Surgical Suites’ Operations Management

Diwakar Gupta
Graduate Program in Industrial & Systems Engineering, Department of Mechanical Engineering, University of Minnesota, 111 Church Street S.E., Minneapolis, Minnesota 55455, USA
guptad@me.umn.edu

Surgical suites are a key driver of a hospital’s costs, revenues, and utilization of postoperative resources such as beds. This article describes some commonly occurring operations management problems faced by the managers of surgical suites. For three of these problems, the article also provides preliminary models and possible solution approaches. Its goal is to identify open challenges to spur further research by the operations management community on an important class of problems that have not received adequate attention in the literature, despite their economic importance.

Key words: surgical suites’ capacity management; elective surgery booking control; surgery sequencing and scheduling; health care operations

Submissions and Acceptance: Received June 2007, Revised October 2007, Accepted October 2007.

1. Introduction

Surgical suites’ operations directly consume 9–10% of a hospital’s budget (Viapiano and Ward 2000). Hospital statistics from 2004 and 2005, available through the Hospital Cost & Utilization Project of the Agency for Healthcare Research and Quality (see http://www.ahrq.gov/data/hcup/), indicate that deferrable surgery procedures may be responsible for up to 52% of all hospital admissions. This estimate is obtained by subtracting the sum of admitted patients who come from the emergency department, another hospital, and long-term care facilities from the total number of admissions. Surgical suites’ operations management (OM) affects costs, patient flow, and resource utilization throughout a hospital. In this article, we identify a variety of decision problems that are faced by surgical suites’ managers, select three examples for closer scrutiny, and propose preliminary models to support those decisions. Our goal is to encourage further research on these topics. Because this requires a deeper understanding of the various types of decisions that affect a hospital’s costs and revenues, we begin by describing the environment in which the management of surgical operations occurs.

Hospitals perform surgeries either in response to an urgent or emergent need or on an elective basis. The term elective does not mean that the procedure is optional, but rather that its scheduling may be delayed to accommodate patient and provider convenience, as well as other more urgent cases. In the elective mode, the need for surgery is often identified after consultations with a patient’s primary care physician and possibly one or more specialists. In contrast, urgent need may arise as a result of a sudden acute episode, e.g., the occurrence of a minor stroke may identify the need for an immediate arterectomy. Finally, emergencies are usually associated with motor vehicle accidents, acute trauma, or medical conditions (e.g., stroke or heart attack) that create a life-threatening situation for the patient. Surgeries may be performed either on an inpatient or on an outpatient basis. In the inpatient setting, patients are admitted to the hospital prior to surgery and assigned a hospital bed, whereas in the outpatient setting they stay in the hospital only until recovery is complete and thereafter return to their usual residence.

Surgery planning includes both pre-operative and post-operative patient-care management. In the pre-operative phase, patients may receive education on how to care for themselves after surgery, as well as any necessary medical tests. Post-operative care occurs either in a post anesthesia care unit (PACU), or an intensive care unit (ICU), or else in some other post-procedure recovery area. Different levels of care are provided in these recovery areas, depending on the acuity of the patients’ needs. After recovery is com-
complete, inpatients return to their hospital beds, where they are cared for until the attending physician authorizes their discharge. Outpatients are discharged immediately after recovery.

Managing surgical operations requires an understanding of the hierarchy of decisions that affect these operations. At the strategic (highest) level of this hierarchy, hospital administrators must decide which types of surgeries will be performed at the facility, how many surgical suites will be built, and which types and what quantities of equipment and diagnostic tools will be purchased. Surgical suites can be specialized either by the type of procedure (medical specialty) or by the urgency level. The former is more common because many procedures require specialized medical equipment. Closely tied to these capacity decisions are decisions concerning the number of nursing staff and the degree to which they should be cross-trained to support different types of surgeries.

Access to elective surgery capacity is managed in one of two ways: block-scheduling and open (or non-block) scheduling. The choice between an open or a block-scheduling paradigm is also a strategic decision. Under a block-scheduling system, either individual or groups of surgeons are assigned blocks of operating room (OR) time in a periodic schedule (weekly or monthly). The surgeons may book cases into their assigned blocks subject to the condition that the cases “fit” within the block time. Mean surgery durations (obtained from historical records) are typically used to determine whether the cases fit. For cases that do not fit, the surgeon must request an allowance to overbook. In an open scheduling system, surgeons submit requests for OR time and an OR schedule is created prior to the day of surgery. The schedule specifies which surgeries are assigned to which ORs and their start times.

In many cases, hospitals use a hybrid of these two approaches in which surgical groups (or subspecialties) may share time depending on the demand that arises for their scheduled block time. This sharing is achieved by defining a deadline (a particular number of days prior to the day of surgery), at which time the unutilized block time of a surgical group becomes available for use by other groups. Another way to realize a hybrid system is to let a subset of ORs follow block schedules and let the remaining ORs have an open schedule. The latter are used to accommodate overflow and more urgent cases.

At the tactical (next) level of decision making, surgical suites’ managers must determine how much elective capacity to allocate to each surgeon or to each subspecialty. Although these problems arise only when block schedules are used, they are commonly encountered in U.S. and Canadian hospitals as a result of the widespread use of block scheduling. This statement is based on the author’s interactions with several health care networks in the United States and Canada. The author could not find published statistics on the proportion of hospitals that allocate OR time in blocks. Block allocations are periodically revisited when new capacity is added, when there is a change in the number of surgeons, or when medical-technology innovations alter the capacity usage of certain types of procedures. Tactical decision making also involves deciding how many surgeries to schedule in a particular block and assigning ORs and start times to each planned surgery so that the members of the surgical team (e.g., anesthesiologists, surgeons, and residents) know when and where to show up for each procedure.

Tactical scheduling decisions are sometimes performed in two steps. In the first step, patients’ medical priority levels, convenience, and various booking rules are used to determine an approximate time window (say, a week) for performing each procedure. In the second step, blocks of surgeries are assigned to specific ORs and start times are set for each procedure. Note that the latter has two parts: surgery sequencing and start times selection. The two-step procedure provides flexibility in scheduling tests and, when needed, appointments with certain specialists to clear the patients as candidates for surgery. A complication that arises when the same operating rooms are also used to handle urgent or emergent cases is that the OR manager must decide how much capacity to leave open for such cases.

The operational level of decision making involves dealing with deviations from plans. Such deviations occur when surgeries take longer than planned, certain equipment or key personnel are not available at the scheduled start time, there are a large number of emergency cases, or downstream resources are not available to take care of patients coming out of the surgical suites. It is the OR manager’s responsibility to try to achieve on-time closure of ORs while ensuring patients’ safety.

This paper describes challenges in modeling and solving three examples of surgical suites’ OM problems. Rather than provide complete solutions, the goal is to draw the OM community’s attention to these problems in the hope that this will spur further research. The problems are as follows.

1. How should a hospital allocate (redistribute) OR time to surgical subspecialties, either after expanding capacity or on a periodic basis?
2. How should a hospital manage elective surgery bookings?
3. Barring any medical constraints, what is the optimal sequence for performing surgeries? What is the efficiency loss from following heuristic rules for sequencing surgeries?

Each of these issues has attracted recent attention
either in the popular press or in the medical literature. For the last problem mentioned above, there are a few articles in the OM literature as well. The above list is not a complete list of important problems facing surgical suites’ managers, who use a variety of heuristics to find practical solutions to these and other problems (see Magerlein and Martin 1978; Blake and Carter 1997 for more details). These problems form the focus of our attention because there is insufficient research on modeling these problems.

Our goal is to propose preliminary models and explore potential solutions. Some of these problems can be discussed only after specifying a particular structural construct. In such cases, we have assumed characteristics that might represent a commonly occurring situation. Those results are, however, dependent on the assumed structure and cannot be generalized to all situations. Throughout this article, we use the following notational conventions. Uppercase letters indicate random variables; to avoid confusion between the numeral 1 and lowercase L, we use script ℓ. Vectors are denoted by the character accent ‘^-’ and matrices by boldface letters. Furthermore, for any a and b, a ∧ b = min {a, b}, a ∨ b = max {a, b}, and (·)^+ indicates max (·, 0).

Sections 2–4 describe each problem, its key trade-offs, and a preliminary formulation. Within each section, we also provide a review of relevant literature, possible solution approaches, model variants, and some open challenges for the OM community. Finally, we conclude the paper in Section 5 with certain problems for which no models have been proposed so far.

2. Elective Surgery Capacity Allocation

We consider a situation where a hospital has decided to increase either the number of ORs or the number of hours of operation because the existing OR time is fully used and demand is expected to increase in the future. Consequently, the hospital administration must decide how much block time to assign to different surgeons or services/subspecialties. A similar approach is also relevant for periodic review of OR time allocations. The OR time must be assigned well in advance to equip the ORs with the right equipment and to hire appropriate staff to support the subspecialties targeted to receive additional capacity.

The problem can be resolved without a model-based analysis, e.g., the hospital may deem it necessary to expand in certain service areas as a matter of strategy or it may decide to use the additional OR time to accommodate overflow from all services under the existing allocation. However, in many situations, the hospital administration does not have a clear idea about how OR time should be allocated (Oliver 2004). In such cases, it may be appropriate to allocate extra capacity to maximize the hospital’s total contribution, while meeting certain constraints. A possible set of constraints might be that no subspecialty should be worse off, in terms of capacity allocation, on account of the capacity expansion. Alternatively, during a periodic review exercise, it may be necessary to impose minimum allocation constraints for each subspecialty to maintain its clinical and economic viability.

A formulation of the resulting problem faced by the hospital administration, under certain simplifying assumptions, was recently proposed in a paper by Dexter, Ledolter, and Wachtel (2005), who also describe related literature. However, Dexter et al. make certain assumptions that restrict the applicability of their formulation to specific scenarios. In what follows, we first provide a general-purpose formulation of the problem. We discuss a solution procedure, variants of this formulation, and practical challenges in Section 2.2.

2.1. Model Formulation

Suppose there are i = 1, . . . , m subspecialties (or surgeons if OR blocks are assigned to individual surgeons). Let Ti = (T1, . . . , Tm) denote the demand for OR time by each of the m subspecialties. The support of each Ti is assumed to be the set of non-negative real numbers. The hospital’s contribution (revenue) per unit of allocated time is £, the existing allocations are q0, and the current capacity is κ0 = m

\[ \sum_{i=1}^{m} q_{i}^{0} \].

In this paper, we do not distinguish between revenue and contribution because a large proportion of a hospital’s costs are fixed (Canales, Macario, and Krummel 2001), although technically, contribution equals revenue minus variable costs. The hospital is adding new capacity such that the total available capacity after this addition will be κ = κ0. The problem facing the hospital is to decide how to allocate this capacity to maximize its total contribution, while meeting certain constraints. If the new capacity allocations (decision variables) are denoted by q̂ then, as discussed above, a possible constraint placed upon the allocation decision might be that q̂i ≤ qi for each i. Additional constraints may come from the availability of downstream PACU or bed capacity.

We rearrange subspecialty indices such that r1 ≥ r2 ≥ . . . ≥ rm. At least one of the preceding inequalities must be strict; otherwise, all subspecialties can be treated as a single specialty for the sake of modeling. We assume that all unutilized OR time can be assigned to miscellaneous procedures that earn an average contribution of r per unit time. [We defer the discussion of other reasonable assumptions regarding the use of excess allocation until Section 2.2.] Moreover, r < r1 because otherwise it would be optimal to allocate all extra capacity to miscellaneous procedures...
(equivalently, general overflow). The OR manager’s objective is to maximize
\[ z(\tilde{q}) = \sum_{i=1}^{m} E[r_i(T_i \cap q_i) + r(q_i - T_i)^+] \]
\[ = \sum_{i=1}^{m} r_i q_i - \sum_{i=1}^{m} (r_i - r) E(q_i - T_i)^+, \quad (1) \]
subject to the following constraints:
\[ \sum_{i=1}^{m} q_i \leq \kappa, \quad \text{and} \quad (2) \]
\[ q_i^0 \leq q_i \quad \text{for each } i = 1, \ldots, m. \quad (3) \]

The formulation in (1)–(3) is relatively easy to solve. We present a solution technique in Section 2.2, which generalizes the approach outlined by Dexter, Ledolter, and Wachtel (2005). This formulation may provide a first-cut solution to the hospital’s problem. However, there are a variety of issues not addressed in the formulation that may require a different approach. Such issues are also discussed in Section 2.2. The size of the problem (number of decision variables) faced by a typical mid-size hospital depends on whether the allocations are to individual surgeons or to subspecialties. In the former instance, the formulation may have several dozen decision variables—one for each surgeon who has operating privileges at the hospital—whereas in the latter scenario, the number of decision variables may be in the low double digits—one for each service or subspecialty.

### 2.2. Solution Approaches, Extensions, and Open Challenges

From the first-order optimality equations (see pp. 314–315 in Luenberger 1984), there exist \( q_i, \lambda, \mu_i, \ldots, \mu_{m\ell} \) such that \( \lambda \geq 0, \mu_i \geq 0, \text{ and} \)
\[ r_i F_i(q_i) + r F_i(q_i) + \mu_i - \lambda = 0, \quad (4) \]
\[ \mu_i(q_i^0 - q_i) = 0, \quad \text{for each } i = 1, \ldots, m \quad (5) \]
\[ \lambda \left( \sum_{i=1}^{m} q_i - \kappa \right) = 0. \quad (6) \]

In the above optimality equations \( F_i \) and \( \tilde{F}_i \) denote, respectively, the cumulative distribution function (CDF) and the complementary cumulative distribution function (CCDF) of the demand from subspecialty \( i \), with \( F_i(0) = 0 \). Equations (5) and (6) are also called the complementary slackness conditions.

From (4), we obtain
\[ \lambda = r_i - (r_i - r) F_i(q_i) + \mu_i \]
for each \( i = 1, \ldots, m \). Furthermore, because \( F_i \leq 1 \) for every \( i \), and \( r_i > r \), it follows that \( \lambda \geq r + \mu_i \geq r > 0 \).

An immediate consequence of this observation and equation (6) is that \( \sum_{i=1}^{m} q_i \) must equal \( \kappa \). Put differently, the OR manager should allocate all capacity among the \( m \) subspecialties. This is optimal because \( r_i > r \).

We now develop an algorithm to solve for optimal \( q_i \). Our algorithm is an iterative algorithm in which we successively refine values of \( \lambda \). The value of this parameter is in iteration \( k \) will be denoted by \( \lambda_k \) and the corresponding OR time allocations are denoted by \( q_i^{(k)} \). Initialize the iterative procedure by choosing \( \lambda_1 \in [r_{L+1}, r_1] \) for \( \ell_1 \) such that \( r \in [r_{L+1}, r_1] \). Following earlier arguments, \( \lambda_1 > r \) must also hold. That is, for indices \( i = 1, \ldots, \ell_1, r_i > \lambda_1 > r \) and for indices \( i = 1, \ldots, m, r_i \leq r < \lambda_1 \) for subspecialty indexed \( i > \ell_1 \), consider the following operations on equation (4):
\[ \mu_i = -r_i F_i(q_i) - r F_i(q_i) + \lambda_i = (\lambda_i - r_i) \]
\[ - (r_i - r) F_i(q_i) \geq (\lambda_i - r_i) - (r_i - r) > 0. \quad (7) \]

Therefore, from (5), \( q_i^{(1)} = q_i^0 \) for \( i = \ell_1 + 1, \ldots, m \).

For the remaining subspecialties, we assume that some new capacity is added to their previous allocation. This is obtained by solving
\[ q_i^{(1)} = F_i^{-1}(r_i - \lambda_1)/(r_i - r) \]
for each \( i = 1, \ldots, \ell_1 \). If we find a \( \lambda_{i+1} \in [r_{L+1}, r_{L+2}] \), such that
\[ \sum_{i=1}^{\ell_1} F_i^{-1}(r_{i+1} - \lambda) = \kappa - \sum_{i=\ell+1}^{m} q_i^0, \quad (8) \]
then we have the best allocation when subspecialties \( 1, \ldots, \ell_1 \) receive additional allocations. This procedure can be repeated for each range of possible values of \( \lambda \). Specifically, this means setting \( \lambda \) in range \([r_{L+1}, r_{L+2}], [r_{L+2}, r_{L+3}], \ldots\) successively. Arguments similar to those behind inequality (7) can be used to show that when \( \lambda \in [r_{L+1}, r_L] \) for some \( \ell < \ell_1 \) only subspecialties \( i = 1, \ldots, \ell \) receive additional capacity allocations. To see this, consider the following variant of arguments embodied in (7) for \( i > \ell \) such that \( r_i > r \):
\[ \mu_i = -r_i F_i(q_i) - r F_i(q_i) + \lambda = (\lambda - r_i) \]
\[ + (r_i - r) F_i(q_i) > 0. \quad (9) \]

For \( i > \ell \) such that \( r_i \leq r \), the arguments in (7) continue to hold. Thus, the complete solution procedure involves a finite number of steps, each of which can be solved relatively quickly. The overall best solution is the one that maximizes hospital’s profit, \( z(\tilde{q}) \), among all feasible solutions.

There are a variety of ways in which the formulation in Section 2.1 can be enhanced to match the realities faced by an OR manager. We present first an alternate formulation that can be solved using a variant of the above methodology. We also discuss other formul-
tions that could form the basis of future work in this area.

Suppose the OR manager could use leftover capacity from any subspecialty, but only to the extent that there is additional demand in another subspecialty to take advantage of this capacity. When this happens, it is generally not possible to utilize all available capacity (because of stacking loss) or the OR manager may need to schedule some overtime to make use of the OR time most effectively. The latter happens when a surgery that usually takes 1 hour must be fitted into a leftover time slot of 45 minutes, whereas the former occurs when a decision is made to idle the OR for the 45 minutes. The overall effect of this inefficiency can be modeled using historical data to estimate the per-unit time contribution, \( r \), from surgeries that are scheduled in this way. It makes sense to assume that \( r \leq r_i \) for all \( i \); otherwise, it will be beneficial for the hospital not to allocate any time to some subspecialties and earn a higher contribution by giving them OR time on an ad hoc basis.

The overall excess capacity in a particular planning period is \( \kappa - \sum_{i=1}^{m} (q_i \land T_i) \), whereas the overall excess demand is \( \sum_{i=1}^{m} (T_i - q_i)^+ \). Using \( T = \sum_{i=1}^{m} T_i \) to denote the convolution of OR usage by all subspecialties, we obtain the following formulation of the OR manager’s problem,

\[
\text{Maximize } \bar{z}(\bar{q}) = E \left[ \sum_{i=1}^{m} r(T_i \land q_i) \right] \\
+ r E \left( \sum_{i=1}^{m} (T_i - q_i)^+ \land \left( \kappa - \sum_{i=1}^{m} (q_i \land T_i) \right) \right) = \sum_{i=1}^{m} (r_i - r) q_i \\
- \sum_{i=1}^{m} (r_i - r) E(q_i - T_i)^+ + r E(T \land \kappa), \tag{10}
\]

subject to constraints (2) and (3). The above formulation is similar to the one presented earlier and the earlier solution algorithm can be adapted to solve this formulation. However, because the leftover capacity results in a lower contribution (because \( r \leq r_i \) for all \( i \) and it may not be possible to utilize all excess capacity), we expect that the block allocations will be relatively higher.

The formulation in (10) is based upon the assumption that the contribution from excess OR time is \( r \), regardless of which subspecialty makes use of this capacity. This makes it easier to write (10) because \( z(\bar{q}) \) is independent of how different subspecialties are given access to the excess capacity and what rules are used to manage stacking loss. Absent this assumption, we will need to model both the excess capacity allocation decision and the decisions about which procedures should be permitted to be scheduled. Our approach is not unreasonable from a practical perspective because hospitals may release excess space to surgeons on a first-come-first-served basis to avoid the appearance of favoring certain subspecialties. However, by aggregating contributions from different subspecialties, our formulation can potentially hide inefficiencies in the management of excess OR time.

Next, consider some enhancements that require a different formulation and solution approach. First, the hospital incurs certain non-recurring costs of equipment and staff hires that are dependent on capacity allocation decisions. However, the formulation in (1)–(3) ignores such costs. Fixed costs may be important to model in some situations. Models with fixed costs lead to discrete optimization problems that are significantly more difficult to solve. Second, the notion of a fixed contribution per unit of allotted time is a modeling convenience. In reality, surgical services are scheduled and billed in discrete units. Therefore, how well the \( i \)th surgical subspecialty can utilize an increase in \( q_i \) may affect \( r_i \). This means that per-unit revenue is a function of allocation. Similarly, the minimum allocation may also depend on clinically acceptable maximum delays in scheduling high-priority cases. Such dependencies must be modeled using historical data on procedure times and surgery-scheduling delays and then factored into capacity allocation decisions. Third, it may be difficult to obtain reliable data on the true demand for each subspecialty. Estimating true demand is particularly difficult in the healthcare setting where supply-induced demand is a well-documented phenomenon (Manning et al. 1987; Evans 1973).

Dexter, Ledolter, and Wachtel (2005) address the demand estimation problem by assuming that only those subspecialties that earn more than the average contribution would be eligible for additional capacity. Then, they assume that the demand for the remaining subspecialties is uniformly distributed between \( T_j \) and \( 2T_j \), where \( r_j > \bar{r} \) and \( \bar{r} = (1/m) \sum_{i=1}^{m} r_i \). One could take a variety of different approaches to solve this problem. For example, investing in some dedicated and some flexible equipment or staff hires may provide the opportunity to learn more about the true demand during a trial period after which allocations can be adjusted to find a better match between supply and demand. Problems of this nature can be formulated as two-stage stochastic programs with recourse.

Yet another alternative might be to design truth-revealing schemes where surgeons, who are likely to have more accurate demand information, are given incentives to report the true demand. Incentives need not be monetary. The hospital could reward truthful reporting by prioritizing excess capacity allocations to subspecialties who report demand truthfully or by
taking back block time from subspecialities who are consistently unable to fill their blocks. Redistribution of OR time can be carried out at predetermined periodic review intervals. In some health care networks, surgeons are not held accountable for underutilizing the allotted block time, with the result that they do not have a strong reason to report true demand.

OR managers can use a principal-agent framework with asymmetric information (examples of this approach can be found in Fudenberg and Tirole 1991; Salanié 1997; Laﬀont and Martimort 2002; Bolton and Dewatripont 2005) to model the problem of choosing incentives. In this approach, each surgeon is offered a menu of allotments, along with incentives for effective utilization of the allotted block time. The OR managers’ problem is that of finding a menu of capacity allocations and corresponding payment functions for each surgeon such that it is in the best interest of the surgeon to choose the overall best capacity allocation, i.e., one that maximizes the combined beneﬁt of the hospital, the surgeons, and the patients. The unique features of the OR capacity allocation problem, which include multi-priority demand, provider preferences, and staffed-bed capacity constraints, present opportunities both for translational research and for building new models that capture the realities of the application domain.

3. Elective Surgery Booking Control

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. This can cause a variety of problems in patient ﬂow. For example, it can lead to ORs getting blocked, increased patient waits, excessive use of overtime, and inability to handle emergencies, which causes ambulance diversion (Litvak et al. 2001, 2005). We describe a stylized model for surgery booking control, which takes into account limited capacity of a critical downstream resource.

3.1. Model Formulation

For modeling the booking control decision, we will assume that the surgeons have agreed to divide all surgery types into \( n \) classes. A class can be the combination of urgency, patient characteristics, revenue, or some other basis of classifying patients. For patient class \( i \), there is a maximum delay \( \tau_i \) and classes are indexed such that \( 1 = \tau_1 \leq \tau_2 \cdots \leq \tau_n \). We use a day as a unit of time for convenience. In our taxonomy, class 1 is the emergent class, i.e., patients in this class are served on the day their request is received. The maximum delays could be determined either by consensus of the operating physicians or by ascertaining acceptable delay norms within surgery departments across a number of different hospitals. A more sophisticated approach is described by Alter et al. (1999), who use clinical and nonclinical determinants to recommend a patient-speciﬁc maximum waiting time for coronary angiography.

To determine the day on which to book a type-\( i \) surgery, our model trades oﬀ the cost of making downstream resources available (regular and overtime) and the costs induced by the need to serve future high-priority cases. The model determines the amount of downstream capacity (PACU, ICU, or beds) that should be reserved exclusively for more urgent cases. Thus, a request may not be scheduled on the ﬁrst available date. However, we assume that the hospital is obligated to meet this demand within \( \tau_i \) time periods, even if that means scheduling overtime in the downstream units. A type-\( i \) procedure earns an average revenue (or contribution) of \( r_i \). The hospital’s regular-time daily capacity for postoperative care is \( \kappa \) (measured in minutes).

A tractable model of the surgery booking control problem is diﬃcult to formulate because, following surgery, patients may require care for several days in a downstream unit and the lengths of stay are not known with certainty. Thus, each booked surgery consumes an unknown and discrete chunk of the downstream unit’s resources. We simplify this problem in a variety of ways. Our aim is not to capture every detail of surgical operations. Rather, we want to demonstrate that models can capture features that have ﬁrst-order effect on operating policies and identify good operating policies. A detailed computer simulation is often necessary to carefully study the implication of using such policies before implementation.

Let \( D_{it} \) denote class-\( i \) demand (measured in minutes of the critical downstream resource) that arrives in period \( t \). We assume ﬂuid demand—speciﬁcally, any portion of \( D_{it} \) can be scheduled on days \( t, t + 1, \ldots, t + \tau_i \). Moreover, we do not model multi-period demand. This assumption is reasonable for recovery areas such as PACUs, but not for downstream units with staffed beds where a patient may be cared for over several days. Models with multi-period demand represent an open research problem that deserves attention in future studies.

The booking control problem is modeled on a rolling horizon basis with a planning horizon of \( \tau = \tau_n \) at each iteration of this process. At the start of each new day, one day is added at the end of the planning horizon and the ﬁrst day drops oﬀ. The booking state of the hospital at the start of a day is denoted by \( s = [s_1, \ldots, s_{\tau - 1}] \), where \( s_i = (s_{1,i}, \ldots, s_{n,i}) \) and \( s_{ij} \) is the amount of day-\( t \) downstream capacity previously committed to type-\( i \) surgery. Let \( s_{ij} = \sum_{i=1}^{\tau} s_{ij,t} \). Then, \( (\kappa - s_{ij}) \) is the amount of capacity available \( t \) days from today. Because the \( th \) day is added at the be-
ginnning of each new day, the available capacity on that day is always $\kappa$. Put differently, $s_i = (0, \ldots, 0)$. Therefore, it is not necessary to include $s_i$ in the system state vector.

The decision variables are $\{a_{ij}\}$, $i = 2, \ldots, n$ and $t = 1, \ldots, \tau$, where $a_{ij}$ is the amount of type-$i$ demand that is scheduled for day $t$. Note that type-1 demand must be satisfied in the same period in which it arrives. Therefore, $a_{11} = d_1$ and $a_{1t} = 0$ for each $t > 2$. Demand allocations must satisfy the following constraints: $a_{ij} \geq 0$ for each $i$ and each $t$; and $\sum_{t=1}^{\tau} a_{ij} = d_{ij}$. Because of the hard constraints imposed by medically acceptable delays, it is possible to have large overtime on certain days. We do not place a limit on the amount of available overtime. However, to prevent excessive use of overtime, the overtime cost, denoted by $h(\cdot)$, is assumed to be an increasing convex function. That is, per-unit overtime cost is increasing in the amount of overtime used, which reflects staff preference for a stable work schedule and the increasing cost of exceeding certain thresholds. A hospital manager can obtain good booking policies by choosing an overtime cost function that matches the realities of OR suites’ operations.

Let $s$ and $s'$ denote the state on two consecutive days, where $s'$ follows $s$. Then, the system evolves according to the system dynamics equations,

$$s'_{i,t} = s_{i,t+1} + a_{i,t+1}, \quad \forall i = 1, \ldots, n \text{ and } t < \tau,$$

with the understanding that $s_{i,0} = 0$ for each $i$. We are now ready to write Bellman’s optimality equations. In the sequel, $j$ denotes the iteration index ($j = 1, 2, \ldots$), $v_j(\cdot)$ denotes the optimal value function, and $\beta$ denotes the discount rate.

$$v_j(s) = E \left[ \max_{\{a_{ij}\}} \left( r_i D_1 + \sum_{i=2}^{n} r_i (s_{i,1} + a_{i,1}) - h \left( \left( D_1 + \sum_{i=2}^{n} [s_{i,1} + a_{i,1}] - \kappa \right)^+ \right) + \beta v_{j+1}(s') \right) \right]$$

Terms on the right-hand side of (12) can be explained as follows. The first two terms capture the current period revenue. Note that the revenue from demand allocations to future days and future demand arrivals, net of any overtime costs, is included in term $v_{j+1}(\cdot)$, which represents maximum net benefit from all future surgical procedures. The third term is the current-period overtime penalty. Solving for the optimal actions $a_{ij}$ subject to the constraint that $\sum_{t=1}^{\tau} a_{ij} = D_{ij}$, can be challenging. The difficulty comes from having a vector-valued state space and multi-dimensional action space. For certain procedures, the maximum permissible delay can be in months; thus, $\tau$, the maximum number of periods in the formulation, can be in the hundreds. A typical hospital performs hundreds of surgical procedures and many use three priority levels—deferrable, urgent, and emergent—for each procedure type. Thus, the number of classes, $n$, can be in the hundreds to the thousands for a typical problem. The number of decision variables in the model formulation is $n\tau$.

### 3.2. Solution Approaches, Extensions, and Open Challenges

Notwithstanding the complexity of the formulation in Section 3.1, it is possible to tease out some useful properties of the optimal policy that provide guidelines about what types of implementable heuristics to evaluate in detailed models (e.g., computer simulation). We devote the bulk of this section to illustrate how to determine useful properties of the optimal policy. But first, we present a review of the related literature.

Bruin, Koole, and Visser (2005) use a queueing model to study emergency cardiac patient flow and find that the source of the bottleneck is the lack of availability of beds downstream in the care chain. However, they do not propose a method for addressing this problem. Chase (2005) develops a simulation model for Vancouver Coastal health to reduce wait times, improve resource utilization, and determine the right number of downstream resources needed. This simulation model provides a virtual test bed for testing various strategies for smoothing demand and managing variability in procedure times. However, it does not contain any guidelines about what types of booking control policies should be considered.

Rohleder, Sabapathy, and Schorn (2005) use goal programming to find the optimal lengths of block schedules to minimize the sum of under- and overutilization costs of the OR. This is a trial-and-error method coupled with a discrete-event simulation model with the goal of finding the optimal block lengths. The booking control problem is not modeled. In fact, there are a limited number of articles focusing on the allocation of medical service capacity among distinct demand streams.

Gerchak, Gupta, and Henig (1996) consider the problem of reserving surgical capacity for emergency cases when the same operating rooms are also used for elective surgeries and surgery durations are random. They formulate the problem as a stochastic dynamic program and characterize the structure of the optimal policy. Patients are not given a particular appointment for surgery; rather, they reside in a queue until it is their turn to be treated. In contrast, we model different urgency levels and a procedure in which requests must be assigned to a particular day as they arrive.

Green, Savin, and Wang (2006) analyze the problem of scheduling demand for a diagnostic medical facility
that is shared by outpatients, inpatients, and emergency patients. Their analysis focuses on establishing optimal service priority rules for admitting patients into service on a particular day of operations, given an arbitrary outpatient appointment schedule. In contrast, our focus is on determining how best to assign arriving surgery requests to future days.

We illustrate a possible approach by considering a special case in which $n = 2$, $\tau_1 = 1$, and $D_{j,t} = D_t$ are i.i.d. random variables. Thus, type-1 demand is from urgent cases, whereas type-2 refers to all non-urgent cases that must be served in $\tau_2 = \tau$ days after a request is made by a surgeon. Several system parameters can be simplified as a result of having a two-urgency-class model. Action $a_t$ now denotes the amount of type-2 demand assigned to the $t$th day, whereas all type-1 demand must be met on the day it arrives. The system state can be captured by a vector $\hat{s} = (s_1, \ldots, s_{\tau-1})$ because only type-2 demand can be deferred. Finally, we can write (12) as

$$v_j(\hat{s}) = E \max \left\{ \sum_{i=1}^{\tau} r_i D_i + r_j (s_j + a_j) - h((s_j + a_j + D_j - \kappa)^+) + \beta v_{j+1}(\hat{s}') \right\}. \quad (13)$$

Using the value iteration algorithm, it can be demonstrated that when $\beta < 1$, the value function converges in the limit as $j \to \infty$ (see details in Puterman 1994). We use the notation $v(\hat{s})$ to denote the limiting value function and work only with $v(\hat{s})$ in the remainder of this section.

Define $s_k = 0$ and $b_1 = s_1 + a_1$. Then, the limiting value function can be rewritten as

$$v(\hat{s}) = r_1 E(D_1) + E \max_{b_1 \geq 0} \left\{ r_2 b_1 - h((b_1 + D_1 - \kappa)^+) + \beta v(b_2, \ldots, b_\tau) \right\}. \quad (13)$$

The maximization above is performed under the additional constraint that $\Sigma b_i = \Sigma s_i + D_\tau$. Given realization $b_1$, let $\phi(b) = r_2 b_1 - h((b_1 + D_1 - \kappa)^+) + \beta v(b_2, \ldots, b_\tau)$. Then, using induction arguments, it can be shown that $v(\cdot)$ and $\phi(\cdot)$ are concave functions. We omit the details to keep this presentation brief.

The OR manager’s problem is to find a set of booking controls, represented by $\{b_t\}$, that maximize the overall hospital revenue while meeting the access targets imposed by the parameter $\tau$. This problem is well behaved because $\phi(b)$ is concave and the constraints are linear in $b_t$. Note that $\phi(b)$ is increasing in $b_1$ so long as $b_1 + D_1 \approx \kappa$. This means if $s_k + d_1 < \kappa$, type-2 demand should be assigned to the current day until the capacity limit is reached or all of the new type-2 demand is satisfied. Put differently, we always set $b_1 = \min \{d_2, \kappa - (s_1 + d_1)\}$ when $s_1 + d_1 < \kappa$. This makes intuitive sense because any unused current-day capacity perishes at the end of the day. The booking control problem is non-trivial only when $d_2 > \kappa - (s_1 + d_1)$.

When $d_2 > \kappa - (s_1 + d_1)$, the problem faced by a hospital manager is a concave maximization problem subject to linear constraints. Therefore, standard nonlinear optimization techniques can be applied to study the nature of the optimal solution. Using the first-order optimality conditions (which are necessary and sufficient in this instance; see details in Luemberger 1984, Section 10.8), we obtain the following simultaneous equations:

$$\mu_1 = \lambda - (r_2 - \nabla_{b_1} h((b_1 + D_1 - \kappa)^+))$$

$$\mu_k = \lambda - \beta \nabla_{b_k} v(b) \quad \text{for each } k = 2, \ldots, \tau$$

Note $\nabla_s$ denotes the gradient with respect to $x$, $\lambda$ and $\mu_k$ are Lagrange multipliers, $\lambda$ is unrestricted in sign, $\mu_k \geq 0$ for each $k = 1, \ldots, \tau$, and at least one of the following equalities holds for each $k$: $\mu_k = 0$, $b_k = s_k$. Suppose realized demand $d_2$ is assigned to days belonging to the index set $J$. Then, the above arguments imply that for any $j$ and $k$ belonging to the set $J$, $\nabla_{s_j} v(b) = \nabla_{b_k} v(b)$. Put differently, optimal allocations are such that the incremental benefit of assigning an infinitesimal portion of $d_2$ to all future days is the same.

To understand how this may help identify a good booking control policy, consider the magnitude of benefit from assigning a portion of $d_2$ to the $k$th day in the future. We know that this benefit is negative when $b_k$ exceeds a certain threshold and therefore further assignments imply the use of expensive overtime. We also know that this benefit is positive when $b_k$ is small and that similar arguments also apply to day $k + 1$ with the difference that the benefit is discounted by a factor $\beta$. Thus, when $s_k$ is small and $s_k = s_{k+1}$, the incremental benefit of assigning to day $k$ is larger. Similarly, when $s_k$ is large and $s_k = s_{k+1}$, the incremental benefit of assigning to day $k$ is smaller. Therefore, to equalize the marginal benefits, we will assign more demand to the $k$th day when $s_k$ is small. Conversely, we will assign more demand to the $(k+1)$th day when $s_k$ is large.

The properties of the optimal booking policy described above can be explained on an intuitive level. If the previous commitments of capacity are relatively small, assign new requests to earlier days. If previous commitments are relatively large, push demand out into the future. The former allows us to better utilize available capacity, which will otherwise go unused. The latter allows us to push overtime costs into the future and preserve capacity in the near future for urgent arrivals. The latter approach also provides more reaction time to adjust staffing levels, which may help to drive down overtime costs even further. The
model helps by showing that booking control on day \( t \) for the \((t + k)\)th day should vary by the current state of bookings for that day, relative to the amount of capacity that is normally booked by time \( t \). Fixed booking controls may not be optimal. In particular, the amount of capacity available for a surgical procedure with a particular priority level changes over time and it is possible for availability to increase as we approach the day of the surgery.

One drawback of our formulation is the lack of concern for fairness. It can result in a later-arriving patient of equal urgency rank receiving treatment earlier than an earlier-arriving patient. This happens when not enough more-urgent demand materializes, which releases near-term capacity for later arrivals. In practice, this problem can be overcome by alerting patients of the possibility of an earlier availability of resources and allowing them to be placed on a voluntary call list. When earlier capacity becomes available, they can be offered the possibility of an earlier treatment first.

Some straightforward enhancements of the model we presented above include analysis for the case when \( n > 2 \). The trade-offs are less transparent in such cases and detailed numerical analysis may be necessary to identify heuristic policies that work well. The above model can be viewed as an aggregate capacity planning model because it ignores the discreteness, the timing, and the multi-period nature of demand for downstream resources. Within each period, the discrete nature of demand and the bunching of departures from ORs can cause the downstream resources to experience unmanageable demand peaks even when the overall capacity during the period is sufficient. Multiple period demand introduces an additional dimension to the action space—when a type-\( i \) demand occurring in period \( t \) is scheduled starting in period \( t' \), it may consume downstream capacity in periods \( t', t' + 1, \ldots, t' + \xi_i \), where \( \xi_i \) is the length of stay of type-\( i \) patients.

Another important problem is that of linking demand with staffing plans. The number of staff can be changed only in discrete units. Staff planning is periodic and often shift schedules are fixed for 4–6 weeks at a time, which limits the degree of flexibility available to operations managers to respond to changing staffing requirements. Linking elective surgery booking control with staffing models can yield substantial efficiencies and represents an untapped opportunity for OM researchers.

4. Elective Surgery Sequencing

Next, we turn to the problem of sequencing booked surgeries for a particular period of OR time. This problem arises after the OR manager has decided how many elective procedures to book. Whether this period corresponds to a block-booking environment or to a non-block environment is immaterial. In the ensuing formulation, we assume that all sequences are medically feasible. When medical constraints or provider preferences can be articulated, they can be added to the formulation as appropriate constraints.

4.1. Model Formulation

Given that \( p \) surgeries, with durations \( \bar{Z} = (Z_1, \ldots, Z_p) \), must be performed during a fixed session of length \( d \), a concomitant problem is one of determining the start times (or, equivalently, allowances) at which surgical teams and patients are told to arrive at the OR. We assume that the hospital administrator has a procedure for determining optimal start times for any given \( p, d \), and a sequence in which surgeries will be performed. One such procedure is described by Denton and Gupta (2003); these authors also provide a summary of other approaches that have appeared in the literature. We also assume that patients and service provider teams arrive punctually at scheduled appointment times. We shall discuss a variety of enhancements to our formulation in Section 4.2.

Let \( i_k \) be the index of the \( k \)th surgery in the sequence and \( x_{ik} \) be the corresponding allowance. That is, the surgeries are performed in the sequence \( \{i_k\}_{k=1}^p = (i_1, \ldots, i_p) \), where the sequence \( \{i_k\} \) is a permutation on the set \( \{1, \ldots, p\} \). We use \( \mathcal{S}_p \) to denote the set of all possible sequences when there are \( p \) surgical procedures. Note that we must specify allowances only for the first \( (p - 1) \) procedures. Assuming that the first procedure starts on time, start times \( a_i \)'s can be determined from \( x_i \)'s as follows: \( a_i = 0 \) and \( a_i = \Sigma_{j=1}^{k-1} x_j \) for \( k = 2, \ldots, p \).

Given \( (\{i_k\}_{k=1}^p, \bar{c}, \bar{Z}) \), key performance measures are \( \bar{W} \), the waiting times, \( \bar{S} \), the OR idling times, and \( \bar{L} \), the tardiness with respect to \( d \), which can be determined from the recursions

\[
W_i = (W_{i-1} + Z_{i-1} - x_{i-1})^+, \quad k = 2, \ldots, p, \quad (14)
\]

\[
S_i = (-W_{i-1} - Z_{i-1} + x_{i-1})^+, \quad k = 2, \ldots, p, \quad (15)
\]

\[
\bar{L} = \left( W_p + Z_p + \sum_{i=1}^{p-1} x_i - d \right)^+, \quad (16)
\]

where we set \( W_1 = S_1 = 0 \) as a consequence of on-time arrivals of surgeons and patients. Note that waiting and idling satisfy a parity relationship, i.e., \( W_i \cdot S_i = 0 \), \( k = 1, \ldots, p \).

Vectors \( \bar{c}^w \) and \( \bar{c}^s \) specify the cost coefficients for waiting by the patient or provider team and OR idling. Similarly, \( c_e \) is the unit cost of tardiness with respect to \( d \). Our goal is to determine the optimal (cost minimizing) sequence, which we shall denote by \( (i^*_1, \ldots, i^*_p) \).
Let $\phi(\tilde{x}; \{i_1, \ldots, i_p\})$ denote the expected total cost when surgeries are scheduled in the sequence $(i_1, \ldots, i_p)$ and the allowances are given by the vector $\tilde{x}$. Then,

$$\phi(\tilde{x}; \{i_1, \ldots, i_p\}) = \sum_{k=1}^{p} c_{ik}^w E[W_k] + \sum_{k=1}^{p} c_{ik}^s E[S_k] + c_i E[L]. \quad (17)$$

The OR manager’s first task is to find allowances $\tilde{x}^*[i_1, \ldots, i_p]$ that minimize the expected total cost, $\phi(\tilde{x}; \{i_1, \ldots, i_p\})$, for each fixed $(i_1, \ldots, i_p)$. Let function $\phi(\tilde{x}^*[i_1, \ldots, i_p])$ denote the minimum expected cost when the sequence $(i_1, \ldots, i_p)$ is chosen. That is,

$$\phi(\tilde{x}^*[i_1, \ldots, i_p]) = \min_{\tilde{x}} \{ \phi(\tilde{x}; \{i_1, \ldots, i_p\}) \}. \quad (18)$$

Finally, the problem of finding the optimal sequence $\{i^*_p \}_{i=1}^p$ can be written as

$$\{i^*_p \}_{i=1}^p = \arg \min_{\{i_p \} \in \mathcal{S}_p} \phi(\tilde{x}^*[i_1]). \quad (19)$$

This problem is difficult because of two reasons: (1) the set $\mathcal{S}_p$ grows rapidly with $p$ (note that the size of $\mathcal{S}_p$ is $p!$), and (2) the optimal allowances do depend on the sequence in which surgeries are performed. The procedure for determining optimal allowances is algorithmic and computationally intensive (see Denton and Gupta 2003 for details). The number of surgeries that could be performed in a particular session of the OR vary a great deal from one subspecialty to another. However, for many subspecialties, a typical session of an OR may be able to accommodate surgeries at most in the double digits.

### 4.2. Solution Approaches, Extensions, and Open Challenges

The problem formulated in Section 4.1 is a combinatorially hard problem. It is further complicated by the fact that optimal start times depend on the sequence and that finding optimal start times requires intensive computations. Therefore, we will demonstrate an approach that exploits properties of stochastically ordered surgery duration vectors. We make certain simplifying assumptions that $p = 2$, $d = 0$, $c_{i1}^w = c_{i2}^w$, and $c_{i1}^s = c_{i2}^s$ for all $i$. The last assumption implies that all surgeries have the same waiting and idling penalty, whereas $d = 0$ implies that the OR manager attempts to minimize the OR completion time. Of all these assumptions, the assumption that $p = 2$ is a serious limitation of our approach because, in reality, the number of surgeries scheduled in a single session of an OR can easily exceed 2. The assumption that $d = 0$ is not as serious a limitation from a practical viewpoint. OR managers may accept a methodology for finding a sequence that minimizes completion time because then any slack can be utilized to accommodate unanticipated delays and urgent cases. Also, $c_{i1}^w > c_{i2}^w$ and $c_{i1}^s > c_{i2}^s$ are likely to be assumed in many practical situations anyway because OR managers do not typically have data upon which to base choices of different values of these parameters for different types of surgeries.

Our goal is to identify stochastically ordered surgery durations for which the optimal sequence can be identified without performing any calculations. However, as we demonstrate below, this approach works only in a limited number of cases, leaving the vast majority of the cases as a future challenge. We begin by reviewing some definitions from the theory of stochastic comparisons (see Shaked and Shanthikumar 1994 or Müller and Stoyan 2002 for extensive background). For random variables, $A$ and $B$ with distribution functions $F_A$ and $F_B$, we say that $A$ is stochastically smaller than $B$ in the usual order, written as $A \preceq_x B$, if $E\phi(A) \leq E\phi(B)$ for all non-decreasing functions $\psi$ for which the expectations exist. If a similar inequality holds for all convex (equivalently, increasing convex) functions, then we say that $A$ is smaller than $B$ in the convex (equiv. increasing convex) order. These orders are denoted $A \preceq_{cx} B$ and $A \preceq_{icx} B$. It can be demonstrated that $A \preceq_{cx} B$ is equivalent to the condition that $F_B(x) \leq F_A(x)$ for each $x$. Similarly, $A \preceq_{cs} B$ implies that $A$ and $B$ have the same mean, but $A$ has a smaller variance. If $A \preceq_{cs} B$, then $A \preceq_{icx} B$, but the opposite is not, in general, true. $A \preceq_{icx} B$ implies that the random variable $A$ is stochastically smaller than another variable $B$, which is less variable than $B$. That is, loosely speaking, $A \preceq_{icx} B$ occurs when $A$ is smaller in both magnitude and variability than $B$.

With $p = 2$, there are only two possible sequences, $(1, 2)$ or $(2, 1)$, with the result that we can exhaustively compare all sequences. The expected costs corresponding to the sequence $(1, 2)$ are

$$\phi(x_{12}; \{1, 2\}) = c_{12}^w E[(Z_1 - x_1)^+] + c_{12}^s E[(x_1 - Z_1)^+] + c_{12}^E E[Z_2 + (x_1 \lor Z_1)] = (c_{12}^w + c_{12}^s + c_{12}^E) E[(Z_1 - x_1)^+]$$

$$- (c_{12}^w + c_{12}^s) E(Z_1 - x_1) + c_{12}^s E(Z_1 + Z_2) = (c_{12}^w + c_{12}^s + c_{12}^s) E[Z_1 + x_1(c_{12}^w + c_{12}^s) + c_{12}^s E Z_2]. \quad (20)$$

Similarly,

$$\phi(x_{21}; \{2, 1\}) = (c_{21}^w + c_{21}^s + c_{21}^E) E[(Z_2 - x_2)^+] - c_{21}^s E Z_2 + x_2(c_{21}^w + c_{21}^s) + c_{21}^s E Z_1. \quad (21)$$

**Proposition 1.** If $Z_1 \preceq_{cx} Z_2$, then the sequence $(1, 2)$ is optimal.

**Proof.** We must show that $\phi(x_{21}^*[1, 2]) \leq \phi(x_{12}^*[2, 1])$. This is proved through the following series of arguments.
\(\phi(x_2^{*}[1, 2]) \leq \phi(x_2^{*}[1, 2]) = (c^w + c^t + c')
\times E[(Z_1 - x_1^{*})^+ - c'EZ_1 + x_2^t(c^t + c') + c'EZ_2
\leq (c^w + c^t + c')E[(Z_2 - x_2^{*})^+] - c'EZ_2
+ x_2^t(c^t + c') + c'EZ_1 = \phi(x_2^{*}[2, 1]) \) (22)

The first inequality above is trivial (follows from optimality of \(x_1^t\)), whereas the second inequality comes from the fact that for each realized value \(z\) of the random variable \(Z\) and \(x \in \mathbb{R}\), the functions \((z - x)^+\) are convex functions. Therefore, if \(Z_1 \leq_{x_1} Z_2\), it follows that \(E[(Z_1 - x)^+] \leq E[(Z_2 - x)^+]\) for each \(x\) and \(E[Z_1] = E[Z_2]\). Hence proved. \(\Box\)

Within uniform, gamma, Weibull, lognormal, and beta families of distributions, if the parameters of two distributions (chosen from the same family) are such that both have the same mean and one has a smaller variance, then the other with the smaller variance is smaller in the convex order. That is, from a practical viewpoint, Proposition 1 implies that it is optimal to schedule procedures with smaller variance first.

In related previous work, Weiss (1990) conjectured that sequencing surgeries in order of increasing variances is a good idea, perhaps even optimal. The intuition behind this idea might have been that, by being last, “problematic” (high variance) surgeries cannot affect the start times of other surgeries. Yano and Zhou (1994) proposed a similar heuristic. Lebowitz (2003) argues that scheduling short procedures first limits variability and leads to a better operating room schedule. Robinson, Gerchak, and Gupta (1996) showed that in the case of two surgeries, Weiss’ conjecture is indeed true for location-scale duration distributions (e.g., normal, uniform). We have further generalized this result to all convexly ordered surgery durations.

Can a weaker ordering of surgery durations lead to an optimal sequence? A possible choice here is the increasing convex order. However, note from the right-hand side of (20) that the expected cost is decreasing in \(EZ_1\) and increasing in \(EZ_2\). Therefore, the impact of changing the sequence cannot be determined without specifying the mean surgery durations. This is somewhat surprising because for the purpose of finding the optimal surgery start times for a given sequence, Robinson and Chen (2003) have demonstrated that the mean surgery durations can be netted out of the relevant optimization problem.

In other enhancements, it is easy to see from Proposition 1 that if \(Z_1 \leq_{x_1} Z_2\) and \(c^t_2 \leq c^t_1\), the sequence \([1, 2]\) continues to be the optimal sequence. That is, the OR manager should continue to schedule the less variable surgery first when it also has a higher waiting cost. Our attempts to further generalize the results shown here, e.g., to cases involving \(p > 2\), have not been fruitful even after placing additional restrictions on the \(Z_i\)’s. Thus, finding the optimal sequence in which surgeries should be performed remains an important open problem.

5. Concluding Remarks

In addition to issues discussed in the previous three sections of this paper, there are a variety of problems related to surgical suites’ management that have not been addressed in the OM literature. An important problem is the decision concerning specialization of ORs, which we describe below.

Different surgical procedures take different amounts of time and have inherently different variability associated with completion times. Mixing different types of surgeries or allowing urgent cases to be added to the daily schedule of an OR can increase variability, which leads to more schedule changes (cancellations and postponement) and more overtime use. Schedule changes can also increase the difficulty of staffing downstream resources such as PACU or hospital beds. For these reasons, some have suggested that specializing certain ORs to handle certain types of procedures can increase efficiency (Landro 2005). As mentioned earlier, specialization can be either by surgery type or by the urgency of the surgery within each type. However, there are no model-based guidelines on when it makes sense for a hospital to specialize its ORs either by medical subspecialty or by urgency of care. This problem is particularly difficult because the decisions also depend on how well the allotted OR time is utilized and how surgeries are scheduled.

Surgeons with operating privileges in a hospital also maintain office practices where they provide consultations, determine whether the patient is best treated via a surgical procedure, and order or review medical tests to check a patient’s fitness for undergoing the procedure. These practices are housed in specialty-service clinics that may be attached to the hospitals. The author has observed that some surgeons prefer to spend time in the OR rather than hold clinic hours. Therefore, when unanticipated OR time becomes available due to cancellations or lack of demand for someone else’s block, it is not uncommon for some surgeons to opportunistically cancel clinic hours and perform surgeries (usually this happens 1 or 2 days in advance of the scheduled clinic time). This practice leads to a better utilization of OR time, but causes considerable difficulties for the clinic managers who must scramble to reschedule appointments. This suggests that management rules around how unused capacity may be utilized can have a significant impact on hospital operations. Dexter and Macario (2004) carry out an empirical study of the effect of releasing allocated OR time on efficiency. They argue that when the OR time is released has negligible effect on OR efficiency. However, there are no model-based approaches for addressing this problem. Moreover, Dexter...
and Macario (2004) did not take into account possible effects on clinic efficiencies and staff planning for downstream resources.

Finally, we wish to draw attention to the dearth of models that deal with day-of-surgery decisions. As the day unfolds, a variety of decisions can affect OR efficiency, waiting times, and patient safety. Examples of such decisions include the allocation of ORs to scheduled procedures, cancellations and postponement of procedures, and scheduling overtime. The OM literature is largely silent on this class of problems, although there have been some attempts to understand the problem via statistical analysis (see, e.g., Dexter et al. 2004). Attempts to provide model-based solutions to such problems can have a significant impact on OR utilization and efficiency.

Acknowledgments

The author is grateful to an anonymous referee for constructive feedback on an earlier version of this paper.

References


Dexter, F., A. Macario. 2004. When to release allocated operating room time to increase operating room efficiency. Anesthesia and Analgesia 98(5) 758–162.


Weiss, E. N. 1990. Models for determining estimated start times and case orderings in hospital operating rooms. IIE Transactions 22(2) 143–150.