Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and should not be used to distribute the papers in print or online or to submit the papers to another publication.

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

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

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

Brett Anderson, MD
Department of Pediatrics, Columbia University, New York, NY, USA, bra2113@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 increases with the state of the system and 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 real 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 a three-fold increase in total expenses.

Key words: Allocation Scheduling; Hospital Operations; Healthcare Management; Service Operations Management; Markov Decision Process

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 (El-Darzi, Vasilakis, Chaussalet and Millard 1998). These two-stage systems share a common feature that patients are often scheduled to arrive at the upstream stage, flow spontaneously to 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 (Koizumi, Kuno, and Smith 2005).

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, 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).

In reality, the effective operation of these two-stage healthcare systems should depend on the “balanced” use of capacity at both stages to ensure smooth patient flow. Indeed, evidence indicates that making scheduling decisions for the upstream stage that account for patient LOS in the downstream stage can bring 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 these evidences, 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).

With more hospitals adopting electronic health record systems following the Federal Government’s “meaningful use” initiative, 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 such a model from being easily analyzed or solved.

In this paper, we study a surgical unit as a canonic 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. 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 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 LOS in the downstream stage.

Our model is an aggregate planning model similar to those studied in Gerchak, Gupta and Henig (1996), Ayvaz and Huh (2010) and Huh, Liu and Truong (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 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.
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 in a 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 advance scheduling models, 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.

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 can be used to approximate the system by focusing on two main stages. 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 stages. Take the pediatric population as an example. Patients $\leq$ 21 years of age receive cardiac surgery in one of two pediatric operating rooms (OR) and recover in the pediatric cardiac intensive care unit (CICU). Many then move to a step-down unit, after the CICU, before being 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 on which surgery can be performed 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. 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 other stage, a result that has been confirmed earlier by simulation studies (Bowers 2013). Finally, through numerical experiments based on data collected from CUMC, we show that there are substantial values to making 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 inefficiency 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 presents numerical results on the value of integrated scheduling. Section 8 discusses several extensions of our model. Finally, Section 9 summarizes our work and discusses potential implications. All proofs of the technical results are shown in the Online Appendix.

2. Related Literature
Our work draws upon several streams of literature. First, it draws upon 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. 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.

Second, 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 who 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 are not appropriate to the healthcare context we consider, in which patients will be automatically added to a waitlist and providers usually do not “reject” patients. In addition, patients, once admitted, will usually stay in the hospital and not leave until discharge. More importantly, capacity in a hospital is often flexible rather than fixed. For instance, additional surgeries can be conducted with staff overtime (at some cost); patients may be placed in “swing beds” if regular beds are all occupied (Griffin, Xia, Peng and Keskinocak 2012). These important features of healthcare delivery lead us to develop the new model presented in this work.

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 analytic, dynamic model of scheduling. Previous work has mainly focused on simulation or static optimization models. Ridge, Jones, Nielsen, and Shahani (1998) and Kim and Horowitz (2002) use simulation to study bed capacity issues in hospital. A number of recent studies use static optimization model to determine the (weekly)
schedule of elective hospital admissions, taking into account their impact on the use of various hospital resources; see, e.g., Hsu, Ning, de Matta, and Lee (2003), Conforti, Guerriero, Guido, Cerinic and Conforti (2011), Helm and Van Oyen (2012), Fügener, Hans, Kolisch, Kortbeek and Vanberkel (2014), and Gartner and Kolisch (2014). Nunes, Carvalho and Rodrigues (2009) propose a dynamic decision model for elective patient admission control for distinct specialties. Samiedaluie, Kucukyazici, Verter, and Zhang (2013) consider patient admission policy in a neurology ward, where there are two stages of service, the ED 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.

3. Model

In this section, we describe our modeling framework. By convention (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 \( \delta_t \) and \( \epsilon_t \), respectively. We assume that \( \delta_t \) and \( \epsilon_t \) 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, 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 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. The patient stays in the downstream stage for a random number of
days before she is finally discharged. Specifically, 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 (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. Nevertheless, in our numerical study later we will test our model in general settings in which patient LOS does not follow the geometric distribution.

We assume that there is a single resource that is consumed by patients in each stage \(i \in \{0, 1\}\). (The case of multiple resources will be discussed in Section 8). 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 \(\upsilon_0\) of capacity in stage 0, and \(\upsilon_1\) of capacity in each day that she remains in stage 1. For each \(i \in \{0, 1\}\), we assume that \(\upsilon_i\) is i.i.d. 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 \(\upsilon_0\) and \(\upsilon_1\), respectively. Conceptually, our model assumes that one procedure/treatment begins when the previous one ends. We do not explicitly model within-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, Gupta, and Henig (1996) and Choi and Wilhelm (2014), such approximation of the workload within a day is reasonable and often used in aggregate-planning models.

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 cost structure is in line with those in the earlier literature, which charge costs for deviations from target capacity levels in order to stabilize average hospital resource utilization to desired levels; see, e.g., Adan and Vissers (2002) and Nunes, Carvalho and
Rodrigues (2009). To further explain our motivation, let us take the hospital’s perspective and assume that there is a sunk fixed cost, for example facility maintenance costs, 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. As discussed below, $P \geq R$ is a quite reasonable assumption in reality. Thus we have chosen to assume $P \geq R$ to reduce the number of parameters for the problem. (If this condition is not satisfied, then we can directly model the costs at stage 0 as $AC + P(S^0(k) - C)^+ - RS^0(k)$. None of our structural results will change.) Following a similar argument, this cost structure also applies to stage 1.

To give a sense the magnitude of these costs, we refer the readers to Macario (2010). Adjusting for inflation, one can estimate that in the US, the average OR cost is about $3000 per hour in 2015 US dollars, approximately 50% of which is the variable staffing cost (excluding fixed per-case surgeon and anesthesia-related professional fees). Thus, the constant $A$ above can be estimated at $1500 per hour. Overtime hourly staffing costs are at least 1.5 times the regular per hour costs due to the federally mandated overtime rate of 1.5 (Dexter Epstein, and Marsh 2001). Factoring in the increased staffing costs and other tangible as well as intangible costs due to overtime, the overall overtime unit cost $P$ is likely to be higher than the revenue rate $R$. Of note, for these estimates we use costs, as derived from cost-to-charge ratios. It is important to note that this is not equivalent to either hospital reimbursement or hospital expenditures, but, in the economics literature, is typically considered a reasonable proxy for resource utilization within a single center.

The expected overtime and idling costs at stage 0, given that $k$ patients are served, is $O_0E[S^0(k) - C_0]^+ + L_0E[S^0(k) - C_0]^−$. 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 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 $\delta_t$ 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 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 $\epsilon_t$ 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. (The overtime cost at stage 1 is the cost associated with using extra capacity, e.g., “swing beds” as mentioned earlier, to serve patients.)

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 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 $(w_t, n_t, \delta_t)$, 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:

$$w_{t+1} = w_t + \delta_t - q_t, \quad (1)$$
$$n_{t+1} = \xi_t(n_t + q_t + \epsilon_t). \quad (2)$$

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

$$\hat{F}(w_t, n_t, \delta_t, q_t) = Ww_t + O_0 E[S^0 (q_t + \epsilon_t) - C_0]^+ + L_0 E[S^0 (q_t + \epsilon_t) - C_0]^-$
$$+ O_1 E[S^1 (n_t + q_t + \epsilon_t) - C_1]^+ + L_1 E[S^1 (n_t + q_t + \epsilon_t) - C_1]^-, \quad (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 $\hat{V}_t(w_t, n_t, \delta_t)$ 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 $(w_t, n_t, \delta_t)$. The Bellman equation can be written as follows:

$$\hat{V}_t(w_t, n_t, \delta_t) = \min_{0 \leq q_t \leq w_t + \delta_t} \hat{G}_t(w_t, n_t, \delta_t, q_t),$$
$$\hat{G}_t(w_t, n_t, \delta_t, q_t) = \hat{F}(w_t, n_t, \delta_t, q_t) + \gamma E[\hat{V}_{t+1}(w_{t+1}, n_{t+1}, \delta_{t+1})|(w_t, n_t, q_t, \delta_t)],$$
$$= \hat{F}(w_t, n_t, \delta_t, q_t) + \gamma E[\hat{V}_{t+1}(w_t + \delta_t - q_t, \xi_t(n_t + q_t + \epsilon_t), \delta_{t+1})]. \quad (4)$$

For convenience, we take the termination function when $T < \infty$ to be $\hat{V}_{T+1}(\cdot, \cdot, \cdot) = 0$, 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. 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, 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](image_url)

*Figure 1* For a fixed waitlist size $w_0 = 8$, the optimal decision $q_t^*$ is not monotone in the state variable $n_0$. The number of patients is rounded to the nearest integer. $C_0 = 7$, $C_1 = 20$, $W = 2$, $O_0 = 6$, $L_0 = 6$, $O_1 = 5$, $L_1 = 5$, $E[v^0] = 1$, $E[v^1] = 1$, $\gamma = 0.9$, $E[\delta] = 6$, $E[\epsilon] = 0.5$, $T = 50$, $E[\xi] = 0.7$ and $\xi_t$ 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 to 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 as the system state changes. Next, we perform a 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. 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 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 \( n_t \) 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,

\[
a_{t+1} = a_t + \delta_t + \epsilon_t - (1 - \xi_t)(m_t + \epsilon_t) = a_t + \delta_t - m_t + \xi_t(m_t + \epsilon_t),
\]
\[
n_{t+1} = \xi_t(m_t + \epsilon_t).
\]

The single-day cost function can be written as

\[
F(m_t, a_t, n_t, \delta_t) = W(a_t - n_t) + O_0 \mathbb{E}[S^0(m_t - n_t + \epsilon_t) - C_0]^+ + L_0 \mathbb{E}[S^0(m_t - n_t + \epsilon_t) - C_0]^-
+ O_1 \mathbb{E}[S^1(m_t + \epsilon_t) - C_1]^+ + L_1 \mathbb{E}[S^1(m_t + \epsilon_t) - C_1]^-. \]
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 \((a_t, n_t, \delta_t)\). The Bellman equation can be written as follows:

\[
V_t(a_t, n_t, \delta_t) = \min_{n_t \leq m_t \leq a_t + \delta_t} G_t(m_t, a_t, n_t, \delta_t),
\]

(8)

where \( G_t(m_t, a_t, n_t, \delta_t) = F(m_t, a_t, n_t, \delta_t) + \gamma E[V_{t+1}(a_{t+1}, n_{t+1}, \delta_{t+1}) | (a_t, n_t, m_t, \delta_t)] \), 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 \textit{transformed formulation}. 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. \textbf{Structure of Optimal Solutions}

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.

\textbf{Proposition 1.} \textit{For every } \( t = 1, 2, \ldots, T \), \( F(\cdot, \cdot, \cdot, \cdot) \), \( G_t(\cdot, \cdot, \cdot, \cdot) \), and \( V_t(\cdot, \cdot, \cdot) \) \textit{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

\[
g(x) + g(y) \geq g(x \wedge y) + g(x \vee y),
\]

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 $(a_t, n_t)$.

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,

$$m_t^{\min}(a_t, n_t, \delta_t) = \min \arg \min_{m_t \leq m_t \leq a_t + \delta_t} G_t(m_t, a_t, n_t, \delta_t),$$

and

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

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 $\delta_t$, 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 $(a_t, n_t)$.

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 sizes of three nested sets of patients. 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. Similarly, as the smallest set containing $n_t$ recovering patients increases in size, the middle set containing $m_t$ admitted and recovering patients increases. The latter result implies that the possible reduction of new admissions in $q_t$ due to congestion downstream does not increase as quickly as the number of inpatients $n_t$ already in the hospital.
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. Note that these capacities in question are not the “physical” capacity of a hospital, changes to which would require building new space and making capital investment. These capacities are 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 is determined by the regular number of health care providers and resources put to work in a day; at the downstream, the capacity is determined by the number staffed beds (not certified or licensed beds). These capacities, however, can be adjusted by tactical decisions, for example, by the use of overtime and flexible beds as mentioned above.

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 $O_1E[S_1^1(m) - C_1]^+$. 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

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

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

Figures 2 and 3 illustrate 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. 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 \(\gamma\) approaches 1 and the planning horizon increases.
In Figure 2, the downstream capacity is $C_1 = 30$. Each patient is expected to stay $1/(1 - E[\xi]) = 6.67$ days in the system, and therefore, on average, the downstream stage can discharge $C_1(1 - E[\xi]) = 4.5$ patients each day, which is much smaller than the expected daily arrival rate of $E[\delta] + E[\epsilon] = 9.5$. 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 $C_1(1 - E[\xi]) = 4.5$, as the throughput is primarily determined by the downstream capacity.

In Figure 3, the capacity upstream is $C_0 = 5$, which is smaller than the average daily demand of $E[\delta] + E[\epsilon] = 9.5$. 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 $9.5/(1 - E[\xi]) = 63$. 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 $C_0/(1 - \mathbb{E}[\xi_t]) = 33$.

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 $i = 0,1$, 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, 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 $m_t^{\text{max}}(C,a_t,n_t)$ and $m_t^{\text{min}}(C,a_t,n_t)$ 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 $m_t^{\text{max}}(C,a_t,n_t)$ and $m_t^{\text{min}}(C,a_t,n_t)$ 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. In summary, increasing the upstream capacity allows a manager to be more reactive to fluctuations in downstream congestion; the manager may not use the full upstream capacity as the hospital may prefer to ramp up or ramp down admissions in order to “chase” inpatient occupancy.

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 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 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 costs now. However, as $C_0$ increases, 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.

7. 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 up on 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 delineates the value of integrated scheduling.

To study the value of 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 the operating room 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. We test the empirical performance 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 the paper, 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.

SINGLE1: A single-stage “optimal” scheduling policy that treats the upstream capacity as infinity.

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 that there is 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 sets of numerical experiments. In the first setting, we consider homogeneous patient arrivals and constant capacity in both stages over time. We set their values based on the average daily values estimated from data (more on this below). Stationary arrivals and capacity allows us to focus on the potential impact of factors other than non-stationarity on the value of integrated scheduling, for example, idling and overtime costs.

In the second setting, we consider a non-stationary environment populated by the actual data. Here, capacity levels are based on actual staffing levels, which vary over time. On top of the four scheduling policies discussed above, we also evaluate the actual strategy currently used by the hospital. This comparison shows how integrated scheduling could improve practice.

7.1. Distribution of length-of-stay

In our experiments, we generate scheduling policies using our model, based on geometrically distributed LOS (the estimated parameter is about 7%, i.e., 7% of patients on average get discharged every day). However, according to our data, patient LOS downstream in the
Integrated Scheduling

CICU at CUMC does not follow a geometric distribution. Figure 5 in the Appendix shows the histogram of patient LOS data as well as the density curve for the best fitted geometric distribution. Visually, the best fitted geometric distribution does not provide a good fit to the data due to the low densities of short LOS on the left tail. A Kolmogorov-Smirnov test confirms this (p-value < 0.01). Nevertheless, the numerical results show that our model still performs quite well 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.

7.2. Stationary Setting

We use the following parameters in our experiments:

**Arrival Rates.** The average daily number of admissions is 1.23 per day for elective patients and 0.98 per day for emergency patients, averaged over all weekdays in the year 2014. Accordingly, we set the emergency arrival rate $E[\epsilon_t] = 0.98$. For the elective arrival rate $E[\delta_t]$, we consider two different values, 1.23 and $1.23 \times 120\%$, where the latter represents a more congested system. The number of arrivals on each day is assumed to be a Poisson random variable.

**Upstream and Downstream Capacities.** The upstream capacity is $C_0 = 3$, which equals the expected total daily arrivals $E[\epsilon_t] + E[\delta_t]$. The downstream capacity $C_1$ is varied in the numerical experiments.

**Cost Rates.** The waiting cost is normalized at $W = 1$. Overtime costs are set to be $O_0 = 15, 30,$ and $O_1 = 40$. We experiment with different idling costs in two stages by choosing $(L_0, L_1) \in \{(10, 10), (1, 15), (15, 1)\}$.

**LOS.** We assume that, for the optimal policies and single-stage policies, a random fraction $1 - \xi_t$ of patients are discharged from the downstream stage on each day $t$. We use empirical distribution to generate $\xi_t$. In particular, we uniformly sample days in the year 2014, and set $1 - \xi_t$ to be the fraction of patients leaving on those days. The average daily discharge rate is 6.8%, which corresponds to a mean LOS of 14.7 days.

**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 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 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%. 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 19%.

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 329%, 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 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 that when the 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 the decisions of an inpatient unit manager who does not pay attention to the OR usage. Our

<table>
<thead>
<tr>
<th>$C_1$</th>
<th>$L_0 = 10$</th>
<th>$L_1 = 10$</th>
<th>$L_0 = 15$</th>
<th>$L_1 = 15$</th>
<th>$L_0 = 10$</th>
<th>$L_1 = 10$</th>
<th>$L_0 = 15$</th>
<th>$L_1 = 15$</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>3.003</td>
<td>3.214</td>
<td>3.292</td>
<td>1.027</td>
<td>1.010</td>
<td>1.055</td>
<td>1.027</td>
<td>1.010</td>
</tr>
<tr>
<td>31</td>
<td>2.797</td>
<td>2.948</td>
<td>3.166</td>
<td>1.030</td>
<td>1.010</td>
<td>1.066</td>
<td>1.030</td>
<td>1.010</td>
</tr>
<tr>
<td>32</td>
<td>2.563</td>
<td>2.647</td>
<td>2.999</td>
<td>1.034</td>
<td>1.011</td>
<td>1.078</td>
<td>1.034</td>
<td>1.011</td>
</tr>
<tr>
<td>33</td>
<td>2.310</td>
<td>2.335</td>
<td>2.802</td>
<td>1.034</td>
<td>1.012</td>
<td>1.092</td>
<td>1.034</td>
<td>1.012</td>
</tr>
<tr>
<td>34</td>
<td>2.057</td>
<td>2.042</td>
<td>2.582</td>
<td>1.035</td>
<td>1.012</td>
<td>1.104</td>
<td>1.035</td>
<td>1.012</td>
</tr>
<tr>
<td>35</td>
<td>1.820</td>
<td>1.784</td>
<td>2.345</td>
<td>1.035</td>
<td>1.012</td>
<td>1.117</td>
<td>1.035</td>
<td>1.012</td>
</tr>
<tr>
<td>36</td>
<td>1.615</td>
<td>1.572</td>
<td>2.125</td>
<td>1.035</td>
<td>1.012</td>
<td>1.135</td>
<td>1.035</td>
<td>1.012</td>
</tr>
<tr>
<td>37</td>
<td>1.448</td>
<td>1.407</td>
<td>1.907</td>
<td>1.034</td>
<td>1.012</td>
<td>1.148</td>
<td>1.034</td>
<td>1.012</td>
</tr>
<tr>
<td>38</td>
<td>1.317</td>
<td>1.284</td>
<td>1.706</td>
<td>1.032</td>
<td>1.011</td>
<td>1.158</td>
<td>1.032</td>
<td>1.011</td>
</tr>
<tr>
<td>39</td>
<td>1.218</td>
<td>1.194</td>
<td>1.532</td>
<td>1.032</td>
<td>1.010</td>
<td>1.164</td>
<td>1.032</td>
<td>1.010</td>
</tr>
<tr>
<td>40</td>
<td>1.149</td>
<td>1.129</td>
<td>1.387</td>
<td>1.029</td>
<td>1.009</td>
<td>1.171</td>
<td>1.029</td>
<td>1.009</td>
</tr>
<tr>
<td>41</td>
<td>1.099</td>
<td>1.085</td>
<td>1.274</td>
<td>1.027</td>
<td>1.008</td>
<td>1.172</td>
<td>1.027</td>
<td>1.008</td>
</tr>
<tr>
<td>42</td>
<td>1.065</td>
<td>1.055</td>
<td>1.188</td>
<td>1.025</td>
<td>1.007</td>
<td>1.174</td>
<td>1.025</td>
<td>1.007</td>
</tr>
</tbody>
</table>
experiments show that the system could overspend up to 17% of the optimal cost without considering the stage upstream. When $C_1$ is relatively small, the stage downstream is, in fact, the bottleneck. Policy SINGLE1 behaves similarly to OPT, and thus the total discounted cost under SINGLE1 is only a few percentage points higher than the optimal cost under OPT. 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 SINGLE1 and OPT increases as $C_1$ increases. Interestingly, however, we find that after $C_1$ is raised to a certain level, the performance of SINGLE1 and OPT 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 $C_0 = 3$, which is fixed at a higher level compared to the overall demand, would not be a significant bottleneck for patient flow. For the downstream stage not to be the bottleneck, $C_1$ has to be "balanced" with upstream demand (for instance, $C_1$ needs to be roughly $(1.23 + 0.98)/6.8\% = 32.5$ in Tables 1 and Tables 2). As the downstream capacity $C_1$ passes beyond these levels, both stages have relatively sufficient capacity for handling patient demand, and thus SINGLE1 would not behave too differently from OPT. However, as $C_1$ increases, the idling costs downstream bears increasingly more weight on the total cost, which cannot be offset much by integrated scheduling. Therefore, at a very high level of $C_1$, the cost of SINGLE1 is close to that of OPT because idling costs downstream dominate. Indeed, we see that the performance gap between SINGLE1 and OPT is smaller in Table 1 than Tables 2 and 3, as the idling cost downstream is relatively higher compared to other costs in Table 1.

In summary, 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. 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 costs of the bottleneck stage are significantly higher than those 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 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. But when the capacities of both stages are relatively balanced, a dynamic scheduling policy can make real-time adjustment to use capacity more efficiently.

7.3. Non-stationary Setting

In this section, we describe our numerical experiments in a non-stationary setting. In particular, we simulate and evaluate the performances of different scheduling policies (including the actual strategy used in practice) 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 4 summarizes the number of available ORs on each day and the corresponding upstream capacities. The upstream capacity 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. Clearly, the upstream capacity presents a non-stationary pattern. In addition to the baseline scenario, we also consider a large-capacity scenario in which daily capacity is 25-50% higher than the baseline value to study the impact of upstream capacity.

<table>
<thead>
<tr>
<th></th>
<th>1st and 3rd Monday</th>
<th>2nd and 4th Monday</th>
<th>Tue</th>
<th>Wed</th>
<th>Thur</th>
<th>Fri</th>
<th>Sat</th>
<th>Sun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Available ORs</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$C_0$ (Normal-Capacity Scenario)</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$C_0$ (Large-Capacity Scenario)</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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 23.27 over the year 2014. Instead of fixing the downstream capacity, we test various downstream capacity values in the range $[15, 30]$, as they cover most of the range of actual values (see Figure 6 in the Appendix for a histogram of the downstream census). Indeed, in our clinical setting, the exact number of downstream beds available for postoperative patients may not be constant as some of the chronic patients might hold beds there for a long time. For example, 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 stationary setting (i.e., $E[\epsilon_t] = 0.98$ and $E[\delta_t] = 1.23$). The arrival rates during weekends 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). Note that our data does not consider time spent at other hospitals prior to arrival at CUMC.

Cost Rates. We normalize $W = 1$, and set $L_0 = 5$, $L_1 = 15$, $O_0 = 20$, $O_1 = 40$.

Planning Horizon and Other Parameters. We simulate the costs of different scheduling policies using real data on patient surgery dates and LOS within a three-month period from April 2014 to June 2014. Specifically, we apply different scheduling policies to the sample arrivals over the three month period, and then sum up the costs incurred on each day. In the experiments, $v^0$, $v^1$ and $\gamma$ are set to be the same as in the stationary setting. 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 we compute in our simulated experiments will be smaller than real cost (incurred when policies are implemented in reality) by the same amount, for all policies we consider.

2. The optimal policies and single-stage policies will 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 will be 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 5). Since, in this table, the costs are not expected future costs, it is possible that SINGLE1 outperforms OPT marginally under the sampled actual arrivals.

Due to the two implications mentioned above, these cost ratios are more conservative (i.e., larger) than they would be were the optimal policy and single-stage policies implemented in reality. From Table 5, we observe that in this realistic non-stationary environment we tested, integrated scheduling still performs the best and brings significant reduction in overall costs. We also observe that the actual strategy performs quite close to SINGLE0. This conforms with the anecdotal note we received from the providers in
Table 5  Comparison against actual practice.

<table>
<thead>
<tr>
<th>C1</th>
<th>Normal-Capacity Scenario</th>
<th>Large-Capacity Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OPT</td>
<td>ACTUAL</td>
</tr>
<tr>
<td>15</td>
<td>0.51</td>
<td>0.99</td>
</tr>
<tr>
<td>16</td>
<td>0.46</td>
<td>0.99</td>
</tr>
<tr>
<td>17</td>
<td>0.45</td>
<td>0.99</td>
</tr>
<tr>
<td>18</td>
<td>0.44</td>
<td>0.99</td>
</tr>
<tr>
<td>19</td>
<td>0.44</td>
<td>0.98</td>
</tr>
<tr>
<td>20</td>
<td>0.42</td>
<td>0.98</td>
</tr>
<tr>
<td>21</td>
<td>0.37</td>
<td>0.98</td>
</tr>
<tr>
<td>22</td>
<td>0.49</td>
<td>0.98</td>
</tr>
<tr>
<td>23</td>
<td>0.60</td>
<td>0.97</td>
</tr>
<tr>
<td>24</td>
<td>0.74</td>
<td>0.97</td>
</tr>
<tr>
<td>25</td>
<td>0.84</td>
<td>0.97</td>
</tr>
<tr>
<td>26</td>
<td>0.88</td>
<td>0.98</td>
</tr>
<tr>
<td>27</td>
<td>0.93</td>
<td>0.98</td>
</tr>
<tr>
<td>28</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>29</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>30</td>
<td>0.98</td>
<td>1.00</td>
</tr>
</tbody>
</table>

this hospital that patient admission decisions 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 similar to that of SINGLE1 in all cases we tested, due to the fact 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). Also, when the downstream capacity $C_1$ gets smaller, cost savings will be more significant as decisions focus on downstream rather than the OR. These observations echo our earlier findings that misclassifying the bottleneck in the system can lead to a significant loss and that single-stage policies may perform reasonably well if there is an obvious bottleneck in the system.

8. Model Extension

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 costs incurred due to the use of individual resources. 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 and idling costs.

Heterogeneous Resource Usage. Emergency patients might have a different distribution of capacity usage compared to elective patients. Since emergency patients are always assumed to enter the upstream stage on the day that they become emergent, 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. One can then verify that 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 negligible emergency cases.

9. Conclusions

This paper 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 that provide intuitive insights for practitioners. Through numerical experiments on practical data, we demonstrate that using our model enables hospitals to significantly improve their operational efficiency compared to following the current practice, which often uses localized decision rules and only focuses on a single location of the hospital. Our work provides a stepping stone to study 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 lengths of stay, and advance scheduling decisions that assign patients directly to future days.

References


Appendix. Online Supplement Materials

A. Proofs of the Results

Proof of Proposition 1

Proof. We first show the joint convexity of \( F(\cdot, \cdot, \cdot) \). 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, \( \{S^i(k), k = 0, 1, 2, \ldots \} \) is stochastic increasing and linear in sample path sense; or in short, \( \{S^i(k), k = 0, 1, 2, \ldots \} \in \text{SI.L}(sp) \) (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 \( O_t \mathbf{E}(x + S^i_t(\epsilon_t) - C_t)^+ \) is convex in \( x \), we have that \( O_t \mathbf{E}(S^i(m_t) + S^i_t(\epsilon_t) - C_0)^+ \) is convex for \( m_t \in \{0, 1, 2, \ldots \} \). 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 \( O_t \mathbf{E}[S^0(m_t + \epsilon_t) - C_0]^+ \) is jointly convex in \((m_t, n_t)\). The convexity of the other terms in \( F(\cdot, \cdot, \cdot) \) 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

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

is also convex (see Theorem 5.7 in Rockafellar (1972)). Therefore, \( G_t(\cdot, \cdot, \cdot) \) is jointly convex. Because \( V_{t}(\cdot, \cdot, \cdot) \) is obtained by minimizing a convex function \( G_t(\cdot, \cdot, \cdot) \) 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 \( t \). \( \Box \)

Proof of Theorem 1

Proof. We start by proving (1). The first term \( W(a_t - n_t) \) in \( F(\cdot, \cdot, \cdot) \) 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 \( F(\cdot, \cdot, \cdot) \) are submodular. The last two terms are trivially submodular as they are both single variable functions.

We 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 \( F(\cdot, \cdot, \cdot) \) shown above, to prove (2), it remains to check the submodularity of its second term

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

for any realization of \( \delta_t, \xi_t \) and \( \epsilon_t \). We drop \( \delta_{t+1} \) from the expression above for notational convenience. For any \( a_t^+ \geq a_t^-, \) and \( m_t^+ \geq m_t^- \), we have

\[
egin{align*}
V_{t+1}(a_t^+ + \delta_t - (1 - \xi_t)m_t^+ + \xi_t \epsilon_t, \xi_t m_t^+ + \xi_t \epsilon_t) \\
-V_{t+1}(a_t^+ + \delta_t - (1 - \xi_t)m_t^- + \xi_t \epsilon_t, \xi_t m_t^- + \xi_t \epsilon_t) \\
-V_{t+1}(a_t^- + \delta_t - (1 - \xi_t)m_t^+ + \xi_t \epsilon_t, \xi_t m_t^+ + \xi_t \epsilon_t) \\
+V_{t+1}(a_t^- + \delta_t - (1 - \xi_t)m_t^- + \xi_t \epsilon_t, \xi_t m_t^- + \xi_t \epsilon_t) \\
\leq V_{t+1}(a_t^+ + \delta_t - (1 - \xi_t)m_t^+ + \xi_t \epsilon_t, \xi_t m_t^+ + \xi_t \epsilon_t)
\end{align*}
\]
\[-V_{t+1}(a^+_t + \delta_t - (1 - \xi_t)b^+_t + \xi_t \epsilon_t, \xi_t m^-_t + \xi_t \epsilon_t)\]
\[-V_{t+1}(a^-_t + \delta_t - (1 - \xi_t)b^-_t + \xi_t \epsilon_t, \xi_t m^-_t + \xi_t \epsilon_t)\]
\[+ V_{t+1}(a^-_t + \delta_t - (1 - \xi_t)b^-_t + \xi_t \epsilon_t, \xi_t m^-_t + \xi_t \epsilon_t)\]
\[\leq V_{t+1}(a^-_t + \delta_t - (1 - \xi_t)b^-_t + \xi_t \epsilon_t, \xi_t m^-_t + \xi_t \epsilon_t)\]
\[-V_{t+1}(a^-_t + \delta_t - (1 - \xi_t)b^-_t + \xi_t \epsilon_t, \xi_t m^-_t + \xi_t \epsilon_t)\]
\[+ V_{t+1}(a^-_t + \delta_t - (1 - \xi_t)b^-_t + \xi_t \epsilon_t, \xi_t m^-_t + \xi_t \epsilon_t)\]
\[= 0,
\]
where the first inequality follows from the submodularity, the second from the joint convexity of $V_{t+1}(\cdot, \cdot, \cdot)$.

Thus,
\[
E[V_{t+1}(a_t + \delta_t - m_t + \xi_t (m_t + \epsilon_t), \xi_t (m_t + \epsilon_t), \delta_{t+1})]
\]
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 $\delta_t$, the set $\{(m_t, a_t, n_t) \in \mathbb{R}_+^3 | n_t \leq m_t \leq a_t + \delta_t\}$ 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)). □

**Proof of Corollary 1**

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

**Proof of Theorem 2**

**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. □
B. Additional Figures

Figure 5  Histogram of Patient LOS and the Best Fitted Geometric Density Curve.

Figure 6  Frequency of the number of patients staying in the hospital in 2014.