Health Care Manage Sci
DOI 10.1007/s10729-006-9005-4

Optimization of surgery sequencing and scheduling decisions under uncertainty

Brian Denton · James Viapiano · Andrea Vogl

Received: 28 April 2006 / Accepted: 12 October 2006

Abstract Operating rooms (ORs) are simultaneously the largest cost center and greatest source of revenues for most hospitals. Due to significant uncertainty in surgery durations, scheduling of ORs can be very challenging. Longer than average surgery durations result in late starts not only for the next surgery in the schedule, but potentially for the rest of the surgeries in the day as well. Late starts also result in direct costs associated with overtime staffing when the last surgery of the day finishes later than the scheduled shift end time. In this article we describe a stochastic optimization model and some practical heuristics for computing OR schedules that hedge against the uncertainty in surgery durations. We focus on the simultaneous effects of sequencing surgeries and scheduling start times. We show that a simple sequencing rule based on surgery duration variance can be used to generate substantial reductions in total surgeon and OR team waiting, OR idling, and overtime costs. We illustrate this with results of a case study that uses real data to compare actual schedules at a particular hospital to those recommended by our model.

Keywords Surgery · Scheduling · Sequencing · Stochastic · Optimization · Linear programming · Simulation

B. Denton (✉)
Mayo Clinic College of Medicine, 200 First St. SW,
Rochester, MN 55905, USA
e-mail: denton.brian@mayo.edu

J. Viapiano · A. Vogl
Fletcher Allen Health Care, 111 Colchester Ave,
Burlington, VT 05401, USA

1 Introduction

Operating rooms (ORs) have been estimated to account for more than 40% of a hospital's total revenues [1] and a similarly large proportion of their total expenses, which makes them a hospital's largest cost center as well as its greatest revenue source. Recent studies indicate that OR performance measures, such as utilization, overtime, and on-time start performance are well below achievable targets at most hospitals [2]. Therefore they offer the potential for significant improvements in operational efficiency. The purpose of this article is to present results from a study of a stochastic optimization model for daily scheduling of a single OR. We present a two-stage stochastic recourse model and numerical results illustrating its application to real-world problems encountered at a large hospital. We begin by providing some background on OR scheduling.

There are several different environments in which surgical services are delivered. Hospitals provide a broad range of services including an emergency department for handling cases resulting from unpredictable adverse events. Surgery at hospitals may be on an *inpatient* or *outpatient* basis. In the inpatient setting patients are admitted to the hospital prior to surgery and assigned a hospital bed. After their scheduled surgery and postoperative recovery they are returned to their room for the remainder of their recovery. Outpatients, on the other hand, arrive at the hospital on the day of the surgery. After surgery outpatients are held until recovery is complete and then they are released from the hospital. More recently, new delivery systems called Ambulatory Service Centers (ASCs) have emerged [3]. ASCs service elective (scheduled) surgeries that can be

performed safely in an outpatient setting with minimal supporting resources. More complex surgeries that require inpatient services and possibly other supporting services (e.g. emergency services) are performed at hospitals.

Whether surgery is performed on an inpatient or outpatient basis, at an ASC or hospital, many operational aspects of the OR scheduling problem remain the same. ORs in both environments have very high fixed costs, and in many cases the large proportion of the cost is associated with the labor cost of the OR team. Typically ORs have a planned utilization time (e.g. 8 h) beyond which overtime costs for some members of the OR team begin to accrue. Therefore on-time surgery starts are important since late starts result in the OR team waiting, and increase the likelihood of overtime which results in higher direct surgery costs and OR team fatigue. At many hospitals and ASCs the OR availability is limited and therefore there is heightened emphasis on scheduling as many cases in a day as is safe and cost effective. While some surgeries have relatively predictable durations others may have significant (and unavoidable) uncertainty in their duration. The combination of tight schedules and uncertainty in duration creates the need for careful consideration of OR schedules to balance the competing criteria of OR Team waiting, OR idling, and overtime.

There are two well known surgery scheduling systems, *block- scheduling* and *open-scheduling*. Under a block-scheduling system individual surgeons or surgical groups are assigned times in a particular OR in a periodic schedule (typically weekly or monthly). Surgeons book cases into their assigned time subject to the condition that the mean duration of the cases fit within the scheduled time period. For cases that do not fit, the surgeon must request an allowance to overbook. On the other hand, in open scheduling systems the intention is to accommodate all patients. The surgeons submit cases up until the day of surgery and all cases are scheduled in ORs. Individual surgeries are allocated to ORs to create a schedule prior to the day of surgery.

The model we present in this article considers the scheduled time for surgical cases as well as the sequence of cases in a particular OR on a particular day. We limit the scope of decisions to a particular OR/day combination and we do not consider reassigning procedures from one day to another, or from one OR to another. Therefore our model is applicable to both block-scheduling and open-scheduling systems described above. Our model extends the two-stage stochastic linear programming model first presented in Denton and Gupta [4]. Their model was used to compute scheduled time for cases given that the sequence of surgeries is known.

In this article we relax the assumption of a fixed sequence to study a more realistic version of the problem. Based on numerical experiments using real surgery duration data, we compare optimal schedules to actual schedules and show that sequencing decisions also play a significant role in scheduling decisions. We present the results from testing several heuristics including (a) a pairwise interchange heuristic that takes advantage of lower bounds on the optimal solution and (b) heuristics that use the mean and variance of the surgery durations to select a sequence. Our results indicate that the common practice of scheduling longer and more complex cases earlier in the daily schedule may have a significant negative impact on OR performance measures.

The remainder of the article is organized as follows. In the next section we provide a brief review of the literature relating to OR scheduling. Next, in Section 3, we discuss the model and solution methodology used for optimizing surgery schedules. We describe some structural properties of our model, and discuss several heuristics for computing OR schedules. In Section 4 we present the results of numerical experiments based on real data. In Section 5 we discuss some practical challenges to implementing the model. Finally, in Section 6 we summarize our findings and point out future research directions.

2 Literature review

More general reviews than the following can be found in Blake and Carter [5], Przasnyski [6], and Magerlein and Martin [7]. Several papers present quantitative models to allocate time for customers arriving to a stochastic server. Sabria and Daganzo [8] consider the problem of scheduling arrivals of cargo ships at a seaport. Wang [9] discusses the problem in a manufacturing context in which the arrival of parts are scheduled on the shop floor. In the health care context there have also been several articles presenting heuristics for assigning appointments for arrivals at outpatient clinics (for example Bailey [10], Soriano [11], Mercer [12], Charnetski [13], Ho and Lau [14] and references therein). Weiss [15] and Denton and Gupta [4] propose stochastic optimization models for determining OR schedules. Strum, Vargas, and May [16] describe an application of a news-vendor model as a heuristic for determining the planned OR schedule duration to allocate for surgical subspecialties. They fit

probability distributions to historical patterns of surgical demand and combine them with the newsvendor model to minimize costs associated with underutilization and overutilization of OR time. In contrast, in our model we assume the total OR schedule duration is known in advance and is an input into our model.

The problem of simultaneously sequencing and scheduling surgical procedures that we consider in this article is a combinatorial stochastic optimization problem. Few articles have considered both of these factors in constructing OR schedules. Weiss [15] studied sequencing decisions in the two-surgery context and showed, using stochastic dominance arguments, that for certain selective choices of distributions the optimal solution is in order of increasing variance of service durations. Wang [9] conjectured that the optimal rule for $n \geq 3$ customers is also in order of increasing variance; however, no proof is given beyond the case of $n = 2$. Vanden Bosch and Dietz [17] evaluate a pairwise interchange heuristic for sequencing patient arrivals at a medical office with varying service duration distributions. They consider a heuristic based on a discrete lattice of potential service durations, and present encouraging results for small ($n = 6$) problems which consider waiting and idling time costs as the relevant measures. Dexter and Ledolter [18] study prediction bounds for operating room times and consider the effect of sequencing on mean tardiness, pointing out that sequencing less uncertain cases earlier reduces mean patient waiting.

Surgical case sequencing decisions may also have an effect on other activities, such as patient intake and recovery. Dexter and Marcon [19] consider the impact of sequencing on post anesthesia care unit staffing. They find that the *longest case first* rule, often used in practice, performs poorly from a staffing perspective, and that the *shortest case first* rule performs relatively well with respect to a number of other decision rules. As we show in the remaining sections, these results tend to be consistent with our findings about the impact of case sequencing rules on the OR scheduling measures of waiting, idling, and tardiness.

In contrast to the above referenced work, we provide a formulation of the model as a two-stage recourse problem with first stage binary decision variables representing sequencing decisions. While many of the studies above are based on Monte Carlo simulation, our approach seeks to optimize the scheduled time and sequence of surgical cases in an OR. We exploit the structure of the problem to develop and test several easy-to-implement heuristics.

3 Model formulation and methodology

In this section we describe a two-stage stochastic programming model for determining the optimal surgery schedule. In addition to the measures considered in the cited literature, the model we propose also includes *tardiness* (overtime) with respect to a planned OR utilization. Tardiness is important because late completion of surgeries in an OR can have a negative impact on an organization's performance. For instance, depending on the compensation policies, in some cases there can be a direct cost associated with overtime for members of the OR team if the planned completion time is synchronized with a shift end. In other situations, where all members of the OR team are salaried, late completion impacts planned future activities for staff members. For example, if an OR is planned to close at noon a surgeon may have outpatient clinic hours in the afternoon which would be impacted by late closure. In our model we assume that the quality of a schedule can be measured as a weighted sum of the expectation of three measures: waiting time, idling time, and tardiness. We let n denote the number of surgeries to be scheduled in a given session. Uncertainty is denoted by a scenario ω that defines the vector of collective outcomes of the random surgery durations, denoted by $\mathbf{z}(\omega)$, having support $\Xi \subseteq \Re^n$ and probability distribution F on Ξ. Surgery duration random variables are denoted by $z_i(\omega)$ where the subscript i indexes the n surgeries. The decision variable, x_i, defines the scheduled time for surgery i. Note that selecting the scheduled time is equivalent to selecting the start time for surgeries where the first starts at time zero, the second at x_1, the third at $x_1 + x_2$, and so on. The waiting and OR idling time prior to each surgery are represented by $w_i(\omega)$ and $s_i(\omega)$ respectively, where each depends on the random surgery durations through the scenario, ω. Based on this notation the waiting and idling times can be written as the follow recursive functions.

$$w_i(\omega) = \max(w_{i-1}(\omega) + z_{i-1}(\omega) - x_{i-1}, 0), \quad i = 2,...,n, \tag{1}$$

$$s_i(\omega) = \max(-w_{i-1}(\omega) - z_{i-1}(\omega) + x_{i-1}, 0), \quad i = 2,...,n. \tag{2}$$

It is assumed that the first surgery starts on time, i.e., $w_1(\omega) = s_1(\omega) = 0$. Tardiness, denoted by ℓ, is measured with respect to the planned duration for which the OR will be utilized, d, and can be written as

$$\ell(\omega) = \max\left(w_n(\omega) + z_n(\omega) + \sum_{i=1}^{n-1} x_i - d, 0\right)$$

Waiting, idling, and tardiness are all functions of the random surgery durations. Given these definitions the stochastic optimization problem can be written as minimization of the weighted sum of the expectation of waiting, idling, and tardiness, as follows

$$Z = \min_{\mathbf{x}} \left\{ \sum_{i=1}^{n} c_i^w E[w_i(\omega)] + \sum_{i=1}^{n} c_i^s E[s_i(\omega)] + c^\ell E[\ell(\omega)] \right\}. \tag{3}$$

Since the expectations in Eq. 3 are over multiple random variables, evaluation of the objective function is computationally challenging, let alone computing the optimum. Denton and Gupta [4] exploit structural properties of the problem to develop fast solution methods based on an L-shaped decomposition method. We discuss this further in the next section. For now we describe a special case in which Eq. 3 is easy to solve.

3.1 Two surgery model

The special case of Eq. 3 in which $n = 2$ and $c^\ell = 0$ was analyzed by Weiss [15]. For this case it is only necessary to determine the scheduled time for the surgery that is selected to be first since the other case comes second by default. Thus, for simplicity we drop subscript i from x_i, c_i^w, and c_i^s, in the following discussion. Weiss showed the case of $n = 2$ corresponds to the *newsvendor problem* and therefore a closed form expression for the optimal surgery allowance for the first job (the job allowance for the second job is immaterial) is known. Letting $F_i(\cdot)$, $i = 1, 2$, denote the cumulative distribution function (c.d.f.) of the two surgeries, and $\bar{F}_i(\cdot) = 1 - F_i(\cdot)$, the problem for $n = 2$ can be written as follows:

$$Z = \min_{x} \left\{ c^w E_\omega[w_2(\omega)] + c^s E_\omega[s_2(\omega)] \right\}, \tag{4}$$

where

$$\begin{aligned} E_\omega[w_2(\omega)] &= \int_0^\infty (z_i(\omega) - x)^+ dF_i(\omega) \\ &= \int_x^\infty z_i(\omega) dF_i(\omega) - x\bar{F}_i(x) \end{aligned}$$

and

$$\begin{aligned} E_\omega[s_2(\omega)] &= \int_0^\infty (x - z_i(\omega))^+ dF_i(\omega) \\ &= -\int_0^{x_i} z_i(\omega) dF_i(\omega) + xF_1(x_1). \end{aligned}$$

where $E_\omega[w_2(\omega)]$ and $E_\omega[s_2(\omega)]$ both depend on which surgery is selected to be first $i = 1$ or 2. The optimal allowance, x, is the solution to the newsvendor problem obtained as follows:

$$x^* = F_i^{-1}\left\{ \frac{c^w}{c^w + c^s} \right\}.$$

Weiss described properties of this problem related to sequencing including showing that when a convex ordering exists between surgeries it is optimal to sequence surgeries according to that ordering. In our study of the model we extend the objective function to include expected tardiness, in addition to expected waiting and idling. While a closed form expression for x^* is not available in this case, it is straightforward to extend Weiss's original sequencing argument to our model.

Proposition 1 *If* $z_1(\omega) \leq_{cx} z_2(\omega)$ *then the sequence* $\{1, 2\}$ *is optimal.*

Proof Let the objective function for sequence $\{i, j\}$ be denoted by $Z(\mathbf{x}; \{i, j\})$. Since $z_1(\omega) \leq_{cx} z_2(\omega)$ and x_1^*, x_2^*, are the optimal solutions for sequences $\{1, 2\}$ and $\{2, 1\}$ respectively, it follows that

$$\begin{aligned} Z(x_1^*; \{1, 2\}) &\leq Z(x_2^*; \{1, 2\}) \\ &= c^w E[(z_1(\omega) - x_2^*)^+] + c^s E[(x_2^* - z_1(\omega))^+] \\ &\quad + c^\ell E[(z_1(\omega) + z_2(\omega) + (x_2^* - z_1(\omega))^+ - d)^+] \\ &\leq c^w E[(z_2(\omega) - x_2^*)^+] + c^s E[(x_2^* - z_2(\omega))^+] \\ &\quad + c^\ell E[(z_2(\omega) + z_1(\omega) + (x_2^* - z_2(\omega))^+ - d)^+] \\ &= Z(x_2^*; \{2, 1\}). \end{aligned}$$

□

Where the last inequality follows from the convex ordering and convexity of the expectation of waiting, idling, and tardiness times.

The above proposition establishes, for certain special cases, the optimal sequence of surgeries. To the authors knowledge no such results for the appointment scheduling problem for $n > 2$ are known. However, we can use the insight from Proposition 1 to motivate heuristics for $n > 2$. In the next section we present a stochastic program which captures decisions about scheduled surgery time and sequencing decisions. We describe some easy-to-implement heuristics for sequencing surgeries based on the model, and in Section 4 we compare the results of the heuristics to optimal solutions based on actual surgery schedules.

3.2 Stochastic programming formulation

The special case in which sequence is determined a priori can be formulated as a two- stage recourse problem. In this article we assume a discrete finite set of

scenarios, denoted by $\{\omega_k \mid k = 1, ..., K\}$, that are representative of the uncertainty in surgery durations, such as would result from statistical sampling. Given this discrete set of scenarios we can write the deterministic equivalent of the two-stage recourse problem as the following *sample average approximation* [20].

$$Z = \min \left\{ \sum_{k=1}^{K} \frac{1}{K} \left(\sum_{i=2}^{n} c_i^w w_i(\omega_k) + \sum_{i=2}^{n} c^s s_i(\omega_k) + c^\ell \ell(\omega_k) \right) \right\} \tag{5}$$

$$s.t. \quad -w_i(\omega_k) + w_{i+1}(\omega_k) - s_{i+1}(\omega_k) = z_i(\omega_k) - x_i, \quad \forall (i, \omega_k) \tag{6}$$

$$-w_n(\omega_k) + \ell(\omega_k) - g(\omega_k) = z_n(\omega_k) - d + \sum_{j=1}^{n-1} x_j, \ \forall \omega_k \tag{7}$$

$$x_i \geq 0, \ \forall i, \ w_i(\omega_k) \geq 0, \ s_i(\omega_k) \geq 0, \ \forall (i, \omega_k) \ \ \ell(\omega_k), g(\omega_k) \geq 0, \ \forall \omega_k \tag{8}$$

The first constraint balances waiting and idling times with respect to the actual and scheduled time for surgery. Similarly, the second constraint balances the overtime based on the completion time of the last surgery and the planned OR utilization, d. The above formulation includes an additional slack variable g which denotes the *earliness* with respect to the planned OR utilization. It is necessary for an accurate formulation, however, earliness is not explicitly penalized in the objective function.

Relaxing the assumption that the sequence is predetermined makes the problem combinatorial in nature (with $n!$ sequences if all surgeries are distinct). This relaxed version of the problem can also be formulated as a two-stage recourse problem. However, the formulation includes first stage binary decision variables representing sequencing decisions which must be defined in advance of knowing the outcomes of the random surgery durations. The following is the corresponding two-stage stochastic mixed-integer-program

$$Z(\mathbf{x}^*, \mathbf{o}^*) = \min \left\{ \sum_{k=1}^{K} \frac{1}{K} \left(\sum_{i=1}^{n} \sum_{i'=1}^{n} c_{ii'}^w w_{ii'}(\omega_k) + c_{ii'}^s s_{ii'}(\omega_k) + c^\ell \ell(\omega_k) \right) \right\} \tag{9}$$

$$s.t. \quad w_{ii'}(\omega_k) - M_1(\omega_k) o_{ii'} \leq 0 \quad \forall (i, i', k) \tag{10}$$

$$s_{ii'}^k - M_1(\omega_k) o_{ii'} \leq 0 \quad \forall (i, i', k) \tag{11}$$

$$\sum_{i'=1}^{n} o_{ii'} \leq 1 \quad \forall i \tag{12}$$

$$\sum_{i=1}^{n} \sum_{i'=1}^{n} o_{ii'} = n - 1 \tag{13}$$

$$-\sum_{i'=1}^{n} w_{i'i}(\omega_k) + \sum_{i'=1}^{n} w_{ii'}(\omega_k) - \sum_{i'=1}^{n} s_{ii'}(\omega_k) + x_i = z_i(\omega_k) \quad \forall (i, \omega_k) \tag{14}$$

$$\sum_{i=1}^{n} z_i(\omega_k) + \sum_{i=1}^{n} \sum_{i'=1}^{n} s_{ii'}(\omega_k) - \ell^k(\omega_k) + g^k(\omega_k) = d \tag{15}$$

$$o_{ii'} \in \{0, 1\} \ \forall (i, j), x_i \geq 0, \forall_i \ \ell(\omega_k), g(\omega_k) \geq 0 \ \forall \omega_k, \ w_{ii}(\omega_k), s_{ii}(\omega_k) \geq 0 \ \forall (i, i', \omega_k) \tag{16}$$

In the above two-stage stochastic mixed-integer-program $o_{ii'}$ and x_i are the first stage decision variables and $w_{ii'}$, $s_{ii'}$, ℓ, and g are the second stage variables. The variable $o_{ii'}$ is a binary decision variable representing sequencing decisions where $o_{ii'} = 1$ if surgery i directly precedes i' and 0 otherwise. Waiting and idling variables, $w_{ii'}$ and $s_{ii'}$, are now indexed by multiple surgeries, and constraints (10) and (11) require that waiting and idling times between surgeries be zero unless the surgeries are consecutive. Multiple indices are required for $w_{ii'}$ and $s_{ii'}$ in this formulation because waiting and idling are sequence dependent and the optimal sequence is not known in advance. Thus waiting and idling may be non-zero between any surgeries and, furthermore, they are constrained to be zero between all surgeries except those that are consecutive. Similarly, cost coefficients for waiting and idling, $c^w_{ii'}$ and $c^s_{ii'}$ respectively, depend on the sequence of surgeries. Thus, for example, waiting time between two surgeries performed by different OR Teams will have a higher waiting cost since in such instances the OR Team and a patient waits, as opposed to just the patient waiting. Therefore, setting a high cost for OR Team waiting will tend to result in contiguous cases for each team. Constraints (12) and (13) are sequence feasibility constraints enforcing the fact that all surgeries must be sequenced, and that subtours are not permitted. Finally, constraints (14) and (15) balance waiting/idling and tardiness/earliness respectively. The inclusion of sequencing decisions in the above model makes it considerably more complex. Therefore, in the next sections we propose and test several heuristics based on the model.

3.3 Heuristics

The stochastic mixed-integer-program of the previous section is presumably NP-hard, and the combination of stochastic and combinatorial elements of the problem makes it particularly difficult to solve. We consider several simple heuristic rules for approximating the optimal solution and demonstrate their value by comparison with actual schedules used in practice. The following heuristics are motivated by the convex ordering property we discussed in Section 3.1.

Heuristic 1 (H1): Sequence surgeries within each surgeons block of cases in order of increasing mean of durations. Compute the optimal scheduled time as the solution to Eqs. 5–7, and 8.

Heuristic 2 (H2): Sequence surgeries within each surgeons block in order of increasing variance of durations. Compute the optimal scheduled time as the solution to Eqs. 5–7, and 8.

Heuristic 3 (H3): Sequence surgeries within each surgeons block in order of increasing coefficient of variation of durations. Compute the optimal scheduled time as the solution to Eqs. 5–7, and 8.

The above heuristics resequence each surgeon's cases individually based on first and second moment information, thus retaining contiguity of each surgeon's cases. Maintaining contiguity of cases minimizes setup costs that would result from interchanging OR teams. Thus, heuristics $H1$ and $H2$ generate orderings that are identical to a convex ordering, if such an ordering exists. Heuristic $H3$ captures both mean and variance information to generate an ordering. In Section 4 we compare $H1$–$H3$ to the optimal solution for several instances of Eqs. (9–14, and 15) that can be solved to optimality via total enumeration.

For model instances that are too large for total enumeration we compare $H1$–$H3$ to an *interchange heuristic*, HI. Heuristic HI is a local search that starts with an initial sequence (corresponding to the actual sequence used in practice) and computes optimal scheduled times by solving Eqs. 5–7, and 8. At each iteration HI investigates randomly generated pairwise interchanges, accepting an interchange each time it makes an improvement in the solution. By maintaining a feasible choice of $o_{ii'}$ it is possible to take advantage of a fast method for generating optimality cuts in an L-shaped implementation to solve the restricted (fixed $o_{ii'}$) version of formulation (9–14, and 15) (see [4] for details). Furthermore, heuristic HI can take advantage of the lower bound generated from the solution of the master problem at each iteration of the L-shaped method. Thus, if the current bound at any given iteration is worse than the best solution so far, the sequence can be abandoned without fully solving the 2-SLP. With a good initial bound this can significantly reduce average computation time per iteration for the interchange heuristic. Following is a description of the algorithm. Index j denotes the iteration with respect to pairwise interchanges, and index v denotes iterations (cut generation) for the L-shaped algorithm for a particular sequence. The algorithm stops when the maximum number of interchanges, J, have been explored.

3.3.1 Interchange heuristic (HI)

Step 1. Initialize $o_{ii'}$ to a feasible surgery sequence. Set $j = v = 1$, $Z^{UB} = \infty$, and $Z^v = 0$. Initialize the master problem.

Step 2. Add an optimality cut to the current master problem and solve to obtain Z^{v+1}. Set $v = v + 1$.

Step 3. If $Z^v > Z^{UB}$ then set $j = j + 1$, $v = 1$, reinitialize the master problem by removing all optimality cuts, generate a new feasible sequence, $o_{ii'}$, using random pairwise interchange, and return to Step 2.

Step 4. If $Z^v - Z^{v-1} = 0$ set $Z^{UB} = Z^v$. If $j = J$ then stop. Otherwise generate a new feasible sequence, $o_{ii'}$, using random pairwise interchange, reinitialize the master problem, and return to Step 2.

Step 5. Return to Step 2.

Steps of HI such as *initializing the master problem*, and adding an *optimality cut* to the master problem are the same as the steps of the L-shaped method for solving two-stage stochastic linear programs [21].

4 Numerical results

In this section we present numerical experiments based on real surgery scheduling data collected at Fletcher Allen Health Care, a non-profit academic health center serving Vermont and upstate New York. Approximately 20,000 surgical procedures are performed at Fletcher Allen each year. The main hospital campus had 12 operating rooms that were scheduled for outpatient and inpatient surgical procedures in 2004, and the operating schedule was based on a 5-week block schedule. Booking is done through a centralized operating room scheduling center and surgeons specify the sequence for their individual cases. Collected data for this study includes:

- The actual weekly surgery schedule for all ORs in 2004
- Historical surgery duration data for all surgeries scheduled in the ORs in 2004
- Objective function weights based on input from OR scheduling decision makers

We describe the results of an empirical evaluation of confidence intervals for the optimal schedule of a single OR, comparisons of the optimal to actual schedules for 50 model instances, and the results of sensitivity analysis for cost coefficients in the objective function. Model instances were selected to present a broad range of different settings including different numbers and types of surgeries. However, the choices were biased towards cases with a sufficient sample size to construct scenarios for the stochastic program. There were approximately 250 different types of surgical procedures (based on aggregations of CPT codes) performed during the period of our analysis, with an average of 21 samples per surgery type. Scenarios were generated by sampling with replacement from the historical data set of individual surgery procedure durations. Sampling was based on the surgery type alone, and not specific to an individual surgeon. Therefore our results do not control for systematic differences between different surgeons.

4.1 Numerical experiments for a single OR

We begin the presentation of our results with an analysis of a 1-week schedule for a single OR. The purpose of this analysis is to provide summary statistics for surgery durations within a particular OR (referred to as OR1 from this point on), and empirical justification of the sample size we use in our model. The OR1 test models were selected since they have a range of model instances corresponding to 2, 3, 4, 6, and 12 surgeries for the 5 days respectively (Monday through Friday). The data set for OR1 contained 19 distinct procedure types with an average of 47 observations per surgery type, and a minimum of 11 observations for each surgery type. The coefficient of variation across all surgery types varied from 0.17 to 0.88. For each model instance presented in this section the planned duration of OR utilization, d, is defined as the difference between the actual planned end time of the last case and the planned start time of the first case. For instance, if the actual schedule planned to start the first case at 7:30AM and the last case to end at 4:30AM then d was set to 9 h.

The results presented in this section and Section 4.2 are based on cost coefficient settings that were selected through consultation with a senior decision maker and management engineer involved in managing the operating room schedules during the period of the study. The waiting time cost coefficients, $c^w_{ii'}$, had two values. For cases in which the same OR team was operating on a new patient the waiting cost was set as $c^w_{ii'} = 3$. On the other hand, for surgeries in which the surgeon, OR team, and patient wait for the OR to become available the cost was set to $c^w_{ii'} = 8$ to reflect waiting for additional resources (OR Team and patient). Thus the waiting cost coefficients, $c^w_{ii'}$, incorporate a *setup* cost which promotes contiguous sequencing of surgeries for the same surgeon. Operating room time was highly valued and the cost coefficient for idling the OR was set to $c^s_{ii'} = 8$. Tardiness costs were set to $c_\ell = 4$ in this instance of the model. It is important to note that the length d was not selected as the shift end time (e.g. 8 h) which would, for some team members, correspond to

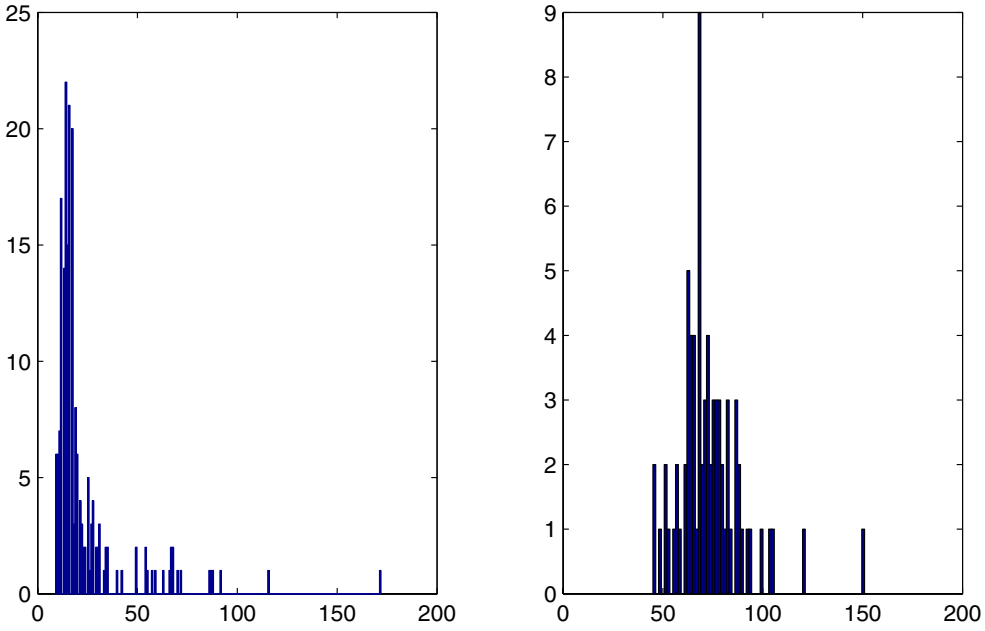

Fig. 1 Illustration of frequency distributions for two surgery types based on a sample of surgery durations

the time at which overtime costs would begin to accrue. Therefore c_ℓ should be viewed as a *penalty* for late completion of surgeries, rather than a precise overtime cost.

Figure 1 illustrates the empirical probability distributions for two specific examples of surgeries in OR1. The structure of these example distributions is typical of uncertainty in surgery durations, where there is a fairly significant mass of probability confined to a predictable range, and a tail indicating a lower probability of extended surgery duration resulting from unexpected complications. Instances of the stochastic linear programming model were created using 10, 000 scenarios. To evaluate the effect of sample size 100 replications of the optimal solution with $K = 10,000$ were performed for each of the 5 daily schedules for the OR1 weekly schedule. The confidence intervals for the optimal solution for OR1 test models ranged from approximately ± 1 to 2.5% relative to the mean. Based on these results we use $K = 10,000$ scenarios for the remainder of our numerical experiments.

4.2 Optimal vs. actual schedules for multiple ORs

In this section we draw comparisons between actual schedules, optimal schedules, and schedules computed using the proposed heuristics. We evaluate a large sample of daily schedules (50 in total) with each one corresponding to a different OR/day combination. The OR/day combinations were selected such that multiple schedules were chosen from each of the ORs, schedules included most of the types of surgeries performed in 2004, and the number of surgeries scheduled varied between 2 and 12 (where 12 was the maximum number scheduled in any OR).

Table 1 compares results for 20 representative samples. In the table we let $\mathbf{x}^0$ and $\mathbf{x}^*$ denote the vectors of actual and optimal scheduled times respectively, and $\mathbf{o}^0$ and $\mathbf{o}^*$ denote the vectors of actual and optimal sequence respectively. The results compare the proposed heuristics to (a) the objective function, $Z(\mathbf{x}^0, \mathbf{o}^0)$, for the actual schedule used in practice for the OR during the period of the study (b) the objective function, $Z(\mathbf{x}^*, \mathbf{o}^0)$, for the optimal scheduled time and the actual sequence and (c) the objective function for the optimal sequence and optimal scheduled time, $Z(\mathbf{x}^*, \mathbf{o}^*)$. Thus, the effects of optimizing scheduled time and sequence of surgeries can be separated.

For the 50 test models considered, 14 had $n > 4$ and were evaluated using the pairwise interchange heuristic rather than total enumeration, due to unacceptable computation time for the latter. The remaining test models were solved to optimality by total enumeration of all sequences. To compare heuristics we define the relative difference between the best solution found, Z^*, and the heuristic, Z^H, as $100 \times (Z^H - Z^*)/Z^*$. Comparing heuristics $H1$–$H3$ to the best solution (the optimal, Z^*, for $n \leq 4$ and the best solution found with HI for $n > 4$) yields the following results

– $H1$, $H2$, $H3$, found the best solution 32, 62, and 58% of the time respectively. For all instances in which $H1$ was optimal, $H2$ was also optimal, i.e, the ordering for $H1$ and $H2$ were the same in these test

Table 1 Numerical example illustrating the effects of optimizing scheduled time and sequence vs. heuristics for an actual OR schedule for varying numbers of surgeries from $n = 2$ to $n = 12$

Model #	n	$Z(\mathbf{x}^0, \mathbf{o}^0)$	$Z(\mathbf{x}^*, \mathbf{o}^0)$	$Z(\mathbf{x}^*, \mathbf{o}^*)$	$H1$	$H2$	$H3$
1	2	886.61	162.73	134.82	162.73	134.82	134.82
2	2	303.01	232.15	232.15	232.15	234.46	234.46
3	2	337.44	297.61	297.61	358.71	297.61	297.61
4	3	759.05	213.87	212.65	213.87	212.65	212.65
5	3	341.33	298.99	248.76	249.26	248.76	248.76
6	3	1,548.85	393.63	245.32	245.32	245.32	245.32
7	4	791.30	436.57	331.07	419.05	331.07	406.68
8	4	806.21	665.89	571.35	571.35	571.35	571.35
9	4	925.61	474.05	304.37	476.49	304.37	304.37
10	5	1,529.66	518.19	430.60	517.96	432.36	430.60
11	5	1,289.29	1,216.11	852.55	976.98	852.55	884.04
12	5	1,290.79	433.06	406.23	485.05	408.12	406.23
13	5	1,173.77	807.36	496.26	597.08	554.66	554.66
14	6	1,209.27	1,058.25	951.00	951.00	951.00	1,000.82
15	6	886.94	686.25	592.10	695.55	592.10	592.10
16	7	1,447.42	1,167.97	886.43	920.89	892.02	910.78
17	7	922.67	386.07	308.34	467.95	308.34	308.34
18	10	2,159.32	931.17	599.96	637.49	603.54	617.93
19	11	1,790.65	737.91	567.10	655.50	567.61	567.10
20	12	1,433.08	1,055.10	868.50	902.38	877.19	910.69

Objective function coefficients: $c^w_{ii'} = 3$ (no OR Team setup between i and i'), $c^w_{ii'} = 8$ (with OR Team setup between i and i'), $c^s_{i'i'} = 8$, and $c^\ell = 4$. Note that for instances where $n > 4$ HI was used to generate the best sequence because these problems could not be solved to optimality in reasonable computation time

models. For all but two test models in which $H3$ was optimal, $H2$ was also optimal.

– The average deviation of $H1$, $H2$, $H3$, from the optimum for all 50 test models is 12.74, 3.57, and 5.58% respectively. The worst case deviation for $H1$, $H2$, $H3$, is , 243, 102, and 102% respectively. These results indicate the performance of $H2$ is quite robust on average, and superior to heuristics $H1$ and $H3$. However, there are specific instances in which the heuristic performs poorly, which is to be expected among a large number of experiments.

Across all test models we found that optimizing scheduled times while leaving the sequence fixed leads to a 40.40% average improvement, with maximum and minimum improvements of 88.57 and 2.76% respectively. Optimizing both scheduled times and sequence results in a 24.17% additional improvement on average, with a maximum and minimum of 243.30 and 0% respectively. In addition to the results presented above we have also found that sequencing alone, without optimizing scheduled time, can have a significant effect, i.e., resequencing cases while assuming the actual scheduled time for cases remains unchanged.

4.3 Cost sensitivity analysis

Recognizing that the calibration of cost coefficients is sensitive to the particular health care environment, and its associated cost structure, we present numerical experiments for the OR1 model instances described above with three choices of cost parameters in Table 2. For each choice we assume idling costs are zero to simplify comparisons between problems with different cost coefficients. This was done for two reasons. First, to differentiate the results from those of Table 1 in which idling costs take nonzero values. Second, because idling and tardiness similarly penalize excess scheduled time for surgeries (note that penalizing tardiness is equivalent to penalizing idling when $d = 0$). Thus, Table 2 corresponds to numerical experiments in which the relative differences between waiting and tardiness costs are set to high/low, equal and low/high, respectively.

The most significant feature of Table 2 is that the relative improvement for the optimal schedule (near optimal in the case of larger instances) is quite large for all choices of cost coefficients. Another significant feature of Table 2 is that heuristic $H2$ outperforms the other heuristics, except in one large instance of the model (Thursday) in which heuristic HI finds a marginally improved sequence. The relative difference between $H2$ and HI in this case is very small (less than 1.1%). Combining these results with the numerical results of Table 1 provides strong evidence that $H2$ is robust with respect to problem size, surgery procedure mix, and cost parameters.

Based on the results in Table 2 it is clear that the relative importance of optimizing scheduled time and sequence depends on the choice of cost coefficients.

Table 2 Numerical example illustrating the effects of optimizing time allocation and sequence vs. heuristics for an actual OR schedule

Cost Coefficients	Heuristic	Monday	Tuesday	Wednesday	Thursday	Friday
$c^w_{ii'} = 1, c^s_{ii'} = 0, c^\ell = 1$	$Z(\mathbf{x}^0, \mathbf{o}^0)$	24.11	28.65	241.03	230.53	131.66
	$Z(\mathbf{x}^*, \mathbf{o}^0)$	11.03	12.66	148.98	142.59	124.66
	$Z(\mathbf{x}^*, \mathbf{o}^*)$	11.03	12.66	100.21	133.08	91.36
	H1	11.03	13.15	100.21	142.77	124.85
	H2	11.03	13.00	100.21	134.61	91.36
	H3	11.03	13.00	100.21	140.77	91.36
$c^w_{ii'} = 1, c^s_{ii'} = 0, c^\ell = 3$	$Z(\mathbf{x}^0, \mathbf{o}^0)$	9.08	35.06	339.80	362.61	262.75
	$Z(\mathbf{x}^*, \mathbf{o}^0)$	3.50	22.12	275.15	275.00	240.21
	$Z(\mathbf{x}^*, \mathbf{o}^*)$	3.50	22.12	225.47	256.48	200.59
	H1	3.50	23.01	225.47	274.73	240.93
	H2	3.50	23.29	225.47	257.65	200.59
	H3	3.50	22.29	225.47	273.07	200.59
$c^w_{ii'} = 3, c^s_{ii'} = 0, c^\ell = 1$	$Z(\mathbf{x}^0, \mathbf{o}^0)$	14.20	79.55	624.32	650.24	312.68
	$Z(\mathbf{x}^*, \mathbf{o}^0)$	10.64	23.07	248.26	291.96	254.38
	$Z(\mathbf{x}^*, \mathbf{o}^*)$	10.64	23.06	143.51	275.89	169.48
	H1	10.64	23.50	143.51	287.58	254.60
	H2	10.64	23.83	143.51	278.89	170.11
	H3	10.64	23.83	143.51	288.39	170.11

Objective function coefficients: $c^w_{ii'} = 1$ (independent of OR Team setup between i and i'), $c^s_{ii'} = 0$, and $c^\ell = 1$. Note that instances for Thursday and Friday, with $n = 6$ and $n = 12$ respectively, were solved using HI

Figure 2 illustrates the difference between the solution with optimal surgery start times, for a particular example (Tuesday for OR1), and assuming that the sequence is fixed a priori to the best and worst sequence. The objective function for the two cases is plotted as a function of α which controls the relative difference between waiting and tardiness cost coefficients as follows

$$c^w_{ii'} = \alpha, \quad c^\ell = 1 - \alpha$$

where α varies from 0 to 1 in Fig. 2. Figure 2 shows that the relative difference is higher when the costs are more evenly matched, and the differences are less substantial when one cost is significantly higher than the other. These results are intuitive for the following reasons. If waiting costs are very high compared to overtime costs then the optimal scheduled times will tend to be chosen so that waiting is minimized, i.e., surgeries will

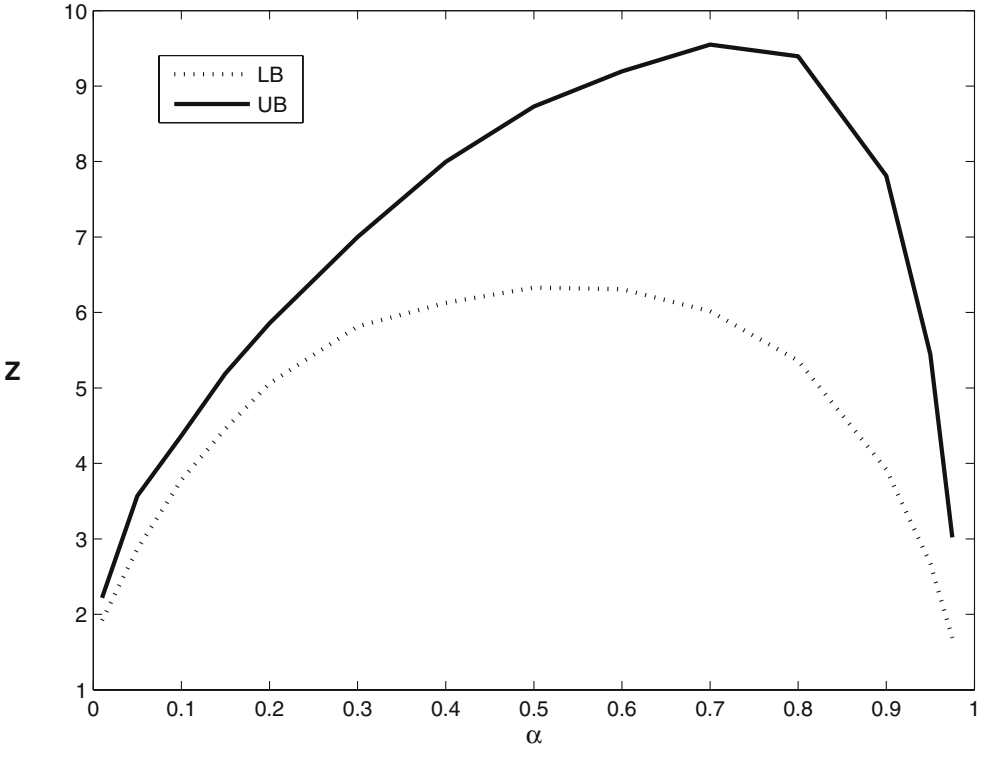

Fig. 2 Illustration of the range in the optimal objective function for the best and worst sequences of surgeries as a function of the relative difference in the waiting cost coefficient, $c^w_{ii'}$ and tardiness cost coefficient, c^ℓ

be scheduled such that they tend to start on time. As a result, surgeries are approximately decoupled, and the importance of sequencing is reduced since each surgery can be approximately treated as independent. On the other hand, if overtime costs are high relative to waiting costs then time will be scheduled such that waiting is likely. Thus surgeries will tend to start immediately with little or no idling between surgeries. In this case the probability distribution for the complete sequence of surgeries can be well approximated by the *convolution* of individual surgery durations. Since the convolution is independent of sequence the effects of sequencing are expected to be low.

5 Practical challenges and open problems

The model we present assumes complete flexibility in making scheduling decisions. In practice there are additional constraints that may affect sequence decisions. For instance, in some situations it is preferable to schedule complex (typically more variable) cases early in the day when more significant resources are available to deal with potential complications. At some facilities *day-of-surgery* cases that will be admitted as inpatients are scheduled later in the day since doing so increases the probability of a hospital bed being available. Also, some cases are preferentially scheduled at certain points in the day (e.g. patients with an infectious disease may be scheduled later in the day to avoid contamination of an OR) and patients that must fast prior to a procedure are preferably scheduled early in the day to avoid the hardship of fasting throughout the day.

The above constraints are based on patient safety and convenience. Additional constraints may arise due to staff and other resource availability constraints. For instance, surgeons must be available at outpatient clinics for pre and post-surgery appointments with patients, thus OR sequencing decisions are coupled with other activities during the day. (These constraints tend to be more significant in block-scheduling than open booking environments, since the latter tend to have surgery and outpatient clinic hours on separate days.) The availability of diagnostic equipment or other finite resources can limit the number of a certain type of procedures that may occur at the same time. Therefore, myopic resequencing of cases in all ORs, without consideration of resource based constraints, may result in infeasible schedules.

The numerical results presented in this paper are based on a relatively comprehensive data set sufficient for modeling probability distribution functions for surgery durations. However, in many environments such historical data may be unavailable or insufficient. When limited data is available subjective estimates of mean and variance in surgery durations may be the only means for implementing the proposed heuristics. Dexter and Ledolter [18] present encouraging results of a bayesian method for developing lower and upper prediction bounds. Combining this approach with the model presented in this article could extend use of this model to environments with less historical data; however, our heuristics were not tested under these assumptions, and therefore the results are an upper bound on the real benefits that could be achieved in such environments.

Based on the results presented in this article we find that the heuristics we present lead to significant improvements when compared to actual schedules used in practice. For the special situation of $n = 2$ we proved a convex ordering of surgeries is optimal, however, more general results for $n > 2$ are not known. Furthermore, no bounds on the quality of the results of the heuristics we present are available, leading to open questions about new heuristics and bounds on the optimal solutions. In addition to new heuristics and bounds for the single OR problem, more general models and methods for optimizing multi-OR schedules represent an open area of research including:

- Consideration of allocation of surgeries to multiple operating rooms to optimize surgery mix, and across multiple days to construct optimal weekly and monthly schedules
- Allocation of scarce resources (e.g. specialized staff, mobile diagnostic equipment)
- Reservation of OR capacity for accommodating urgent and emergent cases (add-ons) that arise on short notice.

6 Conclusions

Based on the analysis presented in this article we draw the following general insights of relevance to OR managers, directors, and surgical planning committees.

- Stochastic models that consider uncertainty in surgery durations offer the potential for significant improvement to daily OR schedules
- Improvements to OR schedules are sensitive to both scheduled time and sequencing decisions
- The effects of optimal sequencing depend on the relative importance of performance measures
- Most of the improvements resulting from resequencing of surgeries can be achieved using the

easy-to-implement heuristic, $H2$, which dominates heuristics $H1$ and $H3$ in nearly all test models

The performance of $H2$ for sequencing decisions is supported by the numerical results presented, and the analytic result for the special situation of $n = 2$ presented in Section 3.1. Heuristic $H2$ seems intuitively reasonable, since positioning high variance surgeries late in the schedule minimizes the potential impact on waiting time for surgeries later in the schedule. In other words, a high variability surgery at the beginning of the day may increase waiting time for the next surgery, as well as all consecutive surgeries. Furthermore, due to its ease of implementation (compared to HI which performs similarly well) heuristic $H2$ demonstrates the best trade-off between solution quality and solution method complexity. Anecdotal evidence indicates this heuristic is not commonly implemented in practice, where it is typical to schedule longer and more variable surgeries earlier in the day. Therefore there is the potential for practical improvements in many surgery scheduling environments. The rule is also directly applicable to other manufacturing and service systems where appointment based models are used.

Acknowledgements The authors are grateful for the comments of three anonymous referees which helped to improve this manuscript. This work was supported in part by National Science Foundation grant DMI-0620573.

References

1. Achieving operating room efficiency through process integration (2005) Technical report, Health Care Financial Management Association Report
2. Surgical services reform: executive briefing for clinical leaders (2001) Technical report, Washington, DC, Clinical Advisory Board
3. Bowers J, Mould G (2005) Ambulatory care and orthopaedic capacity planning. Health Care Manage Sci 8:41–47
4. Denton BT, Gupta D (2003) A sequential bounding approach for optimal appointment scheduling. IIE Trans 35:1003–1016
5. Blake JT, Carter M (1997) Surgical process scheduling: a structured review. J Soc Health Syst 5(3):17–30
6. Przasnyski Z (1986) Operating room scheduling: a literature review. AORN J 44(1):67–79
7. Magerlein JM, Martin JB (1978) Surgical demand scheduling: a review. Health Serv Res 13(4):418–433
8. Sabria F, Daganzo CF (1989) Approximate expressions for queuing systems with scheduling arrivals and established service order. Transp Sci 23:159–165
9. Wang PP (1993) Static and dynamic scheduling of customer arrivals to a single-server system. Nav Res Logist 40:345–360
10. Bailey N (1952) A study of queues and appointment systems in hospital outpatient departments, with special reference to waiting-times. J R Stat Soc A 14:185–189
11. Soriano A (1966) Comparison of two scheduling systems. Oper Res 14:388–397
12. Mercer A (1973) Queues with scheduled arrivals: a correction simplification and extension. J R Stat Soc B 35:104–116
13. Charnetski J (1984) Scheduling operating room surgical procedure with early and late completion penalty costs. J Oper Manag 5:91–102
14. Ho C-J, Lau H-S (1992) Minimizing total cost in scheduling outpatient appointments. Manag Sci 38:750–764
15. Weiss EN (1990) Models for determining the estimated start times and case orderings. IIE Trans 22(2):143–150
16. Strum DP, Vargas LG, May JH (1999) Surgical subspecialty block utilization and capacity planning. Anesthesiol 90:1176–1185
17. Vanden Bosch PM, Dietz DC (2000) Minimizing expected waiting time in a medical appointment system. IIE Trans 32:841–848
18. Dexter F, Ledolter J (2005) Bayesian prediction bounds and comparisons of operating room times even for procedures with few or no historic data. Anesth Analg 103(6):1259–1267
19. Dexter F, Marcon E (2006) Impact of surgical sequencing on post anesthesia care unit staffing. Health Care Manage Sci 9:87–98
20. Kleywegt AJ, Shapiro A, Homem-de Mello T (2001) The sample average approximation method for stochastic discrete optimization. SIAM J Optim 12(2):479–502
21. Van Slyke RM, Wets RJ-B (1969) L-shaped linear programs with applications to optimal control and stochastic programming. SIAM J Appl Math 17:638–663