Operations Research
Vol. 58, No. 4, Part 1 of 2, July–August 2010, pp. 802–816
issn 0030-364X | eissn 1526-5463 | 10 | 5804 | 0802

doi 10.1287/opre.1090.0791
© 2010 INFORMS

# Optimal Allocation of Surgery Blocks to Operating Rooms Under Uncertainty

**Brian T. Denton**
Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina 27695, bdenton@ncsu.edu

**Andrew J. Miller**
IMB, Université Bordeaux 1, and RealOpt, INRIA Bordeaux Sud-Ouest, 33405 Talence, France, andrew.miller@math.u-bordeaux1.fr

**Hari J. Balasubramanian**
Department of Mechanical and Industrial Engineering, University of Massachusetts, Amherst, Massachusetts 01003, hbalasubraman@ecs.umass.edu

**Todd R. Huschka**
Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota 55905, huschka.todd@mayo.edu

The allocation of surgeries to operating rooms (ORs) is a challenging combinatorial optimization problem. There is also significant uncertainty in the duration of surgical procedures, which further complicates assignment decisions. In this paper, we present stochastic optimization models for the assignment of surgeries to ORs on a given day of surgery. The objective includes a fixed cost of opening ORs and a variable cost of overtime relative to a fixed length-of-day. We describe two types of models. The first is a two-stage stochastic linear program with binary decisions in the first stage and simple recourse in the second stage. The second is its robust counterpart, in which the objective is to minimize the maximum cost associated with an uncertainty set for surgery durations. We describe the mathematical models, bounds on the optimal solution, and solution methodologies, including an easy-to-implement heuristic. Numerical experiments based on real data from a large health-care provider are used to contrast the results for the two models and illustrate the potential for impact in practice. Based on our numerical experimentation, we find that a fast and easy-to-implement heuristic works fairly well, on average, across many instances. We also find that the robust method performs approximately as well as the heuristic, is much faster than solving the stochastic recourse model, and has the benefit of limiting the worst-case outcome of the recourse problem.

*Subject classifications*: optimization; stochastic programming; surgery.
*Area of review*: Optimization.
*History*: Received July 2007; revisions received July 2008, May 2009, August 2009; accepted September 2009. Published online in *Articles in Advance* March 24, 2010.

## 1. Introduction

Operating rooms (ORs) have been estimated to account for more than 40% of a hospital's total revenues (HFMA 2005) and a similarly large proportion of its total expenses, which makes them a hospital's largest cost center as well as its greatest revenue source. Furthermore, recent studies indicate that OR efficiency measures, such as utilization, overtime, and on-time start performance, fall short of achievable targets at most hospitals (CAB 2001). Therefore, for health care providers, surgical suite management is an area with significant potential for realizing greater efficiencies. However, improving OR efficiency is a challenging problem for several reasons. First, finding a schedule that balances resource utilization (e.g., ORs, surgeons, nurses, etc.) is a combinatorial problem, which includes decisions such as how many ORs to open, allocation of surgeries to ORs, and surgery sequencing decisions. Second, there is significant uncertainty in several of the activities involved in the delivery of surgical care, including the duration of the surgical procedure itself. This uncertainty leads to unpredictable OR utilization and overtime staffing costs due to late closure of the surgical suite.

The challenge of balancing competing criteria to improve surgery scheduling is not new. (A review of pertinent literature can be found in §3.) However, the extant literature deals largely with a single OR and ignores the difficulties in scheduling surgeries in large multi-OR health care organizations. In this article, we develop and analyze deterministic and stochastic optimization models for a multi-OR allocation problem. The models can be viewed as extensions of the deterministic and extensible bin-packing problem (Dell'Olmo et al. 1998), which arises in a number of industrial contexts. In the OR allocation context, ORs represent bins of a certain size, where size denotes the time the OR is available during a particular day. The bins are extensible in the sense that they can be utilized for longer than the normal available time, but with a cost of overtime. Decisions include the number of ORs to open on a given day and the assignment of surgeries of varying length to open ORs.

We describe structural properties of the models we present, and we use these properties to develop computationally efficient solution methods. From the stochastic perspective, we propose both a stochastic recourse model formulation, and a robust formulation. The latter is applicable to situations in which limited information about probability distributions is available, which is a problem faced by many health care providers. We evaluate and compare each of the models using real data to illustrate the potential benefits in practice. Based on our numerical experimentation we find that a fast and easy-to-implement heuristic works fairly well on average across many instances, with the worst case being within 22% of the optimal solution to the stochastic recourse model. For certain selections of cost coefficients the heuristic is very close to optimal for a large number of test cases. We find that the robust method performs nearly as well as or better than the heuristic in many cases, is much easier to solve than the stochastic recourse model, and has the benefit of limiting the worst-case outcome of the recourse problem.

The remainder of this article is organized as follows. In the next section, we provide some background on the general problem of surgery scheduling. In §3, we provide a review of related literature. In §4, we present the deterministic version of the multi-OR scheduling problem that we propose, including structural properties and related bounds on the optimal solution. Sections 5 and 6 present two-stage stochastic recourse and robust formulations of the problem, respectively, and discuss structural properties of the models that provide computational advantages. In §7 we summarize solution methodology. In §8 we present some numerical results based on the solution of test cases using real data from a large health care provider. Finally, in §9 we summarize our findings and point out opportunities for future research.

## 2. Surgery Scheduling Process

There are different types of surgery delivery systems. Hospitals provide many services and are typically equipped with a broad range of capabilities, including an emergency department for handling cases resulting from unpredictable adverse events. More recently, a new delivery system called an outpatient procedure center (OPC) has emerged (Bowers and Mould 2005). OPCs service elective (equivalently *deferrable* and *scheduled*) surgeries that can be performed safely in an outpatient setting with minimal supporting resources. At hospitals there are different degrees of urgency associated with patient care. Often surgery can be performed on an elective basis on an agreed-upon future date. This is true of many types of surgery in which there is not an immediate need for intervention. Emergent add-on cases, on the other hand, are cases in which the timing is critical. They arise on short notice, and speed of intervention directly affects the patient's safety and potential for recovery. Hospitals typically reserve one or more ORs for emergent cases.

Whether surgery is performed on an inpatient or outpatient basis, at an OPC or hospital, or on an elective or emergent basis, many aspects of the OR environment are the same. From a facilities perspective ORs tend to be housed in a suite, in which several individual ORs share central resources such as an equipment storage area, sterilization resources, preoperative and recovery rooms. From a staffing perspective, the OR team is composed of a variety of uniquely skilled individuals, including the surgeon, one or more surgical assistants, an anesthesiologist, nurse anesthetist, and a scrub person. ORs have very high fixed costs, the large proportion of which is associated with the labor cost of the OR team, and staffing of upstream and downstream areas. Typically, ORs have a planned utilization time (e.g., 8 hours) beyond which overtime costs for some members of the OR team begin to accrue. Therefore, on-time surgery start performance, to the extent it affects overtime, is an important metric.

There are two common processes for advance planning of surgeries, known as *block-booking* and *open-booking*. Under a block-booking system individual surgeons or surgical groups are assigned times in a particular OR in a periodic (e.g., weekly or monthly) schedule. During each period, surgeons book cases into their assigned block time. In open-booking systems, on the other hand, surgeons submit cases up until the day of surgery, and by and large, all accepted cases are scheduled, subject to limits on total capacity. Once the set of surgeries has been collected, individual surgeries are then allocated to ORs to create a schedule immediately prior to the day of surgery. The open-booking policy is more common in destination medical centers. Such health care centers often see complex patients that have travelled long distances, and it might only become certain that they are candidates for surgery upon arrival and completion of a medical examination. Because it is a hardship to patients to return at a later date (and in some cases a safety concern to delay surgery), destination medical centers try to offer surgery on short notice. For instance, the health care provider that motivates the problems we consider has a goal of offering surgery on 24 hours notice for nonregional patients in need of surgery. These policies serve the patient's needs well. However, as a result of such policies there can be significant uncertainty in the type and number of surgeries to be scheduled on any given day. Such uncertainty arises, in part, because of the goal to provide surgery on short notice to visiting nonregional patients who are found to be surgical candidates.

In this article, we concentrate on surgery-to-OR assignment decisions, which play an important role in both block-booking and open-booking processes. We assume that the complete set of surgeries is known in advance (e.g., the evening prior to surgery) and the goal is to construct an assignment that trades off two competing criteria: (1) the fixed cost of opening individual ORs and (2) the total cost of overtime across all ORs. We study these costs because they represent the performance measures with the greatest direct costs. Additional factors that could be considered include idle time between individual surgeries and waiting times for patients, for OR teams, and for other critical resources. Many other studies have considered these additional performance measures in the context of single OR scheduling problems. In this article, we ignore the single OR performance measures in favor of the more strategic measures of total OR fixed costs and overtime costs. Implicit in this is the assumption that individual OR performance measures such as idling and waiting time are considered in assigning start times for individual cases after surgery-to-OR assignments have been made. This parallels the hierarchical planning structure common to other industrial planning and scheduling environments.

## 3. Literature Review

In our review, we focus on the literature relevant to stochastic scheduling and advance planning. More extensive reviews on challenges related to the management of ORs can be found in Goldman et al. (1970), Blake and Donald (2002), Magerlein and Martin (1978), and Przasnyski (1986). We also concentrate on advance scheduling, which involves allocating OR time in advance of the day of surgery. The single-OR scheduling problem is the simplest version of the advance scheduling problem. It concerns the setting of start times in the presence of uncertainty in surgery durations. The objective is to balance relevant metrics, including surgeon and OR team waiting, patient waiting, idling of the OR, and overtime costs for running later than the scheduled closing time.

The single-OR scheduling problem is mathematically similar to problems arising in many contexts in which appointment decisions are economically significant. Sabria and Daganzo (1989) consider scheduling the arrival of cargo ships at a seaport. In the treatment of the problem by Sabria and Daganzo, the costs of underutilization of a seaport are traded off against the cost of cargo ship waiting. On the other hand, Wang (1993) discusses the problem in a manufacturing setting where the objective is to schedule the arrival of parts on the shop floor such that work-in-process inventory and machine idling are minimized. There have been numerous other simulation- and queuing-based studies presented in operations research, statistics, and health care journals over the past several decades on the problem of assigning start time for surgeries and outpatient clinic appointments (for example Bailey 1952, Charnetski 1984, Dexter et al. 1999, Ho and Lau 1992, Jansson 1966, Mercer 1973, Rohleder and Klassen 2002, Soriano 1966, Welch 1964, and references therein). Ho and Lau (1992) used Monte-Carlo simulation to compare the performance of many of the proposed scheduling heuristics.

Another avenue of research for single-OR scheduling is the study of optimization models. Weiss (1990) solves two- and three-surgery scheduling problems, which can be solved relatively easily owing to the low dimensionality. Special cases where the number of surgeries is greater than 3 are considered by Wang (1993) (manufacturing jobs in the context of the paper), wherein job durations are exponentially distributed and computational advantages of phase-type distributions can be exploited. Vanden Bosch and Dietz (2000) present an algorithm for a similar problem for the case of phase-type distributions in which appointment slots are integer multiples of a discrete slot parameter. Denton and Gupta (2003) study a general two-stage stochastic linear programming formulation of the OR scheduling problem and provide efficient methods for solving larger instances of the problem. All these studies concentrate on a single OR and consider the impact of uncertainty on waiting and idling time with respect to planned start times for surgeries. In this article we consider the broader problem of managing multiple ORs, and we consider the most important performance measures for a complete surgical suite.

The literature on advance planning in the context of multiple ORs is sparse. Notable references include the following. Blake and Donald (2002) present a deterministic integer programming formulation of a model for setting block-booking schedules for multiple ORs. Dexter et al. (2004) consider policies under a block-booking schedule in which unutilized OR time is released prior to the day of surgery. They consider the trade-offs regarding the timing of release, where such problems are analogous to those found in the revenue management literature. McIntosh et al. (2006) discuss performance measures for evaluating OR productivity and costs based on hospital data. They provide a detailed review of the literature on discrete event simulation and other quantitative models for OR management appearing in the health care literature.

In contrast to the above-referenced literature, our goal is to study optimization models for planning and scheduling multiple ORs under uncertainty. The models we present are motivated by real problems at Mayo Clinic in Rochester, Minnesota; however, the models are generalizable to other providers of surgical services. We begin by presenting a deterministic optimization model for multi-OR assignment decisions. Next, we extend that to a two-stage recourse formulation that explicitly considers uncertainty in surgery durations. Finally, we present a robust formulation of the problem with two benefits. First, it is easier to solve than the two-stage recourse problem. Second, it requires limited information about surgery durations, which is a realistic limitation for some health care providers. Our model focuses on OR opening and closing decisions and surgery-to-OR assignment decisions (compared to the use of simulation models, which ignore the combinatorial aspect of allocation decisions). For simplicity, we ignore the upstream (intake) and downstream (recovery) resources required to support surgery, under the assumption that they are staffed based on the resulting surgery-to-OR assignment. This is reasonable because ORs tend to be the bottleneck in the overall process, and the cost of upstream and downstream resources can be factored into the fixed cost of opening ORs and overtime costs.

## 4. The Deterministic OR Allocation Problem

The problem we consider involves two important and related decisions: (1) how many ORs to open on a given day and (2) which OR to assign to each surgery in a daily listing. The OR opening and assignment decisions are made to minimize a weighted sum of the total cost of opening ORs and the total overtime due to overbooking of an OR. The decisions are coupled because opening a small number of ORs will tend to lead to greater overtime costs, due to difficulty in fitting surgeries into the available time in each OR, while opening a large number of ORs tends to result in lower total overtime, but at the expense of additional ORs.

We treat the OR opening decisions as having a fixed cost, where the cost is based primarily on the staff required to support the OR itself as well as up and downstream resources (e.g., nurses in a post-anesthesia care unit). These are assumed to be fixed costs because ORs typically are planned to be opened for a full day, or not at all. In the classical extensible bin-packing problem, the number of ORs is fixed. However, we treat the number of ORs to open as a decision variable in our model. Assuming the number of ORs that are opened changes from day to day is reasonable for the following reasons. First, many hospitals and OPCs that provide surgical services augment their nursing staff with contract nurses (often called *travelers*) who have flexible assignments and are paid on short-term contract, often hourly. Some hospitals, on the other hand, use only permanent employees and have less recourse to make short-term changes to staffing. In such cases it is common to have a *float pool* of nurses that can be drawn from other, less critical areas of the hospital. In such cases the fixed cost of an OR can be thought of as incorporating costs that could be saved if the workforce were smaller and/or the opportunity cost of being able to allocate nursing staff to other hospital activities.

The deterministic OR allocation (DORA) problem that we describe in this section seeks an optimal allocation of surgeries to ORs. We use the following notation:

$i$ index for blocks of surgeries $i = 1, \ldots, n$.
$j$ index for ORs $j = 1, \ldots, m$.
$d_i$ duration for surgery block $i$.
$T$ planned session length for each OR.
$c^f$ fixed cost to open an OR.
$c^v$ variable cost per unit time to keep an OR open past time $T$.
$x_j$ binary decision variable representing whether OR $j$ is opened.
$y_{ij}$ binary decision variable representing whether surgery block $i$ is allocated to OR $j$.
$o_j$ decision variable representing overtime for OR $j$.

We assume $i$ indexes *blocks* of surgeries, which we define as groups of one or more surgeries done consecutively in an OR by the same surgeon. We plan the surgeries together rather than independently because in practice, a given surgeon's surgeries are typically scheduled consecutively and done in the same OR. Note that our model is concerned only with the collection of surgeries and not with their specific sequence. This is reasonable because surgeons typically sequence surgeries based on clinical criteria or their own preference. The key decision variables are whether to open a given OR, $x_j$, and which surgery blocks are allocated to each OR, $y_{ij}$. Note that these variables are binary; the only continuous variable in the model is the overtime, $o_j$, for each OR. The model can be formulated as the following minimization problem.

$$\text{(DORA)} \quad Z_D = \min \quad \left\{ \sum_{j=1}^{m} (c^f x_j + c^v o_j) \right\}, \tag{1}$$

$$\text{s.t.} \quad y_{ij} \leqslant x_j \quad \forall (i, j) \tag{2}$$

$$\sum_{j=1}^{m} y_{ij} = 1 \quad \forall i \tag{3}$$

$$\sum_{i=1}^{n} d_i y_{ij} \leqslant T x_j + o_j \quad \forall j \tag{4}$$

$$\begin{aligned} & y_{ij}, x_j \in \{0, 1\} \quad \forall (i, j); \\ & \quad o_j \geqslant 0, \quad \forall j. \end{aligned} \tag{5}$$

The above problem is a variant of the *extensible bin-packing* problem (EBP), which is known to be NP-hard (Dell'Olmo et al. 1998). EBP assumes that the number of bins is known, and that bins may be loaded beyond their nominal capacity at a cost. The objective is to minimize the amount by which bin capacity is exceeded, i.e., there is a variable cost of over-filling bins. In our formulation the number of bins to open is also a decision variable. Given that ORs (bins) may be filled beyond their intended session length, $T$, we assume that

$$c^f < c^v T. \tag{6}$$

If (6) does not hold, then it is optimal to allocate all surgery blocks to a single OR because the per-unit cost of time is lower for overtime than for opening an OR. In other words, we assume that the per-unit cost of overtime is greater than the per-unit cost of *regular* time. Furthermore, we assume that $d_i < T$. In other words, it is assumed that surgery blocks can be completed within the planned OR session length. This assumption can easily be relaxed, but it is reasonable in a practical sense and allows for a simpler derivation of bounds presented in §4.2.

Note that we consider an uncapacitated version of the problem in which any number of ORs is available to open. Of course, in practice there is an upper limit on the number of available ORs. Our consideration of an uncapacitated model implies there is some control on the number of cases that can arise. In practice this control occurs through a constraint on limited surgeon availability, i.e., the number of cases cannot exceed the number of surgeons permitted to book surgeries for that day. Our model also assumes that any surgery can be completed in any OR; however, this is not necessarily true of all surgeries. An example of a surgery that could not be completed in all ORs is robotically assisted surgery. Robots are used in some cases for prostatectomy, hysterectomy, and heart valve repair. Such surgeries use expensive robots that are typically "hard-wired" into the room, i.e., they are not transferrable between ORs. However, these types of surgeries are the exception rather than the norm. Because the rooms are unique, planning typically is done separately for these rooms.

There are some structural characteristics of the above model that are worth mentioning. Because our intention is ultimately to study a stochastic version of the problem, in the remainder of this section we favor properties that can be adapted to the two-stage recourse version of the problem that we propose in §5. The first is that Constraints (4) are 0–1 continuous knapsack constraints with a single continuous slack variable. Mixed-integer-programming (MIP) solvers are able to generate strong inequalities from these, using results from the literature (see for example Marchand and Wolsey 1999, 2001; and Richard et al. 2003). Furthermore, Constraints (3) may be used to define special ordered sets, and Constraints (2) and (3) together define a clique structure in a graph. These latter characteristics do not seem to have as large an impact on computational speed as the first; the valid inequalities that solvers such as Cplex generate based on cliques do not strengthen the formulations of our models—either DORA or their extensions discussed below—nearly as much as do inequalities for knapsack sets. SOS branching is generally inferior to the solver's default branching strategy for our instances.

## 4.1. Symmetry-Breaking Constraints

Because all ORs have the same planned session length, there is complete symmetry with respect to ORs. Thus, for any solution, an equivalent solution can be obtained by switching the sets of surgeries assigned to any pair of ORs. Numerous authors have noted the importance of reducing symmetry in solving MIPs (e.g., Sherali and Smith 2001, Margot 2003, Ostrowski et al. 2009); for our model, symmetry results in the existence, for any given allocation of surgeries, of $m$! solutions to DORA based on permutations of surgery blocks.

To break this symmetry and limit the number of mathematical solutions to the actual number of differing allocations, we first add the constraints $x_1 \geqslant x_2 \geqslant \cdots \geqslant x_m$. (Constraints of this form are suggested in Sherali and Smith 2001.) However, we also need constraints on the $y$ variables that restrict the placement of surgery blocks into open ORs. To do this, we use the fact that we can realize any feasible allocation by assigning the surgery blocks to their ORs in lexicographical order.

We first make the realistic assumption that $n \geqslant m$. Without loss of generality, we can assume that surgery block $1'$ will be placed into OR 1. For block $2'$, we can assume that it is placed either into OR 1 with block $1'$, or into OR 2. We assume similar arguments for the subsequent blocks. We can therefore add the following constraints to restrict the set of ORs to which surgery blocks may be allocated:

$$\begin{aligned} y_{1'1} &= 1\\ y_{2'1}+y_{2'2} &= 1\\ \vdots\quad &\ \ \vdots\\ \sum_{j=1}^{m} y_{m'j} &= 1. \end{aligned} \tag{7}$$

Additional constraints may be added. For example, if the first five surgery blocks are put into the first two ORs, then we can assume that block $6'$ will either (1) be put into an OR in which one of the first blocks have been assigned (OR 1 or OR 2) or (2) it will be put into a different OR than the first five were put. In the latter case, we may further assume that block $6'$ is assigned to OR 3. That is, in case (2) we ensure that block $6'$ will *not* be put into ORs 4, 5, or 6. The following constraints enforce this:

$$\begin{aligned} &y_{6'4} \leqslant y_{3'3}+y_{4'3}+y_{5'3}\\ &y_{6'5} \leqslant y_{4'4}+y_{5'4}\\ &y_{6'6} \leqslant y_{5'5}. \end{aligned} \tag{8}$$

(Recall that $0 = y_{1'3} = y_{2'3} = y_{1'4} = y_{2'4} = y_{3'4} = y_{1'5} = y_{2'5} = y_{3'5} = y_{4'5}$ because of (7).) More generally, we have the following:

$$y_{ij} \leqslant \sum_{u=j-1}^{i-1} y_{u,\,j-1}, \quad \forall\,(i,j)\colon i \geqslant j.$$

Finally, for surgery block $6'$ note that we can combine the above ideas to tighten the first two constraints of (8):

$$\begin{aligned} y_{6'4}+y_{6'5}+y_{6'6} &\leqslant y_{3'3}+y_{4'3}+y_{5'3}\\ y_{6'5}+y_{6'6} &\leqslant y_{4'4}+y_{5'4}. \end{aligned}$$

The general form of these constraints becomes

$$\sum_{v=j}^{\min\{i,m\}} y_{iv} \leqslant \sum_{u=j-1}^{i-1} y_{u,\,j-1}, \quad \forall\,(i,j)\colon i \geqslant j. \tag{9}$$

Constraints (7) and (9), along with $x_1 \geqslant x_2 \geqslant \cdots \geqslant x_m$, are added *a priori* to the formulation of DORA and the other models in our computational experiments.

## 4.2. Upper and Lower Bounds on ORs

In this section we provide some lower and upper bounds on the optimal number of ORs.

Proposition 1. *The following is a lower bound on the optimal number of ORs to open*:

$$L=\left\lceil \frac{\sum_{i=1}^{n} d_i}{T(1+c^f/(c^vT))} \right\rceil.$$

Proof. Recall from (6) that $c^f < c^vT$, $j=1,\ldots,m$; this implies that each OR will be allocated a set of surgeries of duration at most

$$T+\frac{c^f}{c^v}=T\left(1+\frac{c^f}{c^vT}\right); \tag{10}$$

otherwise the solution can be improved by opening an additional OR to accommodate the additional capacity. Therefore, adding the constraint

$$\sum_{i=1}^{n} d_i y_{ij} \leqslant T\left(1+\frac{c^f}{c^vT}\right)x_j \tag{11}$$

to (1)–(5) does not change the optimal solution. Subsequently relaxing Constraint (4) and relaxing the binary allocation decisions, $y_{ij}$, to be continuous, results in a relaxation that is equivalent to the classical bin-packing problem (BP) with the following well-known lower bound:

$$\left\lceil \frac{\sum_{i=1}^{n} d_i}{T(1+c^f/(c^vT))} \right\rceil. \quad \square$$

An upper bound is achieved by any heuristic that generates a feasible solution to DORA. The restriction of DORA in which $o_j=0$, $\forall j$ corresponds to BP. Thus any feasible solution to BP also yields an upper bound on the optimal number of ORs. There are many heuristics for efficiently estimating solutions to BP. The following is a well-known upper bound on BP (and therefore DORA):

$$U=\left\lfloor 2\sum_{i=1}^{n}\frac{d_i}{T} \right\rfloor.$$

This bound follows from the fact that no two ORs can be less than half full, otherwise they could be combined into a single OR. See Coffman et al. (1984) for a survey of BP and a discussion of the above-mentioned bounds. Alternatively, we can develop the following stronger bound.

Proposition 2. *The following is an upper bound on the optimal number of ORs to open*:

$$U=\left\lfloor \frac{2\sum_{i=1}^{n} d_i}{T(1+c^f/(c^vT))} \right\rfloor.$$

Proof. If any two ORs have utilization less than or equal to

$$\frac{1}{2}\left(1+\frac{c^f}{c^vT}\right),$$

then they could be combined into a single OR at the same or a reduced total cost because $c^f \leqslant c^vT$ (by assumption). $\square$

### 4.3. A Simple Heuristic

The following heuristic uses the longest processing time (LPT) heuristic. It generates a feasible packing by iteratively solving problems with numbers of ORs ranging from the lower bound, $L$, to upper bound, $U$.

```
Heuristic:
    $n = L$; $z^{\min} = \infty$;
repeat
    $z$ = LPT($n$);
    if($n = U$ or $o_j = 0$, $\forall j$) stop;
    if($z < z^{\min}$) $z^{\min} = z$;
    $n \leftarrow n + 1$;
end(repeat).
```

LPT($n$) sorts all items in decreasing order of item size (surgery block duration) and allocates each successive item (starting with the largest) to one of the $n$ bins with the lowest current level. For each choice of ORs to open, the LPT heuristic solves the problem with a 13/12 approximation guarantee (Dell'Olmo et al. 1998). This yields the following proposition.

Proposition 3. *The extended LPT heuristic described above finds a solution to DORA that is within* 13/12 *of optimal.*

## 5. Stochastic OR Allocation Problem

In practice, surgery block durations exhibit considerable uncertainty (Denton et al. 2007). The stochastic OR allocation problem (SORA), which we formulate next, introduces uncertainty in the duration of surgeries as a consideration. As in the deterministic version of the problem, the primary decisions in this model are the decisions to open ORs, and allocation of surgeries to individual ORs to minimize total fixed and variable costs associated with the overall daily schedule. In the stochastic version of the model surgery block durations are random variables, denoted by $d_i(\omega)$, where $\omega$ defines the collective outcomes having support $\Xi \subseteq \Re^n$ and probability distribution $P$ on $\Xi$. Instead of the deterministic overtime cost, the stochastic problem considers expected overtime, which can be written as

$$E_\omega\left[\left(\sum_{i=1}^{n} y_{ij} d_i(\omega) - T\right)^+\right].$$

The fixed cost of opening an OR is incurred in advance of the surgery block durations being realized and therefore is deterministic. Thus, the stochastic model can be formulated as the following two-stage stochastic recourse problem:

$$\text{(SORA)} \quad Z_S^* = \min \left\{\sum_{j=1}^{m}\left(c^f x_j + E_\omega[c^v o_j(\omega)]\right)\right\}, \tag{12}$$

$$\text{s.t.} \quad y_{ij} \leqslant x_j \quad \forall (i,j) \tag{13}$$

$$\sum_{j=1}^{m} y_{ij} = 1 \quad \forall i \tag{14}$$

$$\sum_{i=1}^{n} d_i(\omega) y_{ij} - o_j(\omega) \leqslant T x_j \qquad \forall (j,\omega) \qquad (15)$$

$$y_{ij}, x_j \in \{0,1\} \quad \forall (i,j); \qquad o_j(\omega) \geqslant 0, \quad \forall (j,\omega). \qquad (16)$$

The decisions whether to open a given OR ($x_j$) and which surgery blocks to allocate to each OR ($y_{ij}$) are first-stage decisions. The second-stage recourse decisions, overtime $o_j(\omega)$ for each OR, are *simple recourse* decisions, and problems with this simple second-stage structure are known as *simple recourse* problems (see Birge and Louveaux 1997 for a detailed discussion of simple recourse problems). They benefit from second-stage sub problems that are separable and easily solved given the random outcome (surgery block durations) and the first-stage decisions $x_i$, $y_{ij}$.

### 5.1. Stochastic Upper and Lower Bounds

Several of the insights related to DORA discussed above are also applicable to this stochastic version of the problem. For example, the methods described above for symmetry breaking can also be applied to the stochastic version of the problem, for example, by adding the constraints to the master problem of an integer L-shaped decomposition framework (for a description of the integer L-shaped method see, e.g., Kall and Wallace 1994 or Birge and Louveaux 1997). The bounds of §4.2 can also be extended to the stochastic case of the problem. In the stochastic version there are a finite number of scenarios, indexed by $\omega$, defining the collective observation of random variables that represent the duration of surgery blocks. We let $\underline{d}_i$ and $\bar{d}_i$ denote lower and upper bounds on the support of random variables, $d_i$, respectively. Clearly, the minimum of the lower bounds, or the maximum of the upper bounds across all scenarios are valid bounds on the optimal number of ORs for SORA,

$$L_S = \left\lceil \frac{\sum_{i=1}^{n} \underline{d}_i}{T(1 + c^f/(c^v T))} \right\rceil, \qquad (17)$$

$$U_S = \left\lfloor \frac{2\sum_{i=1}^{n} \bar{d}_i}{T(1 + c^f/(c^v T))} \right\rfloor. \qquad (18)$$

Upper and lower bounds on the optimal solution to SORA can be achieved based on the properties of DORA identified in the last section. In the following, we let $\mu_i$ denote the mean duration of surgery block $i$.

Proposition 4. *The following are upper and lower bounds on the optimal value of SORA*:

$$c^f + c^v \sum_{i=1}^{n} \mu_i \geqslant Z_S^* \geqslant \frac{c^f \sum_{i=1}^{n} \mu_i}{T(1 + c^f/(c^v T))}.$$

Proof. The lower bound follows from the *wait-and-see* integer relaxation of two-stage stochastic recourse problems (see Birge and Louveaux 1997, Chapter 6)

$$Z_S^* \geqslant E_\omega[Z_D^*(\omega)] \geqslant \frac{c^f \sum_{i=1}^{n} \mu_i}{T(1 + c^f/(c^v T))},$$

where $Z_D^*(\omega)$ denotes the optimal value to DORA for scenario $\omega$, and the last inequality follows from Proposition 1. The upper bound follows from the restriction $\sum_{j=1}^{m} x_j = 1$, from which it follows for feasible $x_j$ and $y_{ij}$:

$$Z_S^* \leqslant c^f \sum_{j=1}^{m} x_j + c^v E_\omega\left[ \sum_{j=1}^{m} \left( \sum_{i=1}^{n} y_{ij} d_i(\omega) - T \right)^+ \right] \qquad (19)$$

$$\leqslant c^f + c^v E_\omega\left[ \left( \sum_{i=1}^{n} d_i(\omega) - T \right)^+ \right] \qquad (20)$$

$$\leqslant c^f + c^v E_\omega\left[ \sum_{i=1}^{n} d_i(\omega) \right] \qquad (21)$$

$$= c^f + c^v \sum_{i=1}^{n} \mu_i. \quad \square \qquad (22)$$

The upper bound on the optimal value in Proposition 4 can be used to derive a second upper bound on the optimal number of ORs for SORA as follows.

Proposition 5. *The following is an upper bound on the optimal number of ORs to open in SORA*:

$$U_S^2 = \left\lfloor 1 + \frac{c^v}{c^f} \sum_{i=1}^{n} \mu_i \right\rfloor.$$

Proof. From the upper bound in Proposition 4 it follows that

$$c^f + c^v \sum_{i=1}^{n} \mu_i \geqslant c^f m^* + c^v E\left[ \sum_{j=1}^{m} \left( \sum_{i=1}^{n} y_{ij}^* d_i(\omega) - T \right)^+ \right],$$

where $m^* = \sum_{j=1}^{n} x_i^*$. Recognizing that the second term on the right-hand side is nonnegative, it follows that

$$m^* \leqslant 1 + \frac{c^v}{c^f} \sum_{i=1}^{n} \mu_i,$$

and the proposed bound follows from the integer requirement on $m^*$. $\square$

## 6. Robust Formulation

A common problem faced by many health care providers is insufficient data for calibrating a stochastic programming model such as that described above. When limited or no data are available for procedure times, health care providers rely on estimates of surgery block durations. In this section we assume that a decision-maker is able to provide reasonable estimates of lower and upper bounds on durations for each of the surgeries to be scheduled. We assume that surgery block durations are random, with an unknown distribution, such that durations must lie within an *uncertainty set*, and that we want to choose an allocation of surgeries to ORs that minimizes the *worst* possible cost for all realizations of surgery block durations within the uncertainty set. (For research that develops and uses these concepts, see Atamtürk 2007, Bertsimas and Sim 2003, Bertsimas and Thiele 2006.) For each surgery block $i$ assigned to room $j$, let $\delta_{ij}$ be the *actual* duration of $i$ in $j$. We assume that

$$\underline{d}_i \leqslant \delta_{ij} \leqslant \bar{d}_i$$

must be true; that is, we assume that we know lower and upper bounds ($\underline{d}_i$ and $\bar{d}_i$) on how long $i$ will take. This allows us (1) to define a range of possible values for each surgery block; and (2) to define a bound on the total excess over the minimum that the surgeries can take. Thus, we consider a normalized bound on the total realization of demand:

$$\sum_{(i,j):\, y_{ij}=1} \frac{\delta_{ij}-\underline{d}_i}{\bar{d}_i-\underline{d}_i} \leqslant \tau. \tag{23}$$

In the context of our problem, $\tau$ denotes an upper bound on the number of surgeries that will achieve their worst-case upper bound on duration, $\bar{d}_i$, on a particular day. From a decision-maker's perspective $\tau$ is a way of controlling how conservative the worst-case scenario is. It is closely related to the selection of upper and lower bounds on the duration, $\bar{d}_i$ and $\underline{d}_i$, respectively. For instance, if $\bar{d}_i$ and $\underline{d}_i$ are set to 99% confidence intervals, then it is reasonable to set $\tau$ quite low because the probability of scenarios with a high number of surgeries achieving these bounds is very low. On the other hand, if the confidence intervals are narrower (e.g., 80% confidence intervals) then a larger value of $\tau$ is appropriate to reflect the higher probability of achieving such bounds. We revisit this in §6.1, where we discuss a heuristic for setting $\tau$, and in §8, where we provide a sensitivity analysis with respect to $\tau$.

Given the above discussion, intuitively it makes sense to think of $\tau$ as an integer. We can formulate the robust version of the problem as follows:

$$\begin{aligned}
\min_{(x,y)} \quad & \sum_{j=1}^{m} c^f x_j + F(x,y),\\
\text{s.t.} \quad & y_{ij} \leqslant x_j, \quad \forall (i,j)\\
& \sum_{j=1}^{m} y_{ij} = 1, \quad \forall i\\
& x_j \in \{0,1\} \quad \forall j; \qquad y_{ij} \in \{0,1\} \quad \forall (i,j),
\end{aligned}$$

where

$$\begin{aligned}
F(x,y) = \max_{\delta} \quad & \sum_{j=1}^{m} \eta_j,\\
\text{s.t.} \quad & \eta_j = c^v \max\left\{0, \left(\sum_{i:\, y_{ij}=1} \delta_{ij} y_{ij} - T x_j\right)\right\}, \quad \forall j\\
& \sum_{(i,j):\, y_{ij}=1} \frac{\delta_{ij}-\underline{d}_i}{\bar{d}_i-\underline{d}_i} \leqslant \tau\\
& \underline{d}_i \leqslant \delta_{ij} \leqslant \bar{d}_i, \quad \forall (i,j)\colon\, y_{ij}=1.
\end{aligned}$$

Note that in the problem defining $F(x,y)$, $x$ and $y$ are already fixed, and the $\delta$ variables are the decision variables. Note also that the bounds on the $\delta$ variables are only for $i$ and $j$ combinations for which $y_{ij}=1$, i.e., only for those surgery block allocations that have been chosen. The variable $\eta_j$ is the total cost associated with overtime in OR $j$ and is included in the formulation for convenience. Note that $\eta_j$ equals *either* 0 *or* the cost based on the difference between the actual durations in $j$ and the capacity, if the latter is nonnegative. We can represent this reality as a binary choice. Thus, we can reformulate $F(x,y)$ as follows:

$$\max_{(\delta,z)} \quad \sum_{j=1}^{m} c^v \left(\sum_{i=1}^{n} \delta_{ij} - T x_j\right) z_j, \tag{24}$$

$$\text{s.t.} \quad \sum_{j=1}^{m}\sum_{i=1}^{n} \frac{\delta_{ij}-\underline{d}_i y_{ij} z_j}{\bar{d}_i-\underline{d}_i} \leqslant \tau \tag{25}$$

$$\underline{d}_i y_{ij} z_j \leqslant \delta_{ij} \leqslant \bar{d}_i y_{ij} z_j, \quad \forall (i,j) \tag{26}$$

$$z_j \in \{0,1\}, \quad \forall j, \tag{27}$$

where $z_j = 0$ if $\sum_{i=1}^{n} \delta_{ij} - T x_j \leqslant 0$ and $z_j = 1$ if $\sum_{i=1}^{n} \delta_{ij} - T x_j > 0$.

Although this formulation is nonlinear (due to the multiplication of the $\delta$ and $z$ variables in (24)), it has a special structure. In particular, note that in any feasible solution, if $z_j = 0$ for some $j$, then $\delta_{ij} = 0, \forall i$, must be true. Thus, if an OR does not have overtime, the surgeries that have been assigned to that OR are redundant and are assigned a value of 0. Intuitively, this means that in seeking to look for the worst scenario, we do not need to consider assigning any durations that will never result in overtime. This allows us to reformulate $F(x,y)$ as a mixed-integer *linear* program:

$$\begin{aligned}
\max_{(\delta,z)} \quad & \sum_{j=1}^{m}\sum_{i=1}^{n} c^v \delta_{ij} - \sum_{j=1}^{m} (c^v T x_j) z_j,\\
\text{s.t.} \quad & (25)\text{–}(27).
\end{aligned}$$

Finally, we make the change of variable

$$\begin{aligned}
\Delta_{ij} &= \frac{\delta_{ij}-\underline{d}_i y_{ij} z_j}{\bar{d}_i-\underline{d}_i}\\
&= \frac{1}{\bar{d}_i-\underline{d}_i}\delta_{ij} - \frac{\underline{d}_i y_{ij}}{\bar{d}_i-\underline{d}_i} z_j.
\end{aligned}$$

Observe that for each $\Delta_{ij}$, $0 \leqslant \Delta_{ij} \leqslant 1$ must be true. Intuitively, $\Delta_{ij}$ will be 0 if $\delta_{ij}$ is at its lower bound, $\Delta_{ij}$ will be 1 if $\delta_{ij}$ is at its upper bound, $\Delta_{ij}$ will be $\frac{1}{2}$ if $\delta_{ij}$ is halfway between its two bounds, etc. Then the above formulation can be expressed as

$$
\begin{align*}
\max_{(\Delta,z)} \quad & \sum_{(i,j)} c^v(\bar{d}_i - \underline{d}_i)\Delta_{ij} \\
& \qquad - \sum_{j=1}^{m}\left(c^v\left(Tx_j - \sum_{i=1}^{n}\underline{d}_i y_{ij}\right)\right) z_j, \tag{28}\\
\text{s.t.} \quad & \sum_{(i,j)} \Delta_{ij} \leqslant \tau \tag{29}\\
& 0 \leqslant \Delta_{ij} \leqslant y_{ij} z_j, \quad \forall (i,j) \tag{30}\\
& z_j \in \{0,1\}, \quad \forall j. \tag{31}
\end{align*}
$$

The constraint matrix of the LP relaxation of (29)–(31) is not totally unimodular. However, we can still show that the LP relaxation is integer.

Proposition 6. *The polyhedron $X = \{(\Delta, z) : (29)\text{–}(30);\ 0 \leqslant z_j \leqslant 1, \forall j\}$ has integer extreme points.*

Proof. First, observe that $X' = \{(\Delta, z)\colon (30), 0 \leqslant z_j \leqslant 1, \forall j\}$ is an integral polyhedron, because the constraint matrix is totally unimodular ((30) has exactly one coefficient of 1 and one coefficient of $-1$ in each row, and the bounds on $z_j$ define an identity matrix). Next, observe that $X'' = \{(\Delta, z) \in X'\colon \sum_{(i,j)} \Delta_{ij} = \tau\}$ is a subset of $X'$ in which all extreme points are integer. Finally, observe that all extreme points of $X$ are either extreme points of $X''$, or extreme points of $X'$ in which (29) is satisfied by strict inequality. □

We apply LP duality to the LP relaxation of $F(x, y)$:

$$
\begin{align*}
\max_{(\Delta,z)} \quad & \sum_{(i,j)} (c^v(\bar{d}_i - \underline{d}_i))\Delta_{ij} \\
& \qquad - \sum_{j=1}^{m}\left(c^v\left(Tx_j - \sum_{i=1}^{n}\underline{d}_i y_{ij}\right)\right) z_j, \tag{32}\\
\text{s.t.} \quad & \sum_{(i,j)} \Delta_{ij} \leqslant \tau \tag{33}\\
& \Delta_{ij} - y_{ij} z_j \leqslant 0, \quad \forall (i,j) \tag{34}\\
& z_j \leqslant 1, \quad \forall j \tag{35}\\
& \Delta_{ij} \geqslant 0, \quad \forall (i,j); \qquad z_j \geqslant 0, \quad \forall j. \tag{36}
\end{align*}
$$

The dual is

$$
\begin{align*}
\min_{(\alpha,\beta,\gamma)} \quad & \tau\alpha + \sum_{j=1}^{m} \gamma_j, \\
\text{s.t.} \quad & \alpha + \beta_{ij} \geqslant c^v(\bar{d}_i - \underline{d}_i), \quad \forall (i,j) \\
& -\sum_{i=1}^{n} \beta_{ij} y_{ij} + \gamma_j \geqslant -c^v\Big(Tx_j - \sum_{i=1}^{n} \underline{d}_i y_{ij}\Big), \quad \forall j \\
& \alpha \geqslant 0; \qquad \beta_{ij} \geqslant 0, \quad \forall (i,j); \qquad \gamma_j \geqslant 0, \quad \forall j
\end{align*}
$$

where $\alpha$ is the dual variable associated with Constraint (33), $\beta$ are the dual variables associated with Constraints (34), and $\gamma$ are the dual variables associated with the bounds (35). We can then reformulate the original robust problem as follows:

$$
\begin{align*}
\min_{(x,y,\alpha,\beta,\gamma)} \quad & \sum_{j=1}^{m} c^f x_j + \tau\alpha + \sum_{j=1}^{m} \gamma_j, \\
\text{s.t.} \quad & y_{ij} \leqslant x_j, \quad \forall (i,j) \\
& \sum_{j=1}^{m} y_{ij} = 1, \quad \forall i \\
& \alpha + \beta_{ij} \geqslant c^v(\bar{d}_i - \underline{d}_i), \quad \forall (i,j) \\
& -\sum_{i=1}^{n} \beta_{ij} y_{ij} + \gamma_j \\
& \qquad \geqslant -c^v\left(Tx_j - \sum_{i=1}^{n} \underline{d}_i y_{ij}\right), \quad \forall j \tag{37}\\
& x_j \in \{0,1\} \quad \forall j; \qquad y_{ij} \in \{0,1\} \quad \forall (i,j) \\
& \alpha \geqslant 0; \qquad \beta_{ij} \geqslant 0, \quad \forall (i,j); \qquad \gamma_j \geqslant 0, \quad \forall j.
\end{align*}
$$

Gathering terms in (37), we obtain the following formulation, which we will call (RORA):

$$
\begin{align*}
\min_{(x,y,\alpha,\beta,\gamma)} \quad & \sum_{j=1}^{m} c^f x_j + \tau\alpha + \sum_{j=1}^{m} \gamma_j, \\
\text{s.t.} \quad & y_{ij} \leqslant x_j, \quad \forall (i,j) \\
& \sum_{j=1}^{m} y_{ij} = 1, \quad \forall i \\
& \alpha + \beta_{ij} \geqslant c^v(\bar{d}_i - \underline{d}_i), \quad \forall (i,j) \tag{38}\\
& \sum_{i=1}^{n} \beta_{ij} y_{ij} \leqslant c^v\left(Tx_j - \sum_{i=1}^{n} \underline{d}_i y_{ij}\right) + \gamma_j, \quad \forall j \tag{39}\\
& x_j \in \{0,1\}, \quad \forall j; \qquad y_{ij} \in \{0,1\}, \quad \forall (i,j); \\
& \alpha \geqslant 0; \qquad \beta_{ij} \geqslant 0, \quad \forall (i,j); \qquad \gamma_j \geqslant 0, \quad \forall j.
\end{align*}
$$

Due to the multiplication of variables in (39), this is a mixed-integer nonlinear program. However, note that for any $(i, j)$ pair, if $y_{ij} = 0$, then we can set $\beta_{ij}$ to the right-hand side of (38) without affecting the right-hand side of (39). Thus, if $y_{ij} = 0$, there is always a choice of $\beta_{ij}$ that maintains feasibility with respect to (38) and does not increase the objective function. This motivates modifying the formulation so that (38) is enforced if and only if $y_{ij} = 1$. We therefore consider the following modified formulation, which we call (MRORA):

$$
\begin{align*}
\min_{(x,y,\alpha,\beta,\gamma)} \quad & \sum_{j=1}^{m} c^f x_j + \tau\alpha + \sum_{j=1}^{m} \gamma_j, \\
\text{s.t.} \quad & y_{ij} \leqslant x_j, \quad \forall (i,j) \\
& \sum_{j=1}^{m} y_{ij} = 1, \quad \forall i \\
& \alpha + \kappa_{ij} \geqslant c^v(\bar{d}_i - \underline{d}_i) y_{ij}, \quad \forall (i,j) \tag{40}
\end{align*}
$$

$$\sum_{i=1}^{n}\kappa_{ij}\leqslant c^v\left(Tx_j-\sum_{i=1}^{n}\underline{d}_i y_{ij}\right)+\gamma_j, \quad \forall j \qquad (41)$$

$$x_j\in\{0,1\}, \quad \forall j; \qquad y_{ij}\in\{0,1\}, \quad \forall (i,j);$$

$$\alpha\geqslant 0; \qquad \kappa_{ij}\geqslant 0, \quad \forall (i,j); \qquad \gamma_j\geqslant 0, \quad \forall j.$$

Proposition 7. *For any instance of the robust problem, an optimal solution of RORA can be used to construct an optimal solution of MRORA with the same objective function value, and vice versa.*

Proof. We can prove the result by showing that the optimal solution of RORA corresponds to a feasible solution of MRORA with the same objective function value, and vice versa. To do this, consider an optimal solution to RORA. Now define a solution to MRORA by letting all variables $x$, $y$, $\alpha$, $\gamma$, and $\tau$ take the same values, and by letting $\kappa_{ij}=\beta_{ij}y_{ij}, \forall (i,j)$. This solution has the same objective function value and is clearly feasible to MORA. Now consider an optimal solution to MRORA. We can define a solution to RORA by letting all variables $x$, $y$, $\alpha$, $\gamma$, and $\tau$ take the same values, and by letting

$$\beta_{ij}=\begin{cases}\kappa_{ij}, & \text{if } y_{ij}=1,\\ c^v(\bar{d}_i-\underline{d}_i), & \text{if } y_{ij}=0.\end{cases}$$

Again, this solution is feasible for RORA, and it clearly has the same objective function value. □

By recalling the process through which $\alpha$, $\kappa_{ij}$, and $\gamma_j$ were defined, we can give intuitive interpretations. Specifically, we can view $\kappa_{ij}$ (divided by $c^v$) as the total time in excess of $\underline{d}_i$ spent working on surgery block $i$ in OR $j$, and we can view $\gamma_j$ as the total overtime cost associated with each OR. We can interpret $\alpha$ as the marginal value of being less conservative; i.e., if we can be certain that the number of surgeries lasting much longer than the minimum possible time can be assumed to be smaller, then we can expect the cost of the solution provided by this model to decrease by about this difference multiplied by $\tau$. We can therefore solve the robust version of the problem by solving the mixed-integer (linear) program MRORA. Moreover, this formulation is bigger only by a constant factor than the nominal problem, DORA.

The robust problem can be interpreted as an interesting alternative formulation to the stochastic OR allocation problem faced by an OR manager. It attempts to avoid the worst case, while imposing a limit on how conservative the optimal solution is. Alternatively, it can be used as a heuristic to obtain a solution to the stochastic recourse formulation presented in §5. In the following subsection we discuss a simple heuristic for setting $\tau$ to achieve a solution that performs well with respect to minimizing expected cost. In §8 we show that indeed MRORA can be an effective heuristic for computing near-optimal solutions to the stochastic recourse problem.

### 6.1. A Heuristic for Setting $\tau$

In this section we discuss a simple heuristic for estimating $\tau$. Our heuristic selects $\tau$ to try and generate a robust counterpart (MRORA) with an optimal solution that is near optimal for SORA (thus performing well under the worst-case cost and expected cost). To do this we use the following relaxation of SORA:

$$Z_R=\min\left\{\sum_{j=1}^{m}(c^f x_j+E_\omega[c^v o_j(\omega)])\right\}, \qquad (42)$$

$$\text{s.t.} \quad \sum_{i=1}^{n}d_i(\omega)-\sum_{j=1}^{m}o_j(\omega)\leqslant T\sum_{j=1}^{m}x_j, \quad \forall(\omega) \qquad (43)$$

$$x_j, \quad \forall j; \qquad o_j(\omega)\geqslant 0, \quad \forall (j,\omega). \qquad (44)$$

The above relaxation is obtained from the SORA formulation by relaxing integer constraints, aggregating constraints (15), substituting (14) to this new aggregate constraint, and relaxing Constraint (13). Defining the decision variable, $x_R=T\sum_{j=1}^{m}x_j$, the relaxation can be rewritten more concisely as the following single decision variable problem:

$$Z_R=\min\left\{\bar{c}^f x_R+c^v E_\omega\left[\left(\sum_{i=1}^{n}d_i(\omega)-x_R\right)^+\right]\,\middle|\, x_R\geqslant 0\right\}, \qquad (45)$$

where $\bar{c}^f=c^f/T$ is the cost-per-unit time for opening an OR, and $x_R$ is the total time to allot to complete all $n$ surgeries, such that exceeding $x_R$ results in overtime. This relaxation can be interpreted as a problem in which all surgeries are to be completed in a single OR with continuously adjustable duration, $x_R$. The above relaxation is a variant of the well-known *newsvendor problem*, which has the following optimal solution:

$$x_R^*=F^{-1}\left(1-\frac{\bar{c}^f}{c^v}\right),$$

where $F^{-1}(\cdot)$ is the inverse of the probability distribution function for $\sum_{i=1}^{n}d_i(\omega)$.

Next, we consider the optimal solution to the robust counterpart of (45). Let $D(\omega)=\sum_{i=1}^{n}d_i(\omega)$, and in the notation of §6, let $\bar{D}$ and $\underline{D}$ represent the upper and lower bounds of the uncertainty set for $D(\omega)$. Thus, the robust counterpart of (45) can be written as

$$\min\left\{\bar{c}^f x_R+c^v\max_{\delta}\{(\delta-x_R)^+\}\,\middle|\,\frac{\delta-\underline{D}}{\bar{D}-\underline{D}}\leqslant\bar{\tau},\right.$$
$$\left.\underline{D}\leqslant\delta\leqslant\bar{D}, x_R\geqslant 0\right\}. \qquad (46)$$

The optimal solution to (46) is

$$x_R^*=\underline{D}+(\bar{D}-\underline{D})\bar{\tau}. \qquad (47)$$

This follows from the fact that the optimal solution to the inner optimization problem is to set $\delta$ as large as possible subject to the constraint. Because $c^v \geqslant \bar{c}^f$, by assumption it follows that it is optimal in the outer optimization problem to set $x_R^*$ to the largest value of $\delta$. Our proposed heuristic attempts to select $\bar{\tau}$ such that the optimal solution is the same for (45) and (46). Thus, $\bar{\tau}$ can be written as

$$\bar{\tau} = \frac{F^{-1}(1 - \bar{c}^f/c^v) - \underline{D}}{\bar{D} - \underline{D}}. \tag{48}$$

We make note of the following points. First, as $n$ grows large, by the central limit theorem $F^{-1}(\cdot)$ can be approximated as a normal distribution with mean and variance equal to the sum of means and variances for the individual surgery blocks. This provides a computationally efficient means of computing $\bar{\tau}$ given summary statistics for individual surgery blocks. Second, in the case of identically distributed surgery block durations, setting $\underline{D}$ and $\bar{D}$ based on confidence intervals (as suggested in §6) results in the following intuitive relationship between uncertainty sets:

$$\bar{D} - \underline{D} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\bar{d}_i - \underline{d}_i). \tag{49}$$

Thus (49) motivates setting $\tau = \sqrt{n}\bar{\tau}$ in MRORA to reflect the fact that the uncertainty set of MRORA is increasing by a factor of $\sqrt{n}$ with respect to (46). Intuitively this reflects the tighter confidence intervals associated with estimating the sum of many random variables. In §8 we illustrate the performance of this heuristic based on empirical data.

## 7. Solution Methods

SORA is a two-stage stochastic recourse problem with binary decisions in the first stage and a continuous second stage sub problem. To solve it, we use an adapted version of the integer L-shaped method. In our implementation we iteratively solve a master problem of first-stage variables and constraints using branch-and-bound, and we successively add optimality cuts at each iteration. We use an adapted version of the multi-cut implementation of the L-shaped method in Birge and Louveaux (1988) in which we outer-linearize the recourse function for each OR independently, i.e., we can add up to $m$ optimality cuts at each iteration. Our master problem includes the antisymmetry constraints of §4.1, and we preprocess using the bounds, $L_S$ and $U_S$, of §5.1 to fix some of the $x_i$s to 1 (using the lower bound) and 0 (using the upper bound). We also use mixed-integer rounding cuts to solve the master problem.

We have tested two different methods for adding optimality cuts in our integer L-shaped method implementation. In the first, we solve the master problem to optimality at each iteration before adding optimality cuts. In the second, a branch-and-cut implementation, we add optimality cuts using Cplex *user callbacks* each time a feasible solution is found. Thus, the former approach is more selective, while the latter adds many cuts during the solution of the master problem. Through numerical experimentation, we find that the former approach is superior. In our final implementation we have used a trade-off between these approaches in which we initially set the tolerance for solving the master problem to 20%. At each iteration this is successively reduced by an order of magnitude to a final tolerance of 0.0000005. This allows the master problem to be solved quickly and more optimality cuts to be added in early iterations to articulate the recourse function.

## 8. Numerical Experiments

To test the proposed methods, we consider two types of instances: 10- and 15-surgery blocks. We assume $T = 8$ hours, and we consider two different types of fixed cost and variable cost settings: $c^f = 1$ and $c^v = 0.0333$ and $c^f = 1$ and $c^v = 0.0083$. The first choice of $c^v$ assumes 30 minutes of overtime is equivalent in cost to opening a new OR when $T = 480$ minutes. The latter assumes that 2 hours of overtime is equivalent to opening a new OR. These cost coefficients were chosen to approximate relative priorities of fixed vs. overtime costs in typical outpatient and inpatient settings, respectively. For each instance we generate 1,000 different scenarios by sampling with replacement from real surgery block duration data. Ten different instances are generated for each combination of cost values and the number of surgeries. Each instance uses a different seed to sample randomly for the empirical distributions and therefore results in a different set of scenarios.

We compare the solution to SORA with the following approximations:

- The optimal mean value (MV) solution—that is, the solution obtained by DORA using the mean surgery block durations obtained from the empirical data.
- The solution obtained by using the LPT heuristic to solve DORA.
- The solution obtained by solving MRORA, which assumes information about upper and lower limits on the surgery blocks. For the purposes of our experiments, we use the 10th and 90th percentile of the empirical distribution of surgery blocks. We consider several different values of $\tau$ (2, 4, and 6) as examples to evaluate the sensitivity of MRORA.

Results from the test instances are shown in Tables 1–4. All instances were solved on a 1.062GHz Sun Microsystems Sun-Fire-V440 with 8GB of memory, using Cplex version 10 to solve the master problem. For the instances tested in this article, SORA required a mean of 140 seconds of CPU time to produce a solution. However, the time required depended upon $c^v$. Low $c^v$ instances took less than 20 seconds of CPU time while high $c^v$ instances took as much as 12 minutes of CPU time. The DORA approach produced solutions within a minute of CPU time for all instances while MRORA took a mean of approximately 90 seconds of CPU time; as with SORA, the high $c^v$ instances typically took longer. To compare solutions for each method the ratios are calculated as follows. For a given instance and

**Table 1.** Summary of results for 15-surgery blocks with $c^f = 1$ and $c^v = 0.033$ represented as fraction of optimality with respect to SORA.

| | $c^v = 0.033$ | | | | |
|---|---|---|---|---|---|
| | | | Robust IP | | |
| Instance | MV IP | LPT heu | $\tau = 2$ | $\tau = 4$ | $\tau = 6$ |
| 1 | 0.808 | 0.806 | 0.892 | 0.906 | 0.933 |
| 2 | 0.953 | 0.966 | 0.898 | 0.896 | 0.970 |
| 3 | 0.854 | 0.852 | 0.936 | 0.937 | 0.970 |
| 4 | 0.925 | 0.972 | 0.911 | 0.971 | 0.917 |
| 5 | 0.896 | 0.946 | 0.831 | 0.916 | 0.892 |
| 6 | 0.862 | 0.853 | 0.923 | 0.931 | 0.938 |
| 7 | 0.930 | 0.936 | 0.810 | 0.930 | 0.817 |
| 8 | 0.888 | 0.966 | 0.876 | 0.903 | 0.904 |
| 9 | 0.962 | 0.966 | 0.964 | 0.969 | 0.964 |
| 10 | 0.860 | 0.924 | 0.910 | 0.893 | 0.918 |
| Mean | 0.894 | 0.919 | 0.895 | 0.925 | 0.922 |
| Stdev | 0.046 | 0.057 | 0.047 | 0.028 | 0.046 |
| Max | 0.962 | 0.972 | 0.964 | 0.971 | 0.970 |
| Min | 0.808 | 0.806 | 0.810 | 0.893 | 0.817 |

**Table 3.** Summary of results for 10-surgery blocks with $c^f = 1$ and $c^v = 0.033$ represented as fraction of optimality with respect to SORA.

| | $c^v = 0.033$ | | | | |
|---|---|---|---|---|---|
| | | | Robust IP | | |
| Instance | MV IP | LPT heu | $\tau = 2$ | $\tau = 4$ | $\tau = 6$ |
| 1 | 0.887 | 0.884 | 0.805 | 0.894 | 0.802 |
| 2 | 0.776 | 0.774 | 0.886 | 0.888 | 0.786 |
| 3 | 0.780 | 0.780 | 0.889 | 0.789 | 0.789 |
| 4 | 0.930 | 0.974 | 0.849 | 0.912 | 0.914 |
| 5 | 0.902 | 0.902 | 0.934 | 0.898 | 0.808 |
| 6 | 0.904 | 0.906 | 0.805 | 0.894 | 0.805 |
| 7 | 0.970 | 0.954 | 0.908 | 0.908 | 0.908 |
| 8 | 0.968 | 0.973 | 0.946 | 0.904 | 0.907 |
| 9 | 0.946 | 0.996 | 0.826 | 0.830 | 0.744 |
| 10 | 0.776 | 0.774 | 0.886 | 0.888 | 0.786 |
| Mean | 0.884 | 0.892 | 0.873 | 0.880 | 0.825 |
| Stdev | 0.078 | 0.087 | 0.051 | 0.040 | 0.061 |
| Max | 0.970 | 0.996 | 0.946 | 0.912 | 0.914 |
| Min | 0.776 | 0.774 | 0.805 | 0.789 | 0.744 |

method the ratio is the expected cost of the optimal solution of SORA to the expected cost of the solution generated by the approximation method (MV, LPT, MRORA) for SORA.

From Tables 1–4 is clear that LPT is, on average, marginally better than MV; for certain instances, however, when the variable cost is high, LPT is 5%–6% better. An intuitive explanation of this is as follows. Recall that the mean value problem minimizes overtime by setting each surgery block duration equal to the mean; this does not always ensure that the load is leveled across the ORs. Thus an OR that suffers high overtime is likely to do so across many scenarios in the stochastic version of the problem. On the other hand, the LPT heuristic, which also uses the mean surgery block durations, seeks to minimize makespan and therefore levels the load on the ORs. When the LPT solution is evaluated over the scenarios, there is greater likelihood that the ORs will finish before the limit for the day. The LPT heuristic is also easier to implement in practice. Overall, both LPT and MV give very good results when the variable cost is low; indeed, their average performance is within 2% of the optimum in this case.

**Table 2.** Summary of results for 15-surgery blocks with $c^f = 1$ and $c^v = 0.0083$ represented as fraction of optimality with respect to SORA.

| | $c^v = 0.0083$ | | | | |
|---|---|---|---|---|---|
| | | | Robust IP | | |
| Instance | MV IP | LPT heu | $\tau = 2$ | $\tau = 4$ | $\tau = 6$ |
| 1 | 0.999 | 0.998 | 0.880 | 0.948 | 0.948 |
| 2 | 0.999 | 0.999 | 0.999 | 0.999 | 0.980 |
| 3 | 0.999 | 0.999 | 0.929 | 0.952 | 0.944 |
| 4 | 0.999 | 0.998 | 0.930 | 0.930 | 0.929 |
| 5 | 0.990 | 0.996 | 0.932 | 0.938 | 0.924 |
| 6 | 0.989 | 0.990 | 0.886 | 0.881 | 0.881 |
| 7 | 0.973 | 0.993 | 0.844 | 0.974 | 0.927 |
| 8 | 0.966 | 0.966 | 0.966 | 0.987 | 0.939 |
| 9 | 0.975 | 0.993 | 0.847 | 0.960 | 0.957 |
| 10 | 0.997 | 0.996 | 0.900 | 0.901 | 0.903 |
| Mean | 0.988 | 0.993 | 0.916 | 0.951 | 0.933 |
| Stdev | 0.013 | 0.010 | 0.059 | 0.045 | 0.028 |
| Max | 0.999 | 0.999 | 0.999 | 0.999 | 0.980 |
| Min | 0.966 | 0.966 | 0.844 | 0.881 | 0.881 |

**Table 4.** Summary of results for 10-surgery blocks with $c^f = 1$ and $c^v = 0.0083$ represented as fraction of optimality with respect to SORA.

| | $c^v = 0.0083$ | | | | |
|---|---|---|---|---|---|
| | | | Robust IP | | |
| Instance | MV IP | LPT heu | $\tau = 2$ | $\tau = 4$ | $\tau = 6$ |
| 1 | 0.995 | 0.995 | 0.934 | 0.789 | 0.789 |
| 2 | 0.987 | 0.980 | 0.752 | 0.869 | 0.766 |
| 3 | 0.973 | 0.972 | 1.000 | 0.782 | 0.782 |
| 4 | 0.993 | 1.000 | 0.955 | 1.000 | 0.823 |
| 5 | 0.989 | 1.000 | 0.948 | 0.809 | 0.809 |
| 6 | 1.000 | 0.999 | 0.931 | 0.998 | 0.713 |
| 7 | 0.999 | 1.000 | 1.000 | 0.993 | 0.936 |
| 8 | 0.991 | 0.999 | 0.909 | 0.810 | 0.810 |
| 9 | 0.890 | 0.983 | 0.909 | 0.833 | 0.833 |
| 10 | 1.000 | 0.992 | 0.930 | 0.930 | 0.769 |
| Mean | 0.982 | 0.992 | 0.927 | 0.881 | 0.803 |
| Stdev | 0.033 | 0.010 | 0.069 | 0.091 | 0.058 |
| Max | 1.000 | 1.000 | 1.000 | 1.000 | 0.936 |
| Min | 0.890 | 0.972 | 0.752 | 0.782 | 0.713 |

## 8.1. Sensitivity Analysis of MRORA

As discussed above, $\tau$ controls how conservative the optimal solution to MRORA may be with respect to SORA. Based on Tables 1–4 MRORA produces solutions that are better for the 15-surgery block instances than the 10-surgery block instances. This is intuitive because the $\tau$ settings will lead to more conservative solutions for 10-surgery block instances, and hence might not perform as well for the expected cost criterion. While there is sensitivity to the variation in $\tau$, the results are comparable, except for the $\tau = 6$ setting for the 10-surgery block instances, which is 18%–20% higher than the optimal. This is caused by the ratio of $\tau$ to the number of surgery blocks being high, i.e., the worst case is overly conservative for these model instances.

To evaluate the heuristic suggested in §6.1 we varied $\tau$ values (from 0 to 15) for the 15-surgery block instances for the high and low settings of $c^v$. The $\tau = 0$ instance indicates the least conservative setting (each surgery block duration is constrained to be at its lower bound in the worst case), while the $\tau = 15$ setting indicates the most conservative setting (each surgery block duration may be at its upper bound in the worst case). Figure 1 shows the change in solution ratios for five instances with $c^v = 0.033$. Similarly, Figure 2 shows the change in solution ratios when $c^v = 0.0083$. In general, as $\tau$ increases, more ORs tend to be opened in solutions to MRORA because more surgeries reach their upper bounds in the worst case. For the high $c^v$ instance, $\tau = 0$ produces poor solution quality in all instances (this can be attributed to very few ORs being opened, hence resulting in high overtime costs), but as $\tau$ increases, the solution quality improves (particularly in the 2–4 range), and then gradually falls off again at higher values. Note also that the solution remains unaltered for a given instance at high levels of $\tau$ (e.g., for $\tau \geqslant 7$ in Figure 1). This can be explained as follows. Beyond a certain $\tau$ value, the number of ORs opened is high (it approaches the total number of surgeries to be scheduled) and remains constant, resulting in identical schedules. For the low $c^v$ instance (Figure 2) the $\tau = 0$ solution performs better: although very few ORs are opened, the low overtime costs mean that the total cost will remain relatively low. In this case we observe that the best $\tau$ values are in the range 2–5, and solution quality decreases and remains constant at high values of $\tau$. The heuristic of §6.1 provides reasonable results with $\tau$ values between 3–4 for the high $c^v$ instances and 2–3 for the low $c^v$ instances. We have plotted only five instances in each cost setting, but these results are representative of all other instances as well.

**Figure 1.** Illustration of sensitivity of the ratio of the MRORA solution to the optimal solution with respect to $\tau$ for five instances with $n = 15$ and $c^v = 0.033$.

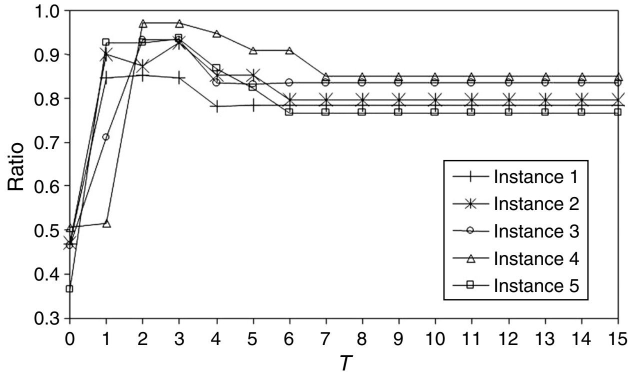
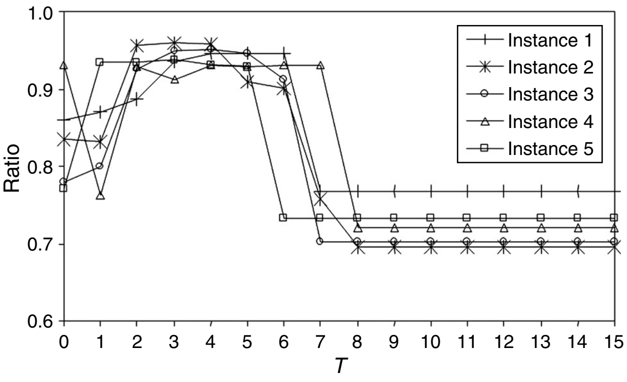


**Figure 2.** Illustration of sensitivity of the ratio of the MRORA solution to the optimal solution with respect to $\tau$ for five instances with $n = 15$ and $c^v = 0.0083$.



In summary, the results show that the benefits of solving SORA are often substantial when $c^v$ is high. In many instances, it is more than 10% better than the next best method. The LPT heuristic is a good alternative. Its performance is consistently good and the computation time needed to generate an LPT solution is a few seconds of CPU time. LPT is also easily implemented in practice. Finally, MRORA, while using only knowledge of upper and lower bounds, is a good alternative when little data are available. It also performs well as a heuristic for SORA and has the the additional benefit of protecting against the worst case.

## 9. Conclusions and Future Research Opportunities

The stochastic OR allocation problem captures the most important aspects of operational planning of daily surgical listings for a health care provider. In environments where there is flexibility in the number of ORs opened from day to day, our model can be used to compute the optimal number of ORs and allocation of surgeries to ORs. In environments where staffing is more rigid, or the goal is to maintain low variation in the ORs utilized each day, our model can be used to find the optimal allocation of surgeries to ORs for a fixed number of ORs. Such a model can play an important operational role in defining the surgery schedule for a particular day, as well as a longer term capacity planning role in evaluating which types and how many of each surgery to schedule for a particular day in the future. More efficient allocation of surgeries can reduce expected overtime and the number of ORs needed over the long term. For hospitals employing short-term contract nurses, the latter can have an immediate effect on costs. For hospitals using only permanent employees, staffing cost reductions might be realized over a longer timeframe through (a) natural attrition of nursing staff or (b) limiting staff expansion during periods of demand increase.

In this article we have examined the structure of the problem and proposed methods for solving it. In addition, we proposed a robust version (MRORA) that is appropriate when limited information is available and offers computational advantages relative to a two-stage stochastic mixed-integer programming formulation (SORA) of the problem. We used real data from a health care provider to evaluate and compare the stochastic programming and robust optimization models, as well as an easy-to-implement LPT heuristic. Based on our numerical experimentation we found that the heuristic works fairly well, on average, across many model instances, with the worst case being within 78% of the optimal solution to the stochastic recourse model. The heuristic works extremely well in situations where the cost of overtime per unit time is low. We found that the robust method performs approximately as well as the heuristic, is much faster to solve than the stochastic recourse model, and has the benefit of limiting the worst-case outcome of the recourse problem. The stochastic recourse model, SORA, was the most computationally intensive of the three approaches, and the value of the stochastic solution was highest (as high as 22% of the optimal solution) when the cost of overtime per unit time was high.

The problem we have studied is quite general. In practice, there are several additional factors that could be considered in the model, which will increase its complexity. Some examples include:

- *Surgery sequencing and start times*, which define the planned start time for each surgery and therefore whether the surgery starts late (causing waiting time for the patient and OR team that planned to use that OR) or if the previous surgery ends early (causing idle time for the OR).
- *Coloring constraints*, which specify that certain surgeries cannot be scheduled simultaneously (e.g., they are performed by the same surgeon but in different ORs, require the same mobile imaging devices, etc.). Such constraints link start times across multiple ORs, further increasing the size and difficulty of the recourse problem.
- *Upstream and downstream resources* that are necessary as part of the patient intake process to prepare for surgery (e.g., nurse evaluation, consent forms) and to recover (e.g., post-anesthesia care unit).

These and other practical and modeling considerations provide ample opportunity for future research on multi-OR surgery scheduling problems. The application of stochastic programming to the above situations is not necessarily straightforward, as the recourse problem is complicated and might not yield the kind of duality information that is often needed to obtain good L-shaped cuts. However, the application of robust optimization for such problems does seem promising, especially a framework such as that proposed in Bienstock and Özbay (2006) in which potential worst-case scenarios can be generated iteratively and incorporated into the decision problem, regardless of the structure of the problem used to generate these scenarios.

## Acknowledgments

This project was funded in part by CMMI-0620573 (Denton) from the National Science Foundation and by FA9550-07-1-0389 of the Air Force Office of Scientific Research (Miller). The authors are grateful for the helpful comments of two anonymous referees that helped improve this manuscript.

## References

Atamtürk A. 2007. Strong formulations of robust mixed 0-1 programming. *Math. Programming* **108**(2-3) 235–250.

Bailey, N. 1952. A study of queues and appointment systems in hospital outpatient departments, with special reference to waiting-times. *J. Roy. Statist. Soc.* **A14** 185–189.

Bertsimas, D., M. Sim. 2003. The price of robustness. *Oper. Res.* **52**(1) 35–53.

Bertsimas, D., A. Thiele. 2006. A robust optimization approach to inventory theory. *Oper. Res.* **54**(1) 150–168.

Bienstock, D., N. Özbay. 2006. Computing robust basestock levels. CORC Report TR-2005-09, Columbia University, New York.

Birge, J. R., F. Louveaux. 1997. *Introduction to Stochastic Programming*. Springer, New York.

Birge, J. R., F. V. Louveaux. 1988. A multicut algorithm for two-stage stochastic linear programs. *Eur. J. Oper. Res.* **34**(3) 384–392.

Blake, J. T., J. Donald. 2002. Mount Sinai Hospital uses integer programming to allocate operating room time. *Interfaces* **32**(2) 63–73.

Bowers, J., G. Mould. 2005. Ambulatory care and orthopaedic capacity planning. *Health Care Management Sci.* **8**(1) 41–47.

CAB. 2001. Surgical services reform: Executive briefing for clinical leaders. Technical report, Clinical Advisory Board, Washington, DC.

Charnetski, J. 1984. Scheduling operating room surgical procedure with early and late completion penalty costs. *J. Oper. Management* **5**(1) 91–102.

Coffman, E. G., M. R. Garey, D. S. Johnson. 1984. Approximation algorithms for bin packing—An update survey. G. Ausiello, M. Lucertini, P. Serafini, eds. *Algorithm Design and Computer System Design*. Springer, New York, 49–106.

Dell'Olmo, P., H. Kellerer, M. G. Speranza, Z. Tuza. 1998. A 13/12 approximation algorithm for bin packing with extendable bins. *Inform. Processing Lett.* **65**(5) 229–233.

Denton, B., J. Viapiano, A. Vogl. 2007. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Sci.* **10**(1) 13–24.

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

Dexter, F., R. H. Epstein, A. Macario. 2004. When to release allocated operating room time to increase operating room efficiency. *Anesthesia & Analgesia* **98**(3) 758–762.

Dexter, F., A. Macario, R. D. Traub, M. Hopwood, D. A. Lubarsky. 1999. An operating room scheduling strategy to maximize the use of operating room block time: Computer simulation of patient scheduling and survey of patient preferences for surgical waiting time. *Anesthesia & Analgesia* **89** 7–20.

Goldman, J., H. A. Knappenberger, W. J. Shearson. 1970. A study of the variability of surgical estimates. *Hospital Management* **110**(3) 46–46D.

HFMA. 2005. Achieving operating room efficiency through process integration. Technical report, Health Care Financial Management Association, Westchester, IL.

Ho, C.-J., H.-S. Lau. 1992. Minimizing total cost in scheduling outpatient appointments. *Management Sci.* **38**(12) 750–764.

Jansson, B. 1966. Choosing a good appointment system—A study of queues of the type ($D$,$M$,1). *Oper. Res.* **14**(2) 292–312.

Kall, P., S. W. Wallace. 1994. *Stochastic Programming*. John Wiley and Sons, New York.

Magerlein, J. M., J. B. Martin. 1978. Surgical demand scheduling: A review. *Health Services Res.* **13**(4) 418–433.

Marchand, H., L. A. Wolsey. 1999. The 0-1 knapsack problem with a single continuous variable. *Math. Programming* **85**(1) 15–33.

Marchand, H., L. A. Wolsey. 2001. Aggregation and mixed integer rounding to solve MIPs. *Oper. Res.* **49**(3) 363–371.

Margot, F. 2003. Exploiting orbits in symmetric ILP. *Math. Programming* **98**(1) 3–21.

McIntosh, C., F. Dexter, R. H. Epstein. 2006. Impact of service-specific staffing, case scheduling, turnovers, and first-case starts on anesthesia group and operating room productivity: A tutorial using data from an Australian hospital. *Anesthesia & Analgesia* **103**(6) 1499–1516.

Mercer, A. 1973. Queues with scheduled arrivals: A correction simplification and extension. *J. Royal Statist. Soc.* **Series 5**(35) 104–116.

Ostrowski, J., J. Linderoth, F. Rossi, S. Smriglio. 2009. Orbital branching. *Math. Programming Ser. A*, ePub ahead of print March 10.

Przasnyski, Z. 1986. Operating room scheduling: A literature review. *AORN J.* **44**(1) 67–79.

Richard, J. P. P., I. R. de Farias Jr., G. L. Nemhauser. 2003. Lifted inequalities for 0-1 mixed-integer programming: Basic theory and algorithms. *Math. Programming* **98**(1–3) 89–113.

Rohleder, T. R., K. J. Klassen. 2002. Rolling horizon appointment scheduling: A simulation study. *Health Care Management Sci.* **5**(3) 201–209.

Sabria, F., C. F. Daganzo. 1989. Approximate expressions for queuing systems with scheduling arrivals and established service order. *Transportation Sci.* **23**(3) 159–165.

Sherali, H. D., J. C. Smith. 2001. Improving discrete model representations via symmetry considerations. *Management Sci.* **47**(10) 1396–1407.

Soriano, A. 1966. Comparison of two scheduling systems. *Oper. Res.* **14**(3) 388–397.

Vanden Bosch, P. M., D. C. Dietz. 2000. Minimizing expected waiting time in a medical appointment system. *IIE Trans.* **32**(9) 841–848.

Wang, P. P. 1993. Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Res. Logist.* **40**(3) 345–360.

Weiss, E. N. 1990. Models for determining the estimated start times and case orderings. *IIE Trans.* **22**(2) 143–150.

Welch, J. 1964. Appointment systems in hospital outpatient departments. *Oper. Res. Quart.* **15**(3) 224–237.