_1\frac{\mathbf{E}[v^1]}{1-\gamma\mathbf{E}[\xi_t]}, \tag{15}$$

where $\mathbf{E}[v^1]\mathbf{E}[\xi_t]^k$ is the expected usage of downstream resource on the kth day after the admission as given by (14).

If $W<W^*$, then $c_1<c_2$. To prove the first statement, note that after sufficiently many patients are admitted in one day, the marginal cost of admitting an additional patient is arbitrarily close to c_2. To be more specific, fix $\epsilon\in(0,c_2-c_1)$, we can always find a threshold independent of (n_t,w_t,δ_t) beyond which if an additional patient is admitted, this patient will lead to an increase in overall

overtime costs of no smaller than $c_2 - \epsilon > c_1$. In this case, keeping this additional patient in the waitlist forever will lead to a strictly smaller overall cost, implying that the optimal policy would not admit this patient at t, i.e., q_t^{max} should not exceed this threshold.

To prove the second statement, note that $m^{max}(a_t, n_t, \delta_t)$ and $m^{min}(a_t, n_t, \delta_t)$ are monotonic functions of a_t as proved in Corollary 1. Since the optimal value of $q_t = m_t - n_t$ is uniformly bounded in (n_t, w_t, δ_t), the limits $\tilde{m}^{max}(n_t)$ and $\tilde{m}^{min}(n_t)$ exist and are finite for any given n_t. □

The limiting function $\tilde{m}^{max}(n_t)$ defined in Proposition 2 is the optimal decision when there are infinitely many patients on the waitlist and $W < W^*$. Consider a scenario in which a_t is very large (the waitlist has blown up). If we apply $\tilde{m}^{max}(n_t)$ in this situation, the process $\{n_t\}$ will satisfy $n_{t+1} \approx \xi_t(\tilde{m}^{max}(n_t) + \epsilon_t)$. Then the process $\{n_t\}$ approximates a Markov chain, for ξ_t and ϵ_t are both exogenous random variables. We can then define the average number of elective admissions per period as

$$\mu = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} (\tilde{m}^{max}(n_t) - n_t), \quad \text{where}$$
$$n_{t+1} = \xi_t(\tilde{m}^{max}(n_t) + \epsilon_t).$$

For any system with $W < W^*$, μ as defined above can be seen as the *limiting service rate of the system*. Note that this may not be the actual rate in which elective patients are admitted, but rather a rate in a hypothetical limiting state where the waitlist always approaches infinity. Therefore, if $\mathbf{E}[\delta_t] > \mu$, the system size a_t will go to infinity in expectation because the limiting departure rate μ is less than the arrival rate. Conversely, if $\mathbf{E}[\delta_t] < \mu$, the system is stable.

As W becomes larger than W^*, our intuition suggests that the system becomes more likely to be stable. It is relatively easy to check if the waitlist explodes via numerical computation, but it is difficult to establish a general condition under which the waitlist is stable. This is because the transition probabilities governing the evolution of the system under the optimal scheduling policy are not explicitly available, making it difficult to check the positive recurrence of the system. Nevertheless, we are able to show one sufficient condition for system stability in line with our intuition.

LEMMA 1. *When $W \geq \frac{W^*}{1-\gamma}$, it is optimal to admit anyone upon his arrival, and thus the waitlist is stable.*

Proof. Note that $W \geq \frac{W^*}{1-\gamma}$ implies $\gamma W \geq c_2$, where c_2 is defined in (15). Recall that c_2 is the highest cost that a patient can incur to the system after being admitted. Consider any time t and any system state at t, keeping one new patient if any in the waitlist at t incurs a waiting cost of γW in period $t+1$, but admitting this patient right away incurs an expected overall cost no larger than c_2. Because $\gamma W \geq c_2$, an optimal policy would not keep any new patient waiting, proving the desired results. □

To further explore these technical results, we numerically investigate how the limiting service rate μ changes as other model parameters change. Figure 5 shows μ as a function of the waiting cost W for different values of the discount factor γ. For each fixed γ, the figure shows that the system tends to admit more patients as W increases, and the limiting service rate approaches infinity as W approaches W^* (the vertical lines), suggesting that the system would be stable when W is large enough. We also observe that the higher the discount factor γ is, the larger the limiting service rate is, and the more sensitive it is to the changes in the waiting cost W. This is because a larger γ leads to higher costs for keeping a waitlist, and thus gives greater incentive to serve patients rather than to make them wait.

Figures 6 and 7 show the limiting service rate μ as a function of the downstream capacity C_1. We find that generally μ increases in C_1 as expected. Moreover, a larger idling cost leads to higher utilization of the system, and thus a higher value of μ. In determining μ, the idling cost regulates two different tradeoffs: (1) the tradeoff against overtime cost in the same stage; and (2) the tradeoff against overtime cost in the other stage.

The second tradeoff becomes more dominant when the capacity gap (the difference between C_0 and $(1-\mathbf{E}[\xi_t]) \cdot C_1$) is larger. This explains the deviation between the two lines in Figures 6 and 7. Consider the region $C_1 \cdot (1-\mathbf{E}[\xi_t]) < 8 = C_0$ in Figure 6. This is the region where capacity downstream is relatively smaller than that upstream. Since the downstream stage is the bottleneck,

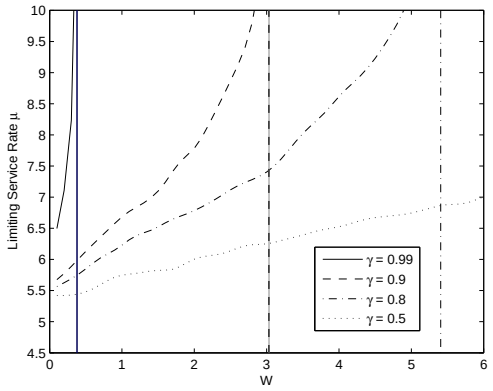

Figure 5 Limiting service rate μ as a function of W, with vertical lines specifying the values of W^*. $O_0 = 6$, $O_1 = 5$, $L_0 = 2$, $L_1 = 1$, $C_0 = 8$, $C_1 = 45$, $\mathbf{E}[v^0] = 1$, $\mathbf{E}[v^1] = 1$, $\mathbf{E}[\epsilon] = 1.5$, $\mathbf{E}[\xi_t] = 0.85$ and ξ_t is uniformly distributed over $[0.7, 1]$. Note that μ is independent of δ_t.

admitting an additional patient will lead to almost the same amount of overtime cost in this stage, but may decrease the idling cost in upstream stage quite substantially, especially when the idling cost there is high. Therefore, idling cost upstream when downstream capacity is tight has a significant impact on the optimal scheduling policy, and thus on the limiting service rate. Similarly, consider the region $C_1 \cdot (1 - \mathbf{E}[\xi_t]) > 8 = C_0$ in Figure 7 where the capacity upstream is tight. Here, the impact of the idling cost downstream becomes important.

8. Numerical Studies

As discussed earlier, previous Healthcare Operations Management literature has either focused on dynamic scheduling decisions for a single stage (e.g., Gerchak, Gupta and Henig 1996), or considered systems design under static scheduling rules in a facility with multiple units (e.g., Helm and Van Oyen 2012). Little is known about how to dynamically schedule patients taking into account downstream capacity and patient census. Our paper is the first to analytically study such integrated decision making. Our numerical experiments in this section follow the theoretical work above to investigate the performance of our proposed scheduling method and compare it with scheduling policies that make decisions independently of operations in other units. This comparison reveals the value of *integrated scheduling*.

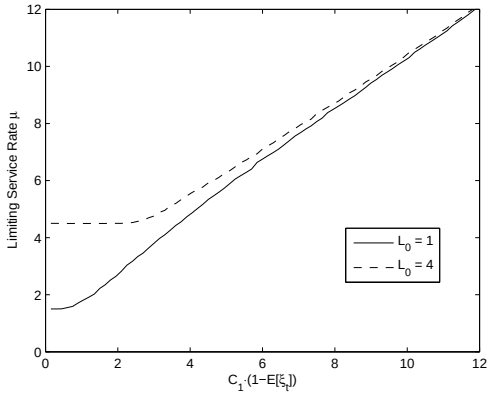

Figure 6 Limiting service rate μ as a function of downstream capacity for different values of upstream idling cost L_0. $W = 3.2$, $O_0 = 6$, $O_1 = 5$, $L_1 = 1$, $C_0 = 8$, $\mathbf{E}[v^0] = 1$, $\mathbf{E}[v^1] = 1$, $\gamma = 0.8$, $\mathbf{E}[\epsilon] = 1.5$, $\mathbf{E}[\xi_t] = 0.85$ and ξ_t is uniformly distributed over $[0.7, 1]$.

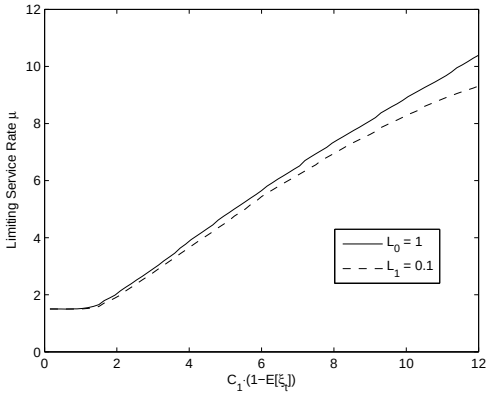

Figure 7 Limiting service rate μ as a function of downstream capacity for different values of downstream idling cost L_1. $W = 1$, $O_0 = 6$, $L_0 = 2$, $O_1 = 5$, $C_0 = 8$, $\mathbf{E}[v^0] = 1$, $\mathbf{E}[v^1] = 1$, $\gamma = 0.8$, $\mathbf{E}[\epsilon] = 1.5$, $\mathbf{E}[\xi_t] = 0.85$ and ξ_t is uniformly distributed over $[0.7, 1]$.

To study the value of integrated scheduling, we compare the numerical performances of the optimal policy Π^* against two single-stage policies Π_i, $i = 0, 1$. Each policy Π_i treats stage i as the only bottleneck resource and assumes that there is infinite capacity at the other stage. A policy Π_i 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 Π^* and Π_i indicates a higher value of integrated scheduling.

To make our comparison realistic, we adopt data from Bowers (2013) in our experiments. In that paper, the author simulates the scheduling system in the Heart and Lung Centre (HLC) at the Golden Jubilee National Hospital in Scotland. The paper generates the optimal staffing levels for the ICU (stage 1), but has not compared the performance of different scheduling heuristics. We use data from Bowers (2013) in the following way.

Upstream and Downstream Capacity. In Bowers (2013), the upstream capacity is 30 surgeries per week and about 12% of the cases are canceled. To model the "effective" capacity level that needs to satisfy patient demand, we set the upstream capacity in our experiments to be $C_0 = 30 \cdot 88\%/7 = 3.77$ per day. In the HLC the average bed occupancy in ICU is 14.9 per day. In our experiment,

we vary C_1 from 9 to 21.

Arrival Rates. Since the HLC has a growing waitlist, we set the total patient arrival rate to be equal or slightly higher than the surgery capacity. We assume the arrival rate for emergency patients $\mathbf{E}[\epsilon_t] = 0.2$, and consider two possible elective arrival rates for $\mathbf{E}[\delta_t]$. First, we assume a balanced system in which $\mathbf{E}[\delta_t] = 3.57 = C_0 - \mathbf{E}[\epsilon_t]$; then, we consider an overloaded system where the demand is on average 10% more than the surgical capacity: $\mathbf{E}[\delta_t] = 3.94 = 110\% \cdot C_0 - \mathbf{E}[\epsilon_t]$.

Resource Usage and LOS. Bowers (2013) assumes that each patient takes a constant unit of resource per day. In our experiment we let v^0 and v^1 to be exponentially distributed with mean 1. The average LOS in the HLC is 3.69 days. Accordingly, we let ξ_t be uniformly distributed in $[0.63, 0.83]$ for every period t, so that the mean LOS is $1/(1 - \mathbf{E}[\xi_t]) = 3.69$.

Cost Rates. We normalize the waiting cost rate to $W = 1$, and set the overtime cost rates to be $O_0 = 5, 10$, and $O_1 = 5$. We experiment with different idling cost rates in two stages by choosing $(L_0, L_1) \in \{(1,1), (0.8, 1.6), (1.6, 0.8)\}$.

Planning Horizon and Other Parameters. We let the planning horizon be $T = 50$, set the discount factor $\gamma = 0.8$, and assume the initial system size is $a_0 = 12$ patients and among them $n_0 = 8$ patients stay in the downstream stage at the beginning.

For each combination of experimental parameters, we calculate the total discounted costs V^{Π_0}, V^{Π_1} and V^{Π^*} under policies Π_0, Π_1 and Π^*, respectively. For easy comparison, we report results as the performance ratio with respect to V^{Π^*}, the total cost under the optimal policy Π^*. We summarize the ratios of V^{Π_0}/V^{Π^*}, V^{Π_1}/V^{Π^*}, and $\min\{V^{\Pi_0}, V^{\Pi_1}\}/V^{\Pi^*}$ in Tables 1 to 3. These cost ratios indicate the value of integrated scheduling. Higher ratios correspond to more cost saving when integrated scheduling is used.

We see that when we misidentify the bottleneck stage, that is, when we use the worse of Π_0 and Π_1, 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 cost by more than 50%. However, even when we correctly identify the bottleneck stage, that is, when

we select the better policy between Π_0 and Π_1, integrated scheduling may still help reduce cost up to 17%.

More specifically, the cost ratio V^{Π_0}/V^{Π^*} represents the performance of a system admitting patients according to a surgical scheduler who ignores the operations in the downstream ICU stage. We see that the largest ratio is 159%, suggesting that the system overspends by almost 60% when not taking downstream stage into account. More importantly, in one third of the scenarios tested, the system overspends by more than 20%. However, as the downstream capacity C_1 increases, we observe that this ratio consistently decreases across the three tables. This is because that stage 0 becomes the bottleneck stage as C_1 gets larger, making Π_0 behave similarly to Π^* and closing the performance gaps of these two policies. Comparing Tables 1 and 2, we also see that Π_0 in general performs better when the overtime cost upstream O_0 is larger. The explanation is that when the upstream cost becomes more significant, a policy that aims to minimize the cost upstream is likely to perform well.

The cost ratio of V^{Π_1}/V^{Π^*} indicates the system performance following the decisions of an ICU manager who does not pay attention to the OR usage. Our experiments show that the system could overspend 2%-18% of the optimal cost without considering the stage upstream. When C_1 is relatively small (e.g., 9 or 10), the stage downstream is in fact the bottleneck. Policy Π_1 behaves similarly to Π^*, and thus the total discounted cost under Π_1 is only a few percentage points higher than the optimal cost under Π^*. As C_1 becomes larger, the value of dynamically balancing the usage of upstream and downstream capacity becomes more significant, and as a result we see that the performance gap between Π_1 and Π^* increases as C_1 increases. Interestingly, however, we find that after C_1 is raised to a certain level ($C_1 = 13$ or 14 in Tables 1 and 2, and $C_1 = 16$ in Table 3), the performance of Π_1 and Π^* starts to converge again as C_1 increases, indicating that a policy which treats downstream as the only bottleneck performs quite well when the downstream capacity is large. This may seem counter-intuitive at first, but a close look at our experiment setup reveals that the upstream capacity C_0, which is fixed at an equal or a slightly lower level compared to the overall demand, would not be a significant bottleneck for patient flow in general.

Table 1 Cost Ratio Under Different Policies ($\mathbf{E}[\delta_t] = 3.57$, $O_0 = 5$)

	V^{Π_0}/V^{Π^*}			V^{Π_1}/V^{Π^*}			$\min\{V^{\Pi_0}, V^{\Pi_1}\}/V^{\Pi^*}$		
C_1	$L_0=1$ $L_1=1$	$L_0=1.6$ $L_1=0.8$	$L_0=0.8$ $L_1=1.6$	$L_0=1$ $L_1=1$	$L_0=1.6$ $L_1=0.8$	$L_0=0.8$ $L_1=1.6$	$L_0=1$ $L_1=1$	$L_0=1.6$ $L_1=0.8$	$L_0=0.8$ $L_1=1.6$
9	1.5925	1.4789	1.5403	1.0193	1.0138	1.0099	1.0193	1.0138	1.0099
10	1.4813	1.3939	1.4360	1.0301	1.0287	1.0215	1.0301	1.0287	1.0215
11	1.3806	1.3168	1.3429	1.0522	1.0382	1.0279	1.0522	1.0382	1.0279
12	1.2853	1.2459	1.2606	1.0444	1.0510	1.0360	1.0444	1.0510	1.0360
13	1.2117	1.1899	1.1954	1.0854	1.0674	1.0462	1.0854	1.0674	1.0462
14	1.1486	1.1412	1.1406	1.0645	1.0827	1.0548	1.0645	1.0827	1.0548
15	1.0973	1.0998	1.0974	1.0519	1.0677	1.0409	1.0519	1.0677	1.0409
16	1.0635	1.0682	1.0647	1.0551	1.0693	1.0424	1.0551	1.0682	1.0424
17	1.0409	1.0448	1.0415	1.0549	1.0683	1.0374	1.0409	1.0448	1.0374
18	1.0254	1.0285	1.0258	1.0449	1.0655	1.0334	1.0254	1.0285	1.0258
19	1.0153	1.0176	1.0156	1.0411	1.0591	1.0283	1.0153	1.0176	1.0156
20	1.0090	1.0105	1.0091	1.0348	1.0480	1.0208	1.0090	1.0105	1.0091
21	1.0051	1.0060	1.0052	1.0298	1.0411	1.0165	1.0051	1.0060	1.0052

Note that the capacity levels for C_1 above are "balanced" levels to match capacity upstream, e.g., $C_0/(1-\mathbf{E}(\xi)) = 3.77 \times 3.69 = 13.9$ in Tables 1 and 2. As the downstream capacity C_1 passes beyond these levels, both stages have relatively sufficient capacity for handling patient demand, and thus Π_1 would not behave too differently from Π^*. However, as C_1 increases, the idling cost of downstream carries increasingly more weights in the total cost, which cannot be offset much by integrated scheduling. Therefore, at very high level of C_1, the cost of Π_1 is close to that of Π^* because idling cost downstream dominates. Indeed, we see that the performance gap is smaller in Table 1 than Tables 2 and 3 as idling cost downstream is relatively higher compared to other costs in Table 1.

In summary, policies Π_0 or Π_1, which make scheduling decisions based on capacity usage at only one stage in the system, can lead to significant efficiency and financial loss. However they might perform well when (1) the bottleneck stage is correctly identified and has much smaller capacity than the other stage; and (2) the cost of the bottleneck stage is significantly higher than that of the other stage. Integrated decision making, conversely, can bring significant values to the system as a whole. It is most beneficial when there are no clear bottleneck stages in the system, This is likely due to that when one of the stages is clearly short of capacity, there is no much room for integrated scheduling to make a difference. By identifying the bottleneck stage correctly,

Table 2 Cost Ratio Under Different Policies ($\mathbf{E}[\delta_t] = 3.57$, $O_0 = 10$)

	V^{Π_0}/V^{Π^*}			V^{Π_1}/V^{Π^*}			$\min\{V^{\Pi_0}, V^{\Pi_1}\}/V^{\Pi^*}$		
C_1	$L_0 = 1$ $L_1 = 1$	$L_0 = 1.6$ $L_1 = 0.8$	$L_0 = 0.8$ $L_1 = 1.6$	$L_0 = 1$ $L_1 = 1$	$L_0 = 1.6$ $L_1 = 0.8$	$L_0 = 0.8$ $L_1 = 1.6$	$L_0 = 1$ $L_1 = 1$	$L_0 = 1.6$ $L_1 = 0.8$	$L_0 = 0.8$ $L_1 = 1.6$
9	1.5670	1.4808	1.5605	1.0465	1.0344	1.0263	1.0465	1.0344	1.0263
10	1.4547	1.3924	1.4499	1.0731	1.0670	1.0535	1.0731	1.0670	1.0535
11	1.3482	1.3095	1.3507	1.1173	1.0890	1.0714	1.1173	1.0890	1.0714
12	1.2518	1.2275	1.2572	1.1132	1.1091	1.0859	1.1132	1.1091	1.0859
13	1.1721	1.1692	1.1883	1.1829	1.1466	1.1140	1.1721	1.1466	1.1140
14	1.1189	1.1213	1.1319	1.1627	1.1777	1.1336	1.1189	1.1213	1.1319
15	1.0805	1.0848	1.0879	1.1546	1.1623	1.1121	1.0805	1.0848	1.0879
16	1.0508	1.0565	1.0554	1.1562	1.1668	1.1134	1.0508	1.0565	1.0554
17	1.0292	1.0347	1.0330	1.1479	1.1633	1.1032	1.0292	1.0347	1.0330
18	1.0154	1.0196	1.0190	1.1268	1.1559	1.0945	1.0154	1.0196	1.0190
19	1.0073	1.0102	1.0108	1.1167	1.1439	1.0847	1.0073	1.0102	1.0108
20	1.0032	1.0049	1.0063	1.1047	1.1263	1.0712	1.0032	1.0049	1.0063
21	1.0013	1.0023	1.0037	1.0952	1.1151	1.0625	1.0013	1.0023	1.0037

Table 3 Cost Ratio Under Different Policies ($\mathbf{E}[\delta_t] = 3.94$, $O_0 = 10$)

	V^{Π_0}/V^{Π^*}			V^{Π_1}/V^{Π^*}			$\min\{V^{\Pi_0}, V^{\Pi_1}\}/V^{\Pi^*}$		
C_1	$L_0 = 1$ $L_1 = 1$	$L_0 = 1.6$ $L_1 = 0.8$	$L_0 = 0.8$ $L_1 = 1.6$	$L_0 = 1$ $L_1 = 1$	$L_0 = 1.6$ $L_1 = 0.8$	$L_0 = 0.8$ $L_1 = 1.6$	$L_0 = 1$ $L_1 = 1$	$L_0 = 1.6$ $L_1 = 0.8$	$L_0 = 0.8$ $L_1 = 1.6$
9	1.5918	1.5245	1.6551	1.0380	1.0324	1.0247	1.0380	1.0324	1.0247
10	1.4877	1.4354	1.5441	1.0689	1.0633	1.0504	1.0689	1.0633	1.0504
11	1.3866	1.3494	1.4402	1.0913	1.0849	1.0679	1.0913	1.0849	1.0679
12	1.2886	1.2633	1.3393	1.1184	1.1056	1.0831	1.1184	1.1056	1.0831
13	1.2135	1.1989	1.2594	1.1509	1.1472	1.1147	1.1509	1.1472	1.1147
14	1.1524	1.1444	1.1904	1.1660	1.1854	1.1399	1.1524	1.1444	1.1399
15	1.1048	1.1017	1.1331	1.1651	1.1925	1.1355	1.1048	1.1017	1.1331
16	1.0674	1.0674	1.0883	1.1780	1.2051	1.1386	1.0674	1.0674	1.0883
17	1.0392	1.0404	1.0558	1.1695	1.1897	1.1222	1.0392	1.0404	1.0558
18	1.0203	1.0219	1.0346	1.1644	1.1834	1.1136	1.0203	1.0219	1.0346
19	1.0093	1.0104	1.0214	1.1532	1.1704	1.1028	1.0093	1.0104	1.0214
20	1.0038	1.0043	1.0131	1.1333	1.1570	1.0914	1.0038	1.0043	1.0131
21	1.0014	1.0016	1.0080	1.1205	1.1396	1.0774	1.0014	1.0016	1.0080

simple policies like Π_0 or Π_1 can already perform reasonably well. But when the capacities of both stages are relatively "balanced," a dynamic scheduling policy can make real-time adjustment to use capacity more efficiently.

9. Conclusions

This paper introduces a centralized scheduling decision model based on capacity usage in different units of a hospital. Using this model enables hospitals to significantly improve their operational efficiency compared to using localized decision rules that only focus on a single location of the hospital. Although the efficient use of healthcare capacity has long been considered as an important

issue, changes in the US practice environment after the recent health care reform have made the decision problem we study here even more timely and crucial. First, providers are under tremendous pressure from payers to improve efficiency and reduce cost due to the ongoing payment reform that shifts from fee-for-service to pay-for-performance. Second, with quick adoption of health information technology spurred by the Health Information Technology for Economic and Clinical Health (HITECH) Act enacted in 2009, more hospital operational data are being collected, making it easier to use decision models like ours. Third, health care organizations are becoming more cognizant of the power of data analytics, and thus there are fewer barriers to implement data-driven operations management approaches than before (Kamath, Osborn, Roger, and Rohleder 2011).

Our work makes an important contribution by introducing the first dynamic multi-day scheduling model that integrates information about capacity usage at more than one stage in hospitals. In particular, we analyze the first dynamic model that accounts for patient LOS and downstream census in scheduling decisions. We develop effective scheduling methods that provide intuitive insights for practitioners. Through numerical experiments on realistic data from practice, we show that there is a substantial value in making integrated scheduling decisions as suggested by our models.

Our modeling framework is quite flexible, and can be extended in a variety of ways to accommodate more details in reality. Specifically, all of our structural results remain valid under a number of extensions:

Multiple Resources. There are multiple resources used in each stage, each with its own overtime and idling cost. In this case, the total overtime and idling cost is the sum of these costs due to use of each individual resource. For example, upstream resources may include those that are essential for performing surgeries, such as ORs, nurses, surgeons and anesthesiologists.

Non-stationary Environment. The exogenous random variables in our models, e.g., elective and emergency demand, can be independent but not necessarily identically distributed. This extension is useful when strong seasonality, e.g., seasonal demand pattern, is observed in the environment.

General Convex Cost Function. We can also use any convex increasing function for the overtime and idling costs instead of assuming constant per unit overtime cost and idling cost.

Heterogeneous Resource Usage. Emergency patients might have a different distribution of capacity usage compared to elective patients. Since emergency patients are always admitted on the day of arrival, the number of emergency patients in each stage at any time is independent of the policy used. Thus, these numbers can be considered as exogenous random variables that reduce the capacity at the stages by random amounts. In this way, all of our structural results remain the same. As a special case, we can also consider the demand process as coming from elective patients alone. This simplified model is especially applicable to elective surgical centers which only accept elective patients or those facilities with only a small amount of emergency cases.

In summary, our model is useful to direct scheduling decisions in a system with two sequential service stages or a system with multiple service stages among which two are the obvious bottlenecks. Our work provides a stepping stone to study scheduling decisions in more complex settings. Future research could focus on determining whether our results extend to systems with more than two stages and multiple patient classes with different urgency for care, and advance scheduling decisions that assign patients directly to future days.

References

Ayvaz, N., W.T. Huh. 2010. Allocation of hospital capacity to multiple types of patients. *Journal of Revenue & Pricing Management* **9**(5) 386–398.

Blumenthal, David. 2009. Stimulating the adoption of health information technology. *New England Journal of Medicine* **360**(15) 1477–1479.

Bowers, John. 2013. Balancing operating theatre and bed capacity in a cardiothoracic centre. *Health care management science* 1–9.

Bretthauer, Kurt M, H Sebastian Heese, Hubert Pun, Edwin Coe. 2011. Blocking in healthcare operations: A new heuristic and an application. *Production and Operations Management* **20**(3) 375–391.

Cardoen, Brecht, Erik Demeulemeester, Jeroen Beliën. 2010. Operating room planning and scheduling: A literature review. *European Journal of Operational Research* **201**(3) 921–932.

Chan, CW, G Yom-Tov, G Escobar. 2011. When to use speedup: an examination of intensive care units with readmissions. Tech. rep., Working paper, Columbia University.

Cochran, Jeffery K, Aseem Bharti. 2006. Stochastic bed balancing of an obstetrics hospital. *Health care management science* **9**(1) 31–45.

Cohen, Morris A, John C Hershey, Elliott N Weiss. 1980. Analysis of capacity decisions for progressive patient care hospital facilities. *Health Services Research* **15**(2) 145.

Dexter, Franklin, Richard H Epstein, H Michael Marsh. 2001. A statistical analysis of weekday operating room anesthesia group staffing costs at nine independently managed surgical suites. *Anesthesia & Analgesia* **92**(6) 1493–1498.

El-Darzi, E, C Vasilakis, T Chaussalet, PH Millard. 1998. A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. *Health Care Management Science* **1**(2) 143–149.

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

Griffin, Jacqueline, Shuangjun Xia, Siyang Peng, Pinar Keskinocak. 2012. Improving patient flow in an obstetric unit. *Health care management science* **15**(1) 1–14.

Guerriero, Francesca, Rosita Guido. 2011. Operational research in the management of the operating theatre: a survey. *Health care management science* **14**(1) 89–114.

Gupta, D. 2007. Surgical suites' operations management. *Production and Operations Management* **16**(6) 689–700.

Harper, Paul R. 2002. A framework for operational modelling of hospital resources. *Health care management science* **5**(3) 165–173.

Helm, Jonathan E, Mark P Van Oyen. 2012. Design and optimization methods for elective hospital admissions. Tech. rep., Indiana University, Kelly School of Busniness, Bloomington, IN.

Hershey, John C, Elliott N Weiss, Morris A Cohen. 1981. A stochastic service network model with application to hospital facilities. *Operations Research* **29**(1) 1–22.

Hsu, Vernon Ning, Renato de Matta, Chung-Yee Lee. 2003. Scheduling patients in an ambulatory surgical center. *Naval Research Logistics* **50**(3) 218–238.

Huh, Woonghee Tim, Nan Liu, Van-Anh Truong. 2013. Multiresource allocation scheduling in dynamic environments. *Manufacturing & Service Operations Management* **15**(2) 280–291.

Johnson, Selmer Martin. 1954. Optimal two-and three-stage production schedules with setup times included. *Naval research logistics quarterly* **1**(1) 61–68.

Kamath, Janine RA, John B Osborn, Véronique L Roger, Thomas R Rohleder. 2011. Highlights from the third annual mayo clinic conference on systems engineering and operations research in health care. *Mayo Clinic Proceedings*, vol. 86. Elsevier, 781–786.

Kim, Seung-Chul, Ira Horowitz. 2002. Scheduling hospital services: the efficacy of elective-surgery quotas. *Omega* **30**(5) 335–346.

Koizumi, Naoru, Eri Kuno, Tony E Smith. 2005. Modeling patient flows using a queuing network with blocking. *Health Care Management Science* **8**(1) 49–60.

Litvak, Nelly, Marleen van Rijsbergen, Richard J Boucherie, Mark van Houdenhoven. 2008. Managing the overflow of intensive care patients. *European journal of operational research* **185**(3) 998–1010.

Macario, Alex, Terry S Vitez, Brian Dunn, Tom McDonald. 1995. Where are the costs in perioperative care?: Analysis of hospital costs and charges for inpatient surgical care. *Anesthesiology* **83**(6) 1138–1144.

May, Jerrold H, William E Spangler, David P Strum, Luis G Vargas. 2011. The surgical scheduling problem: Current research and future opportunities. *Production and Operations Management* **20**(3) 392–405.

Porteus, Evan L. 2002. *Foundations of stochastic inventory theory*. Stanford University Press.

Ribas, Imma, Rainer Leisten, Jose M Framiñan. 2010. Review and classification of hybrid flow shop scheduling problems from a production system and a solutions procedure perspective. *Computers & Operations Research* **37**(8) 1439–1454.

Ridge, JC, SK Jones, MS Nielsen, AK Shahani. 1998. Capacity planning for intensive care units. *European journal of operational research* **105**(2) 346–355.

Robb, WB, MJ Osullivan, AE Brannigan, DJ Bouchier-Hayes. 2004. Are elective surgical operations cancelled due to increasing medical admissions? *Irish journal of medical science* **173**(3) 129–132.

Robinson, Gordon H, Paul Wing, Louis E Davis. 1968. Computer simulation of hospital patient scheduling systems. *Health Services Research* **3**(2) 130.

Rockafellar, R. T. 1972. *Convex Analysis*. Princeton University Press.

Ruiz, Rubén, José Antonio Vázquez-Rodríguez. 2010. The hybrid flow shop scheduling problem. *European Journal of Operational Research* **205**(1) 1–18.

Samiedaluie, Saied, Beste Kucukyazici, Vedat Verter, Dan Zhang. 2013. Managing patient admissions in a neurology ward. Tech. rep., McGill University, Desautels Faculty of Management, Montreal, Quebec.

Shaked, M., J.G. Shanthikumar. 1988. Stochastic convexity and its applications. *Advances in Applied Probability* **20**(2) 427–446.

Topkis, D.M. 1998. *Supermodularity and complementarity*. Princeton Univ Press.

Wang, Yanchao, WL Hare, L Vertesi, AR Rutherford. 2010. Using simulation to model and optimize acute care access in relation to hospital bed count and bed distribution. *Journal of Simulation* **5**(2) 101–110.

Weiss, Elliott N, Morris A Cohen, John C Hershey. 1982. An iterative estimation and validation procedure for specification of semi-markov models with application to hospital patient flow. *Operations Research* **30**(6) 1082–1104.