Fall 2012

Integrated Block Sharing: A Win–Win Strategy for Hospitals and Surgeons

Robert Watson Day

Robert Garfinkel

Steven M. Thompson
University of Richmond

Recommended Citation

Day, Robert, Robert Garfinkel, and Steven Thompson. "Integrated Block Sharing: A Win–Win Strategy for Hospitals and Surgeons." *Manufacturing & Service Operations Management* 14, no. 4 (June 27, 2012): 567-83. doi:10.1287/msom.1110.0372.

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 14, No. 4, Fall 2012, pp. 567–583
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

Integrated Block Sharing: A Win–Win Strategy for Hospitals and Surgeons

Robert Day, Robert Garfinkel
Operations and Information Management, School of Business, University of Connecticut, Storrs, Connecticut 06269
{bob.day@business.uconn.edu, rob.garfinkel@business.uconn.edu}

Steven Thompson
Management Department, Robins School of Business, University of Richmond, Richmond, Virginia 23173,
sthomps3@richmond.edu

We consider the problem of balancing two competing objectives in the pursuit of efficient management of operating rooms in a hospital: providing surgeons with predictable, reliable access to the operating room and maintaining high utilization of capacity. The common solution to the first problem (in practice) is to grant exclusive "block time," in which a portion of the week in an operating room is designated to a particular surgeon, barring other surgeons from using this room/time. As a major improvement over this existing approach, we model the possibility of "shared" block time, which need only satisfy capacity constraints in expectation. We reduce the computational difficulty of the resulting NP-hard block-scheduling problem by implementing a column-generation approach and demonstrate the efficacy of this technique using simulation, calibrated to a real hospital's historical data and objectives. Our simulations illustrate substantial benefits to hospitals under a variety of circumstances and demonstrate the advantages of our new approach relative to a benchmark method taken from the recent literature.

Key words: healthcare management; math programming; production planning and scheduling; service operations
History: Received: November 15, 2010; accepted: November 2, 2011. Published online in *Articles in Advance* April 13, 2012.

1. Introduction

An important problem faced by many hospitals in the United States is how best to balance two competing objectives in the pursuit of efficient management of operating rooms—namely, providing surgeons with predictable, reliable access to the operating room (OR) and maintaining high utilization of capacity. According to a study conducted by the consulting firm Towers Perrin, the provision of surgical services accounts for 20%–40% of hospital expenses and contributes as much as 68% of total revenue through direct and ancillary services (Jackson 2002). Despite this importance of the OR to overall profitability, McKesson (2002) found that even after accounting for emergent and last minute "add-on" cases, the average OR runs at only 68% of staffed capacity. Indeed, our partner hospital for this study, Henrico Doctors' Hospital (HDH) in Richmond, VA, was running at about 55% of staffed capacity prior to our study. HDH is one of over 200 hospitals and outpatient centers owned by the Hospital Corporation of America (HCA), who have indicated that OR suite utilizations in this unsatisfactory range are common among their subsidiaries and are likely typical across peer institutions.

Achieving overall improvement in the efficiency, quality, and accessibility of hospital operating rooms is complex and challenging, because the determination of OR capacity, and the subsequent allocation of that capacity, is a multistep process comprising short-, intermediate-, and long-term decisions where decisions made at one stage can impact decisions at other stages. We propose, evaluate, and implement a new variation on the complex negotiation process between surgeons and the hospital implicit in OR planning, with a particular focus on the tactical problem of *block scheduling*. (As is typical in practice, block scheduling is used for surgeons scheduling nonurgent or nonemergent surgeries, though any improvement to a nonemergent system will free room capacity for emergency cases as an indirect benefit.) Central to this new paradigm is a relaxation of the concept that block time in operating rooms consists *only* of exclusive rights. Instead, we investigate a model where long-term tactical decision making involves a mixture of both exclusive and shared OR access for the surgeons.

In service systems with random demand, and healthcare systems specifically, risk pooling is a well-known way to provide higher customer service without increasing resources. By pooling random demand streams, variability is reduced and performance is enhanced. Taking risk pooling logic to its extreme, the most efficient system (in terms of utilization) for scheduling surgical procedures submitted by surgeons into operating rooms would be one that is completely pooled, open access, and first-come, first served. That, however, would be very undesirable to surgeons who would never be able to predict their schedules, making it difficult for them to handle clinics, office hours, and other aspects of their professional lives.

To accommodate surgeon preferences, most hospitals partition the available OR time for nonemergent cases into dedicated blocks. Each service or surgeon has property rights to a specific block of time. This effectively "depools" the schedule, creating many independent service queues instead of a single pooled one, even though the latter is known to be more efficient. Further, in a blocking system, many low-volume surgeons (we will speak of surgeons as opposed to services) typically get no block time at all. The fact that most hospitals employ such systems, forgoing pooling efficiencies, is testimony to how important it is to cater to surgeons (who have options to operate at other hospitals).

Our integrated block-scheduling system (IBS) combines the best aspects of these two extremes (open access scheduling and dedicated blocks) into a compromise that benefits both surgeons and the hospital. Specifically, the total time allocated to a surgeon is composed of exclusive time (maximizing reliability to the surgeon for those times) and shared time (capturing pooling benefits). The ratio of exclusive to shared time varies among surgeons so that none perceive themselves as worse off, and many are better off, with the new system. For example, high-volume surgeons used to easy access may get most of their time in exclusive blocks, but low-volume surgeons who previously had no reliable access can now get predictable shared blocks of time. In addition to pooling, IBS achieves improved resource utilization and reliable access to operating room time by scheduling these nonurgent procedures across a two-week scheduling horizon, thereby balancing demand.

Methodologically, IBS first generates a set of feasible block schedules for each surgeon via an integer programming formulation. Then another integer program finds an optimal overall block schedule that selects at most one schedule for each surgeon. The objective function maximizes the utility to hospital administration in dollars, and of the surgeons in dollars minus an "inconvenience" penalty. Hence, the solution will be a compromise between the pure preferences of both. The benefits accrued by the hospital and surgeons come from two direct sources and one indirect source. Direct benefits are obtained from reductions in so-called underutilized OR time, or time that is staffed but not needed, and overutilized OR time, which is time that is needed but was not planned so expensive overtime costs are incurred. A substantial indirect benefit, not included specifically in our optimization but of great practical significance, is the potential for adding new surgeons (and demand streams) as the hospital becomes more efficient. Hospitals are largely fixed cost systems, so adding revenue streams with the same resources is highly desirable. Indeed, this was a significant consequence of adopting our system in our client hospital.

Although IBS is sufficiently general to be applied at any hospital, it depends on three things that, if not present, would compromise its ability to perform as well as the case study illustrated later in this paper. First, the hospital should service primarily scheduled, nonurgent procedures rather than emergent cases. Because IBS is essentially an advance scheduling system, the overall benefit that the hospital and surgeons experience is dependent on the proportion of cases that can be submitted to the IBS system. Second, to avoid conflict for shared time, there must be ample OR capacity available (or available to be put online) to accommodate most possible caseload scenarios. Hospitals that are so severely constrained that they cannot do this will have a more difficult time taking full advantage of the pooled part of the schedule. Although the hospital could still service the cases, they would incur overtime costs and lose one of the most desirable benefits of IBS. This may prevent its application in highly resource-constrained hospitals (for example, academic medical centers and some large urban hospitals). However, such hospitals are not characteristic of the norm today. In fact, the average hospital in the United States achieves only 68% utilization of its ORs, so our method should enjoy wide applicability. Finally, an implementable compromise must grant high-volume, highly valuable surgeons a significant amount of dedicated time, leaving the lower-volume surgeons to accept significant amounts of shared time. Because shared time is the key to increased efficiencies, our method is most applicable to hospitals with a relatively large number of low-volume surgeons.

After describing IBS, we compare it to existing methods for determining operating room utilization. We first examine the previous procedure at HDH (prior to the adoption of IBS) consisting of a set of ad hoc rules for assigning blocks and scheduling cases into those blocks. Applying these rules in a simulated environment allows a comparison of IBS to one specific instance of common practice at many hospitals. In contrast to the typical ad hoc approach of HDH, a sophisticated minority of hospitals have utilized a more scientific model of the block-scheduling problem, as noted in a number of papers in anesthesiology and other healthcare-specific academic journals.

This literature stream suggests techniques for allocating block time to surgical subspecialties based on a newsvendor approach, which we will refer to as minimal cost analysis (MCA), borrowing the name used in Strum et al. (1999).

In general, MCA was designed for and implemented at academic medical centers associated with large university medical schools. It should be noted, however, that there are operational distinctions between these large, typically urban hospitals and smaller, suburban hospitals like HDH that make up the vast majority of hospitals in the United States. The former usually have very many operating room suites, and it has thus been reasonable to allocate block time to services (e.g., cardiology). In that setting, the large surgical groups will generally have enough surgeries to fill up very large blocks of time, including, for instance, multiple OR suites for full eight-hour days. On the other hand, smaller hospitals typically have fewer OR suites, fewer subspecialties, and fewer surgeons within each subspecialty. Furthermore, smaller hospitals tend to have a smaller proportion of emergent procedures. The lack of around-the-clock availability of specialists that are typically present at academic medical centers causes smaller hospitals to tend to focus more heavily on nonurgent procedures. As we will see, this setting yields both challenges that make it harder to implement an MCA approach and opportunities to improve both utilization and accessibility. Thus in a second round of simulations, we compare the performance of MCA (designed for academic medical centers) to IBS (designed around the suburban hospital model). Observing 900 simulated days of surgery, we consistently find that IBS offers substantial benefits over MCA for a hospital like HDH and even for hospitals with as much as 50% more surgery per room. A final round of simulations demonstrates the robustness of IBS to increases in surgical arrivals that were not anticipated during the planning phase.

The remainder of the paper is organized as follows. In §2 we review the operational environment and give a general overview of the decision problems faced by nearly all hospitals. In §3 we present a mathematical description of IBS. Simulation results are discussed in §4, and concluding remarks are provided in §5. Further technical details of our particular package-generation scheme, the heuristic online scheduling algorithm, details of our constrained bin-packing model, and a description of the implementation and impact of our research at HDH are provided in Online Appendices §§A.1–A.4 (available at http://dx.doi.org/10.1287/msom.1110.0372), respectively.

2. An Overview

We focus mainly on the tactical decision of assigning block time to surgeons or groups of surgeons (services), and we secondarily address three resulting operational decisions. We do not explicitly consider various possible strategic decisions that are assumed to be exogenous to the process, such as investment in new surgical equipment or infrastructure, recruitment of new specialities, etc.

2.1. The Tactical Decisions

The tactical decision of determining how to allocate blocks to each service or surgeon is generally made by OR managers every 6 to 12 months (Dexter et al. 2005). At minimum, the block schedule specifies the hours allocated to each service or surgeon for every day of a one- or two- week period. For example, a block assignment may designate that Dr. Smith has a block from 8 A.M. to 12 P.M. every Tuesday. The block schedule then serves as a template that recurs over time.

The block schedule serves two main purposes. First, from the perspective of the hospital, it guides the scheduling of surgical procedures and is helpful in identifying gaps in OR staffing needs. The latter is important because intraoperative processes and instrumentation vary considerably among subspecialties, and as a result, the OR staff that support a given surgeon are often trained for that subspecialty. Second, it helps surgeons plan their time efficiently. This is very important because most surgeons maintain private practices and utilize the hospital for the provision of some, but not all, procedures. As a result, from a business perspective, the average surgeon is primarily focused on maximizing the efficiency of a private practice. To balance time among working office hours, performing in the OR, performing rounds, consulting on patients in the hospital, and possibly visiting clinics, surgeons often request block time from hospitals before they will agree to affiliate with them. That is, each surgeon essentially wants a contract that guarantees routine access to an OR suite on certain days of the week during particular, convenient time frames. Typically in practice, only surgeons that perform a large number of cases are granted block time, whereas those that perform fewer cases must wait until after high-volume surgeons have scheduled their cases before they are granted access.

A very general realization of the blocking decision has been modeled as a newsvendor problem (Strum et al. 1999, Spangler et al. 2000). In that model the hospital faces a distribution of demand for OR time and must allocate capacity to satisfy all demand while minimizing the sum of two opposing staffing-related costs. The first, termed underutilization, is associated with staffed but unused capacity, and the second, termed overutilization, is associated with meeting demand that exceeds the planned OR time. Because overutilization costs typically involve overtime, on a per-hour basis, overutilization has a higher cost than underutilization. The actual decisions in the MCA model are how many hours to open individual ORs, each of which is dedicated to a service. In those works, the problem of determining how many hours of operating room time to allocate in order to minimize cost is framed at the level of the surgical service rather than the individual surgeon. In support of conducting the analysis at the service level, Dexter et al. (2003) simulate a scenario in which individual surgeons have stable blocks and assume that case arrivals and durations are independently distributed with Poisson and log normal distributions, respectively. They show that the time series of observed block utilizations (as a proportion of total block time) exhibit autocorrelation. This serves to make attempts to estimate expected block utilization for individual surgeons very difficult, thus motivating analysis at a more aggregate level where demand streams are pooled and the parameter values exhibit lower coefficients of variation. In our simulations we implement a particular realization of this concept and compare it to a corresponding realization of IBS. Our results show that surgeon-specific block-time allocations can work very well, but consistent with Dexter et al. (2003), the actual block-time allocations are not made based on attempts to estimate or predict expected block-time utilization. Rather, the block time is allocated such that very high, and statistically unlikely, levels of demand can be accommodated, and load balancing and block sharing are used to reduce the cost of providing very high service levels.

2.2. Operational Decisions

The first operational decision is to determine how many ORs to actually staff. These decisions are very similar to the tactical decisions but reflect near-term information (e.g., whether certain physicians are on vacation or when, over the next several weeks, a newly recruited surgeon will start to perform cases). Then, the second set of decisions involves the scheduling of surgical procedures, or cases, typically on an ongoing basis. When a request arrives, the hospital must decide whether to accept it and, if so, how much time to allocate to it on the schedule. The final set of operational decisions is made a few days before the day of surgery. Typically, in current practice, the hospital first decides whether to take back time that had been blocked for specific surgeons and allow others to schedule cases into that staffed time. This decision to "release" time is typically made one to five days before the day of surgery and will vary by specialty. The purpose of releasing what had been exclusive time is to allow other surgeons to fill unused time and reduce the underutilization of staffed time. Then, after all cases have been posted, they are assigned to specific rooms. The problem of assigning cases to rooms using the fewest number of suites has been modeled as a bin-packing problem and solved with heuristics (Houdenhaven et al. 2007). We find, however, that these bin-packing problems are typically small enough to solve quickly and optimally, indicating that this step is not a computational bottleneck.

2.3. The Three-Stage IBS Process

We continue in the spirit of papers such as Dexter and Macario (2002), tackling a hospital's tactical and operational concerns together, but here, we relax the traditional definition of block time as a purely exclusive contract with a surgeon. Instead, we allow for combinations of both "primary" or "exclusive" block time (similar to the existing practice) and "secondary" or "shared" block time that is nonexclusive in nature. The advantages of introducing secondary time are expanded upon in §§2.2 and 5.1, but the observed overall effect at HDH has been to increase the usage of staffed OR time in two ways. That is, it first allowed for a reduction in the total hours of staffed time for a fixed set of surgeries, thereby reducing costs, and second, the resulting excess capacity of staffed hours allowed for an expansion in the number of surgeries performed by recruiting new surgeons.

The IBS process is outlined as follows:

Stage 1. *The block-scheduling problem* (*BSP*)

- Actions: Assign surgeons regularly recurring blocks of time.
- Goals: Provide regular access to the OR in a manner that is consistent with surgeons' exogenous needs. Maximize expected value to the hospital by increasing utilization rates and decreasing staffing costs.
- Information: Historical distributions of surgeon caseloads and durations.

Stage 2. *The case-scheduling problem* (*CSP*)

- Actions: As each new case arrives (i.e., is prescribed by the surgeon), assign it a date and time of day as consistently as possible with the result of Stage 1.
- Goals: Schedule all cases in a timely manner, respective of surgeon desire for contiguity and a balanced workload.
- Information: Decisions must be made online, i.e., scheduling each case as it becomes known.

Stage 3. *Room consolidation problem* (*RCP*)

- Actions: Assign each case to a room, consistent with the Stage 2 assignment.
- Goals: Minimize staffing costs while avoiding excess downtime for both surgeons and rooms and respecting the Stage 1 block schedule as much as possible.
- Information: Because we consider only nonemergency cases, all scheduled cases for a particular day of operation will be known a few business days in advance.

We treat these information epochs and the decisions they necessitate as naturally occurring and immutable. Surgeons *demand* long-term scheduling regularity, so block-scheduling decisions *must* be made prior to the arrival of cases. As cases arrive, they *must* be assigned a date immediately to allow the patient and the surgeon to plan efficiently. We therefore address the problem in these three separate but interrelated stages.

The first stage is the main focus of this paper and is modeled as an integer program (IP), the variables of which are first generated via another IP formulation. We solve the first-stage problem to near optimality using CPLEX 11.1. Setting it apart from existing techniques, which solve a set packing of exclusive block times, our model packs exclusive block time deterministically while packing in only the *expected* usage of the shared block time and penalizing expected deviation above this expected usage in the objective function.

The scheduling stage should be flexible and is modeled in our simulation using a few heuristic rules of thumb, based on the schedule generated in Stage 1. In practice, these rules will guide a human scheduler, who can incorporate the varying human needs that would sometimes call for deviations from these guidelines. Then by Stage 3, a great deal of the relevant information is available, and finding the optimal schedule for a single day is a small enough optimization problem to solve rapidly to optimality.

The use of both primary and secondary block times has not been considered in this domain. Based on our simulations and the initial findings at HDH, this key feature of IBS results in what may be considered *Pareto improvements* relative to current practice. That is, surgeons who currently have considerable block time see little change from the status quo, and certain "premier" surgeons can even be guaranteed to receive at least as much exclusive block time, if so desired by the hospital. Simultaneously, the hospital experiences significant improvements in utilization rates and efficiency while lower-usage surgeons gain more regular access to the OR with earlier scheduling and less scheduling volatility.

3. The IBS Model

IBS is a complete three-stage scheduling system, based on the three information epochs described in §2.3, including a tool both for scheduling individual cases as they arrive and for fine-tuning this assignment once all scheduled cases for a particular day become known. As it was clearly necessary to test this design prior to implementation at HDH, we devised a complete simulation of all three stages based on the following framework.

3.1. Stage 1: The Block Scheduling Problem (BSP)

Here, we consider the overall problem of determining a medium- to long-term block schedule. The goal of the BSP is to select at most one personally feasible block schedule for each surgeon, and we expect that a hospital will run this optimization about once every six months, or perhaps once a year, to find a satisfactory block schedule that will remain stable for that length of time. Each personally feasible schedule will be referred to as a *package* of block time, containing the number of primary and secondary block hours assigned to the surgeon in each morning or afternoon of each day in a two-week scheduling window. For example, Dr. X might be awarded a package of two hours of primary time and two hours of secondary time on Tuesday morning. (The basic scheduling units of interest, in this case half-day windows, will be referred to as scheduling *bins*.)

Overall, this problem is an instance of multiobjective optimization, with several performance metrics differentiating some schedules as better or worse than others from the perspective of the different stakeholders. Typical methods for solving multiobjective problems impose a weighting of competing goals into a single aggregate objective function or constrain some objectives to be within reasonable bounds, or some combination of both. In this paper, we identify at least eight different relevant performance metrics or considerations and propose to control some as constrained within reasonable limits while weighting others by their dollar value into a single-objective function. In the latter category are the easily monetized measures: (i) profits to the surgeon, (ii) profits to the hospital for surgery performed, (iii) room staffing costs, and (iv) lost opportunity or inconvenience costs as a result of unavailable time in the schedule. In the former category are measures that are harder to monetize but do help distinguish some schedules as better than others from the surgeon's perspective: (v) the value of having block time on preferred days of the week or time of day, (vi) the value of more total time on the schedule that provides a surgeon more flexibility, (vii) the value of having long consecutive periods of block time, and (viii) the related value of having more time in fewer bins as opposed to spread across bins.

The BSP formulation we now present is focused on optimizing the monetized objectives (i)–(iv), given a set of feasible packages for each surgeon, each one of which has been generated to keep objectives (v)–(viii) within reasonable limits. In practice, surgeons could generate a set of feasible personal schedules that incorporate both their time preferences and constraints, so that they only request amounts of time that are sufficient for operations at the hospital and only on days of the week or times of the day that are relevant, given their office hours, obligations to other hospitals, etc. Alternatively, and more likely, the OR director would use software to generate these packages for each surgeon, provided her scheduling constraints are made known. Our simulations and initial implementation follow the latter approach, and we provide specific details on the generation of feasible packages in Online Appendix §A.1, but note here that the idea is to generate all (and only) combinations of block-time allocations for a surgeon that meet her time constraints and needs. Feasible days of the week (v) are inferred from past history and adhered to strictly, and (vi)–(viii) are given reasonable (perhaps generous) ranges. Pruning out packages that do not keep objectives (vi)–(viii) in reasonable ranges is useful in practice by limiting the size of the resulting difficult-to-solve IP.

We next introduce some of the necessary notation, beginning with the decision variables and parameters.

Decision Variables

x_k: A binary variable equal to 1 if and only if package k is chosen. Note that because each package is associated with a unique surgeon, it is not necessary to specify the surgeon, only the package.

r_j: The number of rooms open during the time period associated with bin j.

Basic Parameters of the Model

d: The number of surgeons. The set of surgeons will be denoted by $D = \{1, 2, \ldots, d\}$, and an individual surgeon will be indexed by $i \in D$.

b: The number of scheduling bins or scheduling blocks. At this point, we can imagine that each bin is a half-day on a particular day of the week (though we leave the model open to the possibility of other partitionings). Thus if we use a 10-workday-scheduling cycle with half-days as bins, we would have $b = 20$. The set B of all bins will be indexed consecutively by $j \in B = \{1, 2, \ldots, b\}$. For instance, for half-days, $j = 1$ refers to Monday morning, $j = 2$ to Monday afternoon, $j = 3$ to Tuesday morning, etc.

l: The time length of a bin. For half-days we set $l = 4$ hours.

R: The number of available rooms. For ease of notation, we assume, without loss of generality, that this is the same for all bins and that all rooms are identical.

c: The cost of opening and staffing a room for the duration of one bin, again assumed equal across bins.

K_i: The set of feasible schedule packages for Dr. i. The cardinality of each K_i is left unspecified for now, but for many practical purposes, we can consider it large enough to enumerate as many distinct combinations of primary and secondary hours as Dr. i could possibly find acceptable.

K: The set of all packages, $\bigcup_{i\in D} K_i$.

p_{ijk}: The amount of primary time awarded to Dr. i in bin j if $x_k = 1$, $k \in K_i$. That is, package k is chosen.

s_{ijk}: The amount of secondary time awarded to Dr. i in bin j if $x_k = 1$, $k \in K_i$.

S_{ik}: The total amount of secondary time awarded to Dr. i in all bins if $x_k = 1$, $k \in K_i$.

u^p_{ik}: The expected primary time usage by Dr. i if $x_k = 1$, $k \in K_i$.

u^s_{ik}: The expected secondary time usage by Dr. i if $x_k = 1$, $k \in K_i$.

σ^+_{ik}: The semistandard deviation of secondary block usage by Dr. i if $x_k = 1$, $k \in K_i$ (defined in detail below).

h_i: The average inconvenience cost to Dr. i for each hour of surgery above $p_{ijk} + u^s_{ik}$ when $x_k = 1$, $k \in K_i$.

v_i: The average value to the hospital for each of Dr. i's surgical hours.

π_i: The average profit to Dr. i of one hour of surgery.

The preceding parameters are essential parts of the model. When we also wish to use historical data to estimate u^p_{ik}, u^s_{ik}, and σ^+_{ik}, we make use of the following notation.

Historical Data Parameters

n: The number of past periods (i.e., scheduling cycles) for which we have historical data. The periods will be indexed by $t \in T = \{1, 2, \ldots, n\}$.

a_{it}: The total hours of surgery performed by Dr. i in past period t.

ξ^s_{ikt}: Inferred usage of secondary time by Dr. i in past period t if package k had been used.

Though we can have experts approximate u^p_{ik} and u^s_{ik} for surgeons without a previous history, or make adjustments to these values based on trends (expected new business, etc.), for surgeons with a stable history, our typical baseline estimates for Dr. i's usage can be computed as follows for any $k \in K_i$:

$$u^p_{ik} = \frac{1}{n}\sum_{t\in T}\min\biggl(a_{it}, \sum_{j\in B} p_{ijk}\biggr), \tag{1}$$

$$\xi^s_{ikt} = \max\biggl[0, \min\biggl(a_{it} - \sum_{j\in B} p_{ijk}, \sum_{j\in B} s_{ijk}\biggr)\biggr], \tag{2}$$

$$u^s_{ik} = \frac{1}{n}\sum_{t\in T}\xi^s_{ikt}, \tag{3}$$

$$\sigma^+_{ik} = \sqrt{\frac{\sum_{t\in T}\max(0, \xi^s_{ikt} - u^s_{ik})^2}{n}}. \tag{4}$$

Thus, our usage estimates represent the average under historical workloads if package k is selected. These equations reflect the notion that (1) primary time is filled before secondary time in general, so primary-time usage is either complete usage or total surgical hours when this is lower; and (2) secondary-time usage typically only occurs when primary time is full (the total amount of surgical time exceeds primary time) and is bounded by the total amount of secondary time allotted.

In (4), σ_{ik}^{+} is computed much like a typical standard deviation, as the square root of the average squared deviation from a mean value, but here, we only recognize positive deviations from the mean, so that this statistic only reflects variability *above* the expected secondary-time usage u_{ik}^{s}. Notice that if every surgeon had a static workload (i.e., each $a_{it} = a_i$, constant for each surgeon), then $\sigma_{ik}^{+} = 0$ for all $k \in K_i$, and the packing in the IP formulation of the BSP below would never involve any conflict. As variability in workloads increases, there is a greater chance that there will be conflict in the secondary-time allocations, but only when variability is in the direction of a workload being more than expected usage. Thus a penalty, increasing in σ_{ik}^{+}, is introduced into the objective function.

Our objective then is to generate a block-time schedule (a collection including at most one accepted package for each surgeon) that maximizes the total expected monetary value to the hospital and surgeons combined. This problem is described by the following IP formulation of the block-scheduling problem, or BSP-IP:

$$\max \sum_{i\in D}\sum_{k\in K_i}\left[(v_i+\pi_i)(u_{ik}^{p}+u_{ik}^{s}) - h_i\sigma_{ik}^{+}\right]x_k - c\sum_{j\in B} r_j \quad (5)$$

$$\text{s.t.} \sum_{i\in D}\sum_{k\in K_i}\left(p_{ijk} + u_{ik}^{s}\frac{s_{ijk}}{S_{ik}}\right)x_k \le l r_j \quad j \in B, \quad (6)$$

$$\sum_{k\in K_i} x_k \le 1 \quad i \in D, \quad (7)$$

$$0 \le r_j \le R \quad j \in B, \quad (8)$$

$$r_j \in \mathbb{Z} \quad j \in B, \quad (9)$$

$$x_k \in \{0,1\} \quad k \in K. \quad (10)$$

If $x_k = 1$, the hospital gains v_i times the expected usage $(u_{ik}^{p} + u_{ik}^{s})$ in hours, and the surgeon gains π_i times the expected usage but loses the inconvenience charge $h_i\sigma_{ik}^{+}$. These package-specific profits are represented by the first terms in (5), which are only achieved if the package is accepted (i.e., if the corresponding $x_k = 1$). Rooms must be opened to accommodate the associated surgical times, and the hospital pays c for each room it opens for an l-hour period of time, reflected in the second set of terms in (5). The penalty for secondary deviations above expected usage is in the objective function with penalty cost h_i, and the penalty for secondary deviations below expected usage is captured in constraint (6) and the objective function, because below-expected usage will mean underutilizing staffed rooms at a cost of c. Hence, the IP will reflect the asymmetry between underutilization and overutilization costs in a natural way.

Constraint set (7) ensures that at most one package is given to each surgeon, and constraints (8)–(10) simply define the nature of the decision variables and give bounds on the number of rooms open at a time. The bin-packing constraints (6) ensure that whenever r_j is increased by 1, another room is open during time bin j, and l new hours are available for fitting in expected surgical time. To understand the term $u_{ik}^{s}(s_{ijk}/S_{ik})$, suppose Dr. i has 10 hours of expected secondary usage and is allotted 15 hours total secondary time, 6 of which are in bin j. Then his expected secondary usage in bin j is four hours.

In Stage 1, we are optimizing the performance of the long-term scheduling plan, in which primary time is awarded as time *guaranteed* to be available to the surgeon, for the sake of consistency in the long-term planning interests of that surgeon. Therefore the entire amount of primary time allocated, p_{ijk}, must fit in the appropriate bin in (6), not just the expected primary-time usage u_{ik}^{p}. For secondary time, on the other hand, we allow for the possibility of more secondary time being awarded than is actually available and only require that the *expected* secondary-time usage fit in the time available, assuming that the fraction of secondary time used in a bin is equal to the fraction of secondary time, s_{ijk}/S_{ik}, awarded in that bin. This model is consistent with using σ_{ik}^{+} to penalize variability over the expected amount of secondary-time usage, because actualizations in which secondary-time usage is low do not experience any difficulty packing the surgeries into the scheduled open room time.

Note that this problem is NP-hard, as the 0–1 knapsack problem can be solved as a special case with $R = b = 1$ and $|K_i| = 1$ for each i (that is, one room, one bin, and a single acceptable package for each surgeon). But with a relatively small number of constraints, this problem seems to lend itself well to a column-generation approach, in which the packages under consideration are carefully managed in order to mitigate overall computational effort.

This brings us to the topic of generating a set of feasible packages for each surgeon. In practice, it would likely be difficult to obtain a large number of feasible packages from each surgeon if generated by hand. It is then useful to have a *package-generation tool* that the OR director can use while only requiring the surgeons to provide information regarding their availability on various days of the week. These simulated requests could be used in lieu of, or in combination with, requests made from participating surgeons. The package-generation technique we implemented could also be used as the basis for a decision-support tool for surgeons. With such a tool, surgeons would provide information regarding their availability on various days of the week, bounds on the total amount of block time they desire, etc. The tool would then recommend a large list of feasible packages, each with varying amounts of primary block time and secondary block time on the available days.

We present one such method for generating feasible packages in Online Appendix §A.1, though we emphasize that other schemes could be used to accommodate the needs of surgeons. The viability of the overall approach we present is based on the premise that there will be *some* method to generate a meaningful set of packages reflecting each surgeon's needs, and we give one such method that performed well for HDH. Our particular method emphasizes day-of-the-week availability, the need of some surgeons to have longer days for longer procedures, and the desire for contiguous primary time when possible (all to satisfy surgeons) while simultaneously providing packages with a wide variety of total block time and mixes of primary and secondary times, providing the flexibility to find better solutions to the BSP.

3.2. Stage 2: The Case-Scheduling Problem (CSP)

Block schedules, such as those generated in Stage 1 are typically used over a period of six months to a year, and are then reevaluated based on actual utilization as part of the standard block-time governance process in place at most hospitals. The problem of scheduling cases into the surgeons' block-time schedules is complicated by the fact that they must be scheduled in real time rather than batched and then scheduled. In addition, because of the amount of time and planning that patients typically need in order to prepare for surgery and postoperative recovery, once a case has been scheduled it cannot be rescheduled except at the request of the patient or surgeon (so the hospital cannot rearrange the schedule for its own benefit). Although the real-life implementation of this phase is performed by a human scheduler who can accommodate the idiosyncrasies of personal requests, etc., we recommend that these schedulers follow a rule of thumb, to first attempt to schedule cases contiguously into days with primary time, spilling over into secondary time on the same day when necessary, before using secondary time on other days. It was necessary, however, to simulate this behavior for our experiments, and we provide the mechanical procedures used in these simulations in §4.2 and Online Appendix §A.2.

3.3. Stage 3: The Room Consolidation Problem (RCP)

In Stage 3 the surgical procedures booked for a given day during Stage 2 are scheduled into rooms, and specific start times are established. In practice, the RCP would be solved one or two days before the actual day of surgery, so that equipment and resources could be planned accordingly, and patients and physicians can be notified when to report to the hospital. The objective of the RCP is to schedule all cases into as few rooms as possible, while adhering to the block-time windows from Stage 1. The complete formulation of RCP is given in Online Appendix §A.3, which tries to pack surgeries of various lengths into bins representing a room open for half of a working day. (In our formulation we assume for simplicity that there are two bins per day—an "A.M." and a "P.M." bin—each of length l, though an extension to the more general case is easily accommodated.) The constraints ensure that no room is assigned more than one surgery at a time and that no surgeon is assigned to be simultaneously present in more than one place. Though a reduction of bin packing to RCP makes it NP-hard, the problem in practice is not difficult. Currently, near-optimal solutions are found easily by hand, allowing a human scheduler to accommodate the idiosyncratic needs of surgeons and staff. An automated version of this problem was needed for our simulations prior to implementation of a block schedule, thus the problem is modeled in Online Appendix §A.3 and solved optimally.

4. Simulations

We set up a number of simulations to demonstrate the effectiveness of IBS in various environments and to compare it to existing techniques. In a first set of experiments, we simulated arrivals of cases based on historic utilizations at HDH, and compared IBS to the historical approach of HDH. We considered the application of various bounds on the proportion of time allocated as secondary time to improve the quality of schedules for more active surgeons under IBS. We also re-ran IBS in the absence of secondary block time to evaluate the improvement due to the use of shared block time.

Starting with records of case durations for $d = 124$ surgeons over an $n = 46$ week observation period at HDH, we first simulated a set of requests for block time using the package-generation technique of Online Appendix §A.1. We then combined these requests with historical patterns of usage to generate the expected usage parameters necessary for the implementation of Stage 1. These packages and usage amounts were then fed to the BSP-IP as described in §3.1, resulting in an OR block schedule, consisting of the amounts of primary and secondary times for each surgeon over a repeating two-week interval.

For all of our simulations, we used Arena version 12.0 from Rockwell Automation to replicate a case arrival process for all surgeons, based on the historical distributions of case timing and duration. The simulation platform allowed us to implement the decision rules for assigning these cases to time windows, the basis of CSP (Stage 2), as detailed in §4.2 and Online Appendix §A.2. Each schedule was then adjusted using the Stage 3 techniques described in §4.3 and Online Appendix §A.3. The results of this first round of simulations are summarized in §4.4.

Using the final parameter settings for IBS from the first round of simulations, we ran a second round implementing the IBS and MCA approaches on the same data in order to compare our technique to a benchmark from the literature. To indicate the robustness of the resulting comparison to hospital congestion, additional simulations were conducted, where surgical volume was increased (both in terms of the number of surgeons and the number of procedures) while holding constant the maximum capacity in terms of available OR rooms. A description of the MCA implementation and the results comparing the performance of IBS and MCA are given in §4.5.

Additionally, we reran the experiments of the baseline (HDH-sized) hospital with increases in the number of arrivals *after* the generation of a block schedule in order to explore the effect of *unexpected* increases in volume on the IBS system. This study of the robustness of IBS is described in §4.6.

4.1. Stage 1: Simulation Details

The first step in running a simulation with historical data for these surgeons was to generate a set of feasible packages for each surgeon. As noted in §3.1, *any* method for generating feasible packages of block time for each surgeon can be accommodated within our framework. Working in consultation with HDH to devise a practical and implementable framework for package generation, it became clear that exclusive block time had been used as an incentive or reward system in the past for the most active surgeons, and that some level of guaranteed primary time should be maintained for more active surgeons to perpetuate the idea of exclusive block time as a reward. Indeed, in some of our initial experiments in which secondary time could be awarded indiscriminately, we found instances in which very fragmented schedules occurred for some higher usage surgeons, thus negating some of the benefit of block time as a reward.

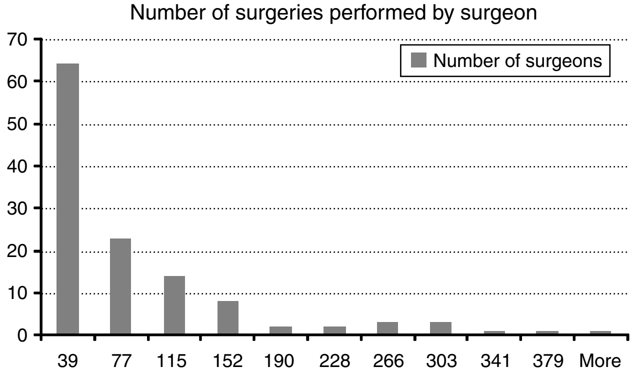
A complete description of the package generation tool we used is given in Online Appendix §A.1, but we may summarize the general idea as follows. We generate all feasible amounts of primary and secondary time for a particular surgeon within an acceptable range of total block time TB_i for surgeon i, not giving allocations that are too small to fit that surgeon's cases and not awarding too much secondary time. The total amount of block time is also varied, allowing for the generation of a large number of possible block schedules for each surgeon. Although the enumeration of feasible solutions is a fairly mechanical process and thus relegated to the online appendix, setting bounds on the proportion of time allocated as secondary time was explored heuristically in consultation with HDH, and thus we describe it in more depth here. Though this step can be adjusted to reflect a particular hospital's objectives, in the simulations that follow, we show that setting these parameters once at reasonable levels resulted in highly satisfactory performance with respect to objective measurements in all our simulations.

To capture the notion of preferred days of the week, we set up our simulation so that a surgeon would not get any block time during a bin unless at least 2% of her total prior history took place in that bin. (This assumed that a surgeon may have performed one or two stray surgeries on unpreferred days/times in the past but would not want to perpetuate these stray surgeries in the block schedule.) Other parameter values were set through discussions with the OR director at HDH to reasonable average levels, without discriminating among surgeons (even though our formulation in §3.1 was general enough to accommodate such discrimination). Clearly, the costs described in this section could be set to any values desired by the hospital administrators. We set $v_i = \$1{,}500$, $\pi_i = \$500$, and $h_i = \$100$, for each surgeon i, based on historical averages. The cost of opening a room for a half day was estimated at $c = \$3{,}000$, implying (together with the v_i value) that the hospital needs to fill two hours of surgical time to break even on opening a room for the half day. The number of bins and the length of a bin were set at typical two-week, half-day levels, $b = 20$, $l = 4$, and the maximum number of available rooms was given by the hospital as $R = 18$.

Controlling the proportion of time awarded as secondary time involved setting the values of the parameter MS_i, representing the maximum amount of secondary time that could be considered acceptable for Dr. i. To do so, we partitioned the surgeons into three basic ad hoc categories of OR usage—high, medium (med), or low—based on the shape of the histogram of Figure 1, of surgeries performed over the observation period.

As we can see from Figure 1, just over half (66 out of 124) of the surgeons performed 39 or fewer surgeries during the 46-week observation period (less than 1 per week), which provided a natural cutoff for the low-usage groups of surgeons. We found our *medium*-usage cluster in the next three histogram buckets, those surgeons performing more than 39 surgeries but fewer than 153, which accounted for 45 of the surgeons in our study. The tail of the distribution consisted of 13 surgeons, each performing more than 152 surgeries in 46 weeks, which we labeled high usage. This illustrates one difference between a suburban community hospital and an academic medical center;

Figure 1 Caseload Distribution at HDH During the 46-Week Observation Period



here, only a handful of surgeons utilize the OR for more than one full day of surgery per week.

In our package generation routine, we allowed the maximum secondary time to vary for the groups *medium* and high, while always allowing for the possibility of 100% secondary time for a surgeon in the low group. Specifically, for each possible amount of total block time, TB_i, we let $MS_i = MS_{\text{med}}$ for $i \in$ med for each of the values $MS_{\text{med}} \in \{0.3TB_i, 0.4TB_i, 0.5TB_i\}$ and let $MS_i = MS_{\text{high}}$ for each of the values $MS_{\text{high}} \in \{0, 0.1TB_i, 0.2TB_i, 0.3TB_i\}$ for $i \in$ high. Without these bounds (i.e., letting $MS_i = TB_i$ for all i), we found very high utilization but very poor schedules for *medium*-usage and high-usage surgeons. Intuitively, if surgeons are forced to be fully flexible in their acceptance of secondary time, then the hospital will benefit greatly at the expense of convenience in the surgeons' schedules. If, instead, the surgeons are urged to accept some but not all (perhaps only up to 20% or 30%, for example) of their schedule as shared time, then utilization rates improve for the hospital with better schedules for the surgeons.

The selection criteria used here for these three groups was primarily driven by the preferences of HDH. Other groupings are possible and can be set at the discretion of the hospital. Surgeons who perform an average of fewer than one surgery a week should be satisfied sharing their reserved time, as these were surgeons who likely had *no block time* in prior practice and had to schedule irregularly based on very near-term block release dates. More active surgeons, on the other hand, should be guaranteed that a significant portion of their reserved block time should belong exclusively to them. Then, surgeons who perform more than three (or perhaps four) surgeries per week may be considered *very* active and should have an even greater portion of their time reserved as exclusive. This suggested practice is indeed consistent with the view of block time as a reward for active surgeons.

Using these surgeon classes, we generated packages using the technique described in Online Appendix §A.1, using CPLEX 11.1 running on a 32-bit, 2 GHz AMD Turion processor with 2 GB RAM. For each $(MS_{\text{med}}, MS_{\text{high}})$ pair, all packages for all surgeons were generated in under one second. In the most flexible case (i.e., $MS_{\text{med}} = 0.5TB$, $MS_{\text{high}} = 0.3TB$), a total of 14,810 packages were generated, with 9,883 generated in the least flexible case, making the average number of packages generated per surgeon range from approximately 80 to around 120. Individual surgeons generated as few as four packages, with one (outlier) surgeon generating over 1,100 packages.

In contrast to the package generation tool, the optimization of BSP represented a significant (yet manageable) computational difficulty, even using the state-of-the-art CPLEX 11.1 MIP solver. We therefore used a faster 2.66 GHz Intel Core2 duo processor with 2 GB RAM, with an additional 10 GB of "virtual RAM" implemented by saving branch-and-cut tree-node files to the hard drive. Thus the instances of BSP that we solved represent fairly difficult problems, requiring a good deal of memory and time to solve to near optimality. Every instance we solved exhausted the memory supply, with average and worst-case runtimes at 3.2 hours and 4.0 hours, respectively. In no case did we verify global optimality of our solution, but in all cases the final optimality gap was below 1.07%, with an average gap of 0.815%. Thus all schedules were at least 98.93% optimal in expected total benefit to the hospital and surgeons.

4.2. Stage 2: Simulation Details

We created a simulation environment to test the real-time scheduling algorithm described in §3.2. The algorithm was customized to reflect the block-time allocations from the schedules that were obtained during the simulation of Stage 1. The first step was to generate case arrivals for the 124 surgeons. To create a realistic environment, it was necessary to simulate both the number of cases submitted by each surgeon and the duration of each case. This was complicated because (1) the duration of a case is specific to the procedure being performed; (2) for a given procedure, there is significant variation in the amount of time it takes different surgeons to perform the procedure; and (3) with only a few exceptions, most surgeons did not perform enough of any given procedure to provide enough data points to create surgeon-specific, procedure-specific distributions for arrival rate or case duration. To overcome this limitation, we adopted a two-step approach when generating case arrivals. Step 1 was to generate case arrivals for each surgeon i without regard to the type of procedure, based on a Poisson distribution with mean λ_i, because we found that the overall historical case arrival rate followed that distribution. Step 2 was to determine a case duration for each newly arrived case. This was done by first creating a discrete distribution that grouped all prior surgeries for each surgeon in 15-minute intervals. A discrete distribution was necessary because the distribution that results from including multiple types of procedures is multimodal and could not be characterized by a standard continuous distribution. The duration of each newly arrived case was then found by generating a random number between 0 and 100, and then drawing the corresponding duration from the cumulative probability mass function for that surgeon.

For each design point in our first set of simulations, Arena simulated the arrival and scheduling of 410 days of activity, and the simulation progressed in one-day time intervals. The first 10 days of simulation time were not included in the results analysis, because that period was used to initialize the simulation environment. The real-time scheduling algorithm was encoded in Arena, and cases that could not be scheduled by the algorithm were recorded but assumed "lost." The rationale was that because we could not accurately simulate the availability of surgeons, any attempt on our part to schedule the cases would bias the results of the simulation. Instead, we recorded the number of cases that the scheduling algorithm could not schedule for each of the design points and used that information as one of our performance metrics. In contrast, for our second round of simulations, the focus was on a comparison of IBS to the MCA approach, which runs overtime when cases cannot be scheduled in the near term. So in the second round of IBS simulation, we also immediately scheduled a case into overtime when it could not be fit into the schedule, allowing for a direct comparison of the approaches.

All simulations were performed on an IBM ThinkPad with an INTEL Centrino Duo Processor at 2.00 GHz with 2 GB RAM.

4.3. Stage 3: Simulation Details

The CSP problem in Stage 2 of IBS does not fully specify the schedule, so to assign cases start times in particular rooms, we used the CPLEX 10.0 MIP solver to create Stage 1-specific formulations that set the corresponding bounds on the starting times of case ω in a room ρ (i.e., $s_{\omega\rho}$ variables in Online Appendix §A.3) to reflect the block-time allocations in each block schedule. For each day of simulation time, Arena exported a case list that included {case number, case duration, surgeon}. Custom code written in VB.NET was used to create problem instances of the corresponding IP that were imported into CPLEX. Because the problem instances were fairly small (40–60 cases, ≤ 18 rooms), we were able to solve each problem instance to optimality in under two minutes. The output of Stag 3 enabled us to record the utilization of each room on each day and the primary and secondary block-time utilization for each surgeon on each day that they were assigned block time.

4.4. Results: IBS vs. HDH

Creating block schedules for each combination of $MS_{\text{med}} \in \{0.3TB_i, 0.4TB_i, 0.5TB_i\}$ and $MS_{\text{high}} \in \{0, 0.1TB_i, 0.2TB_i, 0.3TB_i\}$ provided 12 design points for the simulation. For each we measured the utilization of staffed time to evaluate the impact on the hospital. To measure the impact on the surgeons, we recorded the number of cases that could not be scheduled and the proportion of total surgical time scheduled into secondary time for each group of surgeons. The rationale for using the latter as a measure of surgeon satisfaction is that because secondary time is shared, it is less convenient. Table 1 shows the number of staffed hours for the entire 10-day block schedule and the average utilization of those hours. Average utilization is obtained from the 40 10-day time periods extracted from the 400-day simulation.

Table 1 shows that, as expected, utilization increases as the proportion of block time allocated in the form of secondary time increases for both high-volume and medium-volume surgeons (all block time allocated to low-volume surgeons was in the form of secondary time). In contrast to the hospital's prior utilization level of approximately 55%, all of the utilization levels obtained here represent significant increases in performance. From the standpoint of quality of service provided to the surgeons, there was *never* an instance where a case could not be scheduled. Table 2 shows the proportion of surgical time that is scheduled in secondary time for each group of surgeons for the different combinations of MS_{med} and MS_{high}.

Table 2 shows that for high-volume and medium-volume surgeons, the proportion of surgical time scheduled in secondary time is consistently lower than the proportion of total time allocated in the form

Table 1 Staffed Hours of OR Time and Utilization

MS_{high}	0	0.1	0.2	0.3	0	0.1	0.2	0.3	0	0.1	0.2	0.3
MS_{med}	0.3	0.3	0.3	0.3	0.4	0.4	0.4	0.4	0.5	0.5	0.5	0.5
Staffed hours	1,092	1,056	1,028	992	1,036	1,000	968	940	964	940	904	868
Utilization (average) (%)	65.66	67.71	69.84	72.18	69.31	71.80	73.97	76.38	74.38	76.28	79.42	82.49

Table 2 Proportion of Surgical Time Scheduled in Secondary Time

MS_{high}	0	0.1	0.2	0.3	0	0.1	0.2	0.3	0	0.1	0.2	0.3
MS_{med}	0.3	0.3	0.3	0.3	0.4	0.4	0.4	0.4	0.5	0.5	0.5	0.5
High volume (%)	0.0	2.0	5.0	16.0	0.0	2.1	4.9	16.1	0.0	1.9	4.9	15.8
Medium volume (%)	15.8	16.1	15.9	15.9	22.3	22.9	23.1	23.1	31.1	33.4	31.3	31.2
Low volume (%)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

of secondary time. This is because the total amount of surgical time granted to the surgeon in the package generation tool (see the online appendix) is between $TB_i^- = \mu_i + 2\sigma_i$ and $TB_i^+ = 1.1 \max_{t\in T}(a_{it})$—that is, between two standard deviations above expected use and 110% of the busiest cycle the surgeon had in our data set. The generous time allowance, combined with the Stage 2 scheduling algorithm that actively seeks to schedule cases in primary time, explains the relatively low use of secondary time. In addition, some of the secondary time that is utilized is in the context of scheduling a case in a manner that spans two contiguous bins, where the surgeon has primary time in one bin and secondary time in the other. In those cases secondary time is used to enable the surgeon to achieve full utilization of primary time.

After reviewing the simulation results, HDH felt that the schedule resulting from $MS_{\text{high}} = 0.2TB_i$, $MS_{\text{med}} = 0.5TB_i$ provided the most desirable tradeoff between OR performance and quality of service. Although the schedule obtained from $MS_{\text{high}} = 0.3TB_i$, $MS_{\text{med}} = 0.5TB_i$ had better utilization and a relatively small proportion of surgical time booked in secondary time, it resulted in daily schedules that were fragmented. This took the form of high-volume surgeons coming to the hospital to perform a single case in secondary time. The problem was not evident with the medium-volume surgeons, as they tended to only perform one or two cases on a given day. Based on that selection, we were able to compare the effectiveness of IBS with HDH's prior approach. In terms of surgeon access, IBS provides predictable access to the OR for all surgeons. Under the prior approach, only 51 of the 124 surgeons (41.1%) had been assigned block time. In addition, for HDH's preferred design point, there were no instances where a case could not be scheduled.

We created two additional simulations within this setup to estimate the expected improvement in underutilization of staffed OR time. The first used IBS and HDH's preferred design point. The second scheduled cases based on HDH's most recent block schedule and their scheduling policy at that time. Note that the scheduling policy used by HDH was an "any workday" policy, where no case is ever turned away. Also note that because the prior block schedule called for all suites to be open, we were not able to look at the cost of overutilization (i.e., the additional cost incurred when planning to staff 12 rooms but ultimately needing 14 rooms or keeping a room open until 7 P.M. instead of 3 P.M.) without making additional assumptions. For both simulations we collected the number of suites used each day and how many hours of surgery were performed each day. The results are shown in Figures 2 and 3.

Figures 2 and 3 show that under IBS, we are able to consistently schedule a given volume of surgery into fewer rooms than was previously required. In addition, there is far less variance in the number of rooms needed to provide a given volume of services. For example, using the prior approach, 60 hours of surgery required between 10 and 14 rooms, whereas under IBS, only 9 to 11 rooms were needed. This reduction in the average number of rooms and the variance reflects the more efficient packing of cases into the surgical schedule, greatly reducing the "swiss cheese" problem that has arisen when only certain surgeons

Figure 2 Simulated OR Usage Based on HDH's Prior Block Schedule and Case-Scheduling Rules

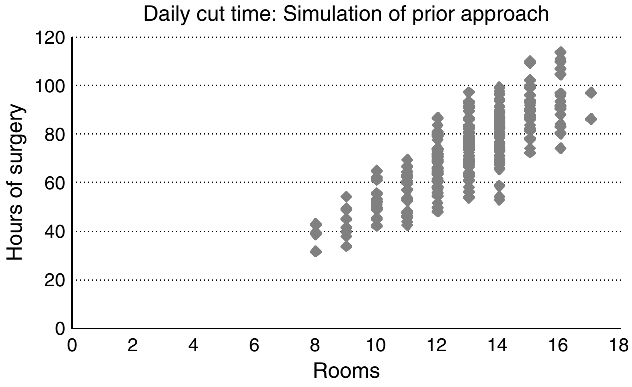
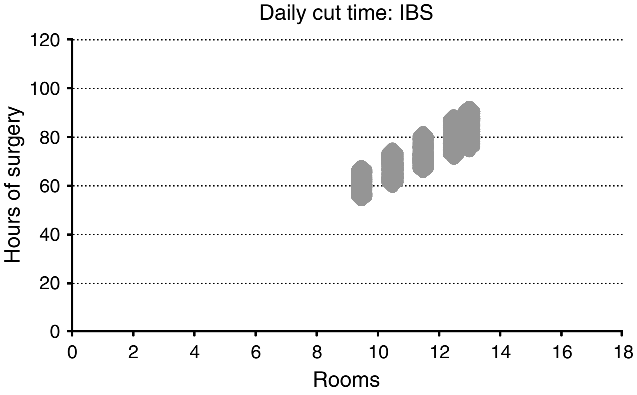


Figure 3 Simulated OR Usage Based on IBS

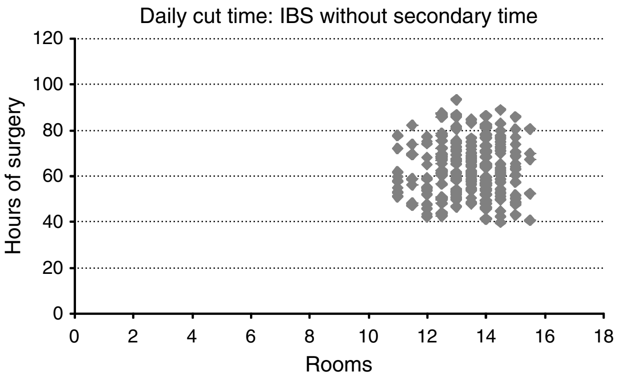


had access to the OR and the rest had to wait until the block release date had passed before they could schedule cases. Looking at the number of hours of surgery performed in a given number of rooms yields a similar result. Under the prior approach, 12 rooms were used to provide between 44 and 85 hours of surgery. Under IBS, 12 rooms provided between 66 and 82 hours of surgery. We also see evidence that load balancing across the two-week scheduling horizon had the effect of smoothing demand over time. Under the prior HDH approach, the busiest day involved 116 hours of surgery in 16 rooms. Under IBS, the busiest day involved 93 hours of surgery in 13 rooms. At the same time, under the old approach, the slowest days involved 38–45 hours of surgery in 8–10 rooms; under IBS, the slowest days involved 52–55 hours of surgery in 9 rooms.

A major contribution of our research to the OR scheduling literature is to introduce a viable mechanism for the allocation of secondary time. To evaluate its impact, a final simulation was performed that compared the performance of the block schedule used above with a block schedule that does not incorporate secondary time. That is, we generated schedule packages where surgeons would not accept secondary time. In that environment, as expected, a number of surgeons (44 out of the 124 surgeons considered in our data) do not receive block time. Therefore, the scheduling algorithm in Stage 2 was modified to include the so-called "block release date," under which all unused OR time on a given day becomes available for use by any surgeon three days prior to that day. (For this simulation we used a block release policy of three days, though this policy varies across hospitals and surgical specialties in practice, usually from one to five days.) Arrivals were generated in the same manner as in the first simulation, and the cases scheduled for each day were assigned rooms and start times. The utilization results are shown in Figure 4.

Without including secondary time, room utilization after Stage 3 room consolidation was 60% on average. This suggests that although block-time governance mechanisms vary, achieving consistently high levels of utilization is not possible if hospitals continue the current practice of offering only exclusive block time to only some surgeons. Comparing this figure with Figure 3, we see that not only does the overall utilization of staffed rooms improve dramatically with the introduction of shared secondary time (performing the same amount of surgery in 9.5 to 13 rooms per day as opposed to 11 to 15.5 rooms per day), but when incorporated into the larger three-stage framework, the introduction of secondary time can facilitate a drastic reduction in scheduling variance, maintaining a more consistent workload for the OR staff. Most importantly, though, in addition to yielding these benefits for the hospital, secondary time provides consistent access to all surgeons who could not be given block time when using only primary time (over one-third of the surgeons in the study), offering categorically new scheduling consistency and predictability for these surgeons and their patients.

Figure 4 Simulated Usage When Secondary Time Is Fixed to Zero Under IBS



4.5. Results: IBS vs. MCA

Our first round of simulations demonstrated significant benefits of IBS compared with an ad hoc block-scheduling procedure based on common practice. In a second round of simulations, we implemented a newsvendor-based approach for additional comparison. Whereas HDH's historical approach may be viewed as unscientific and thus easy to improve upon, the MCA newsvendor-based model offers a more rigorous approach to the allocation of block times, making it a relevant basis for comparison. Combining the ideas from Strum et al. (1999) and Spangler et al. (2000), we simulated an MCA approach to determine the amount of block time to award to surgeons. As in that literature, block time is allocated only as exclusive time, and based on the practice of larger academic hospitals, is allocated to service lines rather than to individual surgeons.

With a few of the details of this approach not fully specified in the existing literature, we made some assumptions allowing for the most direct comparison to IBS. The result is the following procedure, which is henceforth referred to as MCA. According to Strum et al. (1999), block time should be allocated to a service/subspecialty such that $P(X_i \leq B_i^*) = C_o^i/(C_o^i + C_u^i)$, where X_i is the random volume of surgical time for service line i; C_o^i and C_u^i are overutilization and underutilization costs, respectively; and B_i^* is the optimal amount of block time to offer service line i in their model. This critical step appears in Step 3 below.

Step 1. Calculate the amount of OR time (in hours) for each service line for each scheduling period from our historical data. For consistent comparison with IBS, two-week scheduling periods were used.

Step 2. Calculate the mean and standard deviation of usage for each service line.

Step 3. Determine the ratio $C_o^i/(C_o^i + C_u^i)$. We assumed overtime costs are 50% higher than regular operating costs, yielding $1.5/2.5 = 0.60$. Using a normal distribution of X_i, determine a Z-statistic associated with this probability $= 0.60$—in this case, $Z \approx 0.25$. (Historical distributions satisfy the Kolmogorov–Smirnov test for goodness of fit with a normal distribution. These results are consistent with Strum et al. 1999.)

Step 4. Calculate allocations for each service line based on this Z-statistic. For example, if a service line had an average usage of 100 hours with a standard deviation of 40 hours, the service would be allocated $100 + (40 \times 0.25) = 110$ hours of OR time.

Step 5. Calculate the number of rooms allocated to each service line. Since our algorithm was based on $l = 4$-hour scheduling bins, we divided the allocation for the service line by four and rounded up. (This detail allows for a direct comparison to IBS, which opens rooms in four-hour blocks, and rounding up is consistent with the spirit of the MCA literature, that overutilization is more costly than underutilization.)

Step 6. Allocate room blocks to days across the 10-day scheduling period, based on closeness to historical usage patterns.

Step 7. Allocate individual block times to surgeons. The MCA literature suggests that high-volume surgeons receive an individual allocation equal to their expected usage. The remaining time is reserved for above-average demand from high-volume surgeons and demand from low-volume surgeons.

This seven-step procedure was used to devise a block schedule under the MCA approach, and then a random arrival process was simulated as in our first simulations. IBS was then implemented on the random arrival stream using the same procedures as described for the first round of simulations, whereas MCA used the following case-scheduling paradigm, consistent with the descriptions given in the MCA literature.

- Cases for surgeons with individual block-time allocations were scheduled into the next available block.
- Once surgeons with block times reach their time allocations, additional cases are held in queue until the block release date (three days prior to the day of surgery).
- Cases for surgeons without individual block-time allocations are held in queue until the block release date.
- Once the block release date arrives, cases held in queue are scheduled, first filling in free (underutilized) time and then putting cases into overtime (no cases are turned away).

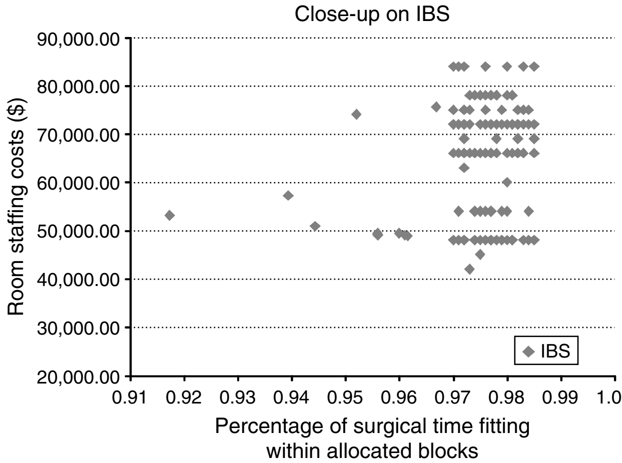
This process was repeated using six different historical usage patterns, with an increasing number of surgeons for the same fixed hospital size, simulating an increasingly congested scheduling environment. The first run used HDH's historical loads, and larger hospitals were simulated by "cloning" randomly selected surgeons until the expected volume increased by a 10% increment. These runs are referred to as Base, Base + 10%, ..., Base + 50%, and the results are summarized in Table 3. Each data point represents the average performance over 150 simulated days of surgery.

Table 3 provides a clear general comparison of the two approaches. Although IBS tends to plan ahead to open and staff more rooms for the same expected surgeries, the better load balancing over the two-week look-ahead typically results in slightly less underutilization, on average, and drastically less overutilization. Intuitively, for the block-time and block release date policy environment in which MCA operates, surgeries that cannot be squeezed into the next few days of regular hours cannot be pushed further forward into the future, as blocks further than three days forward have not yet been released. The result is that these surgeries are performed in overtime, resulting in greater overutilization. IBS, on the other hand, has a coordinated plan for block sharing that does have some time reserved for the surgeon in the next 10 business days, making it easier to spread out over this horizon. So only surgeries that cannot be squeezed into regular block time or the near-term released block time of others need to be handled as overtime cases under IBS.

Table 3 Comparison of IBS and MCA Under Six Congestion Scenarios

	Common arrivals		Rooms		Staffed time		Underutilized time		Overutilized time	
Daily averages	Procedures	Hours	IBS	MCA	IBS	MCA	IBS	MCA	IBS	MCA
Base	36.0	73.5	10.6	10.2	84.5	81.6	11.10	12.45	0.12	4.40
Base + 10%	38.5	78.8	11.4	11.2	91.4	89.6	12.69	14.16	0.06	3.32
Base + 20%	43.8	88.8	12.8	12.4	102.7	99.2	14.00	15.66	0.11	5.27
Base + 30%	47.4	97.1	14.0	13.3	112.1	106.4	15.17	15.62	0.14	6.15
Base + 40%	52.0	104.5	15.3	14.3	122.2	114.4	17.80	16.20	0.09	6.26
Base + 50%	54.9	110.0	16.2	15.1	129.3	120.8	19.40	17.35	0.08	6.56

Figure 5 One Hundred Fifty Days of Simulated Surgeries for IBS



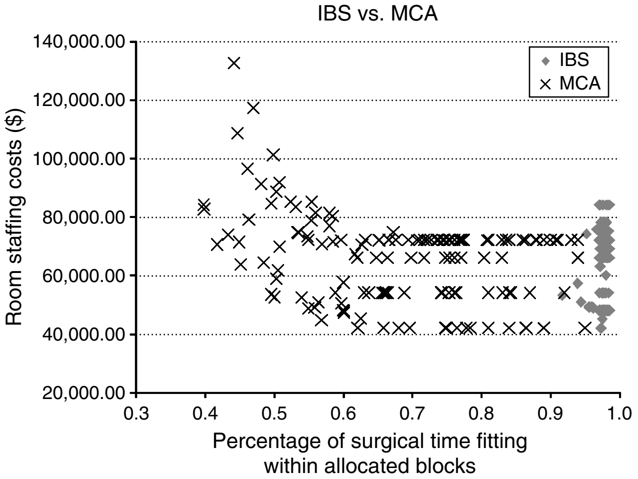
To most clearly demonstrate the win–win nature of the IBS approach, consider Figure 5, which plots each simulated day of surgery for the IBS approach under the Base load, indicating room staffing costs for the day on the vertical axis and percentage of surgical time performed during time preallocated to that surgeon as block time on the horizontal axis. Clearly, reducing room staffing costs represents the interests of the hospital for the same fixed set of randomly generated cases; however, surgeons would prefer to perform surgery during preallocated blocks of time. Thus in this picture, down-and-to-the-right (lower cost and more surgery done in block time) represents the Pareto-improving direction.

Room staffing costs include the fixed $3,000 cost of opening a room for a half day, plus a per-hour overtime charge at time and a half. The percentage of surgical time performed in block time is less than 100% because of cases that were squeezed into overtime or released block time, as well as cases that ran long and spilled out of the block time. Looking at Figure 5, the most striking pattern that emerges is the presence of visually apparent horizontal lines. Days that exactly line up on a horizontal line are days in which costs were strictly integer multiples of the $3,000 room opening cost and are thus days in which overtime charges did not occur. On any day for which overtime did occur, first the scheduled room opening cost is incurred, and then the point moves up and to the left as more overtime occurs, indicating that costs increase (upward movement) and more out-of-block surgery (left movement) is occurring.

Figure 6 plots each simulated day of surgery for the MCA and IBS approaches under the Base load, again indicating room staffing costs for the day on the vertical axis and percentage of surgical time performed during time preallocated to that surgeon as block time on the horizontal axis.

Considering both Figures 5 and 6, a few striking patterns are visible. The first is that MCA incurs drastically more overtime days (53 compared to 10 for this Base load scenario), apparent by their scattering away from the horizontal line tendency. Also, much of the overtime that does occur under MCA results in very expensive days, in which many overtime hours are performed *and* many rooms are open. For IBS, in contrast, overtime days occur only when a small number of rooms were planned to be open, and so the highest daily cost is much more contained. Further, in the case of MCA, the percentage of surgical time performed within previously scheduled block time is brought down by all surgical time by surgeons not receiving any block time, clearly accounting for the wide horizontal range of values for MCA compared with the tightly compact range of values for IBS.

Figure 6 One Hundred Fifty Simulated Days of Surgery Under IBS and MCA, Comparing the Costs to the Hospital and Blocking Convenience for Surgeons (Base Load Scenario)



The other visible pattern is that costs are roughly the same between the two approaches if considering only nonovertime days, with a slight advantage for MCA, as it does not plan to open as many rooms as IBS sometimes does. But overall, this foresight to open more rooms and spread out (load balance) according to an all-inclusive block schedule gives an overall cost advantage to IBS. IBS's total underutilization and overutilization costs are 59% of the MCA costs in the Base load scenario and 65% over all congestion scenarios. Though this cost gap does steadily diminish with increasing amounts of congestion, the corresponding graphical figures for the other scenarios do not tell a drastically different story and have thus been omitted to conserve space. We note, though, that even in the most congested case considered, Base + 50%, the IBS underutilization plus overutilization costs are still 72% of the MCA costs, indicating cost domination by IBS under all congestion scenarios considered.

4.6. Robustness of IBS to Unexpected Arrivals

The IBS technique generates and evaluates block schedules based on historic usage, implicitly expecting that past surgical volume is the best indicator of future usage. But because additional block time may attract more surgeries to the hospital, it is important to consider the possibility of increased arrival rates beyond those observed in historical data. Intuitively, because some surgeons were previously not offered block time, they associated with a few different local hospitals, performing cases at each, depending on the availability of time at each hospital. But now, offered block time at the current hospital, surgeons may in fact divert a larger percentages of their cases to where they now have block time. Because medium- and high- usage surgeons were typically already performing all or almost all of their cases at the hospital under investigation, we expect this kind of effect to be most prevalent among low-usage surgeons.

Thus we ran a final round of simulations in which block schedules were generated based on historic arrival rates of cases, but where arrival rates were drastically increased in the simulations. We did this in two parts, first rerunning 150 simulated days where low-volume surgeons increased their arrival of cases by 50%, and then by 100%. Note that in this case, low-volume surgeons account for 24% of total volume. As a result, 50% and 100% increases in volume from this group of surgeons translate into 12% and 24% increases in total volume, respectively. The results of these simulations are shown in Figures 7 and 8, respectively.

As expected, this unanticipated increase in volume results in more overtime days, from 6.7% to 21.3%, but as we see in Figure 7, the daily cost is still never more than the cost of opening 14 rooms all day with no overtime (that is, the highest horizontal cluster, occurring at cost = $84,000, is the day with the planned highest cost, and no overtime day exceeds this cost).

Figure 7 **One Hundred Fifty Simulated Days of Surgery Under an Unanticipated 50% Increase in Case Arrival Rates for Low-Usage Surgeons**

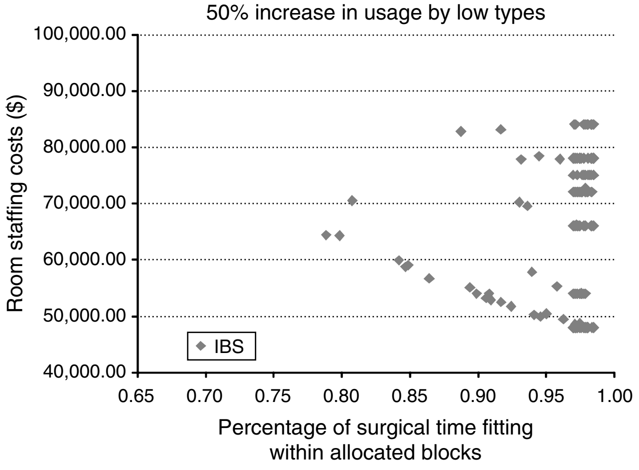
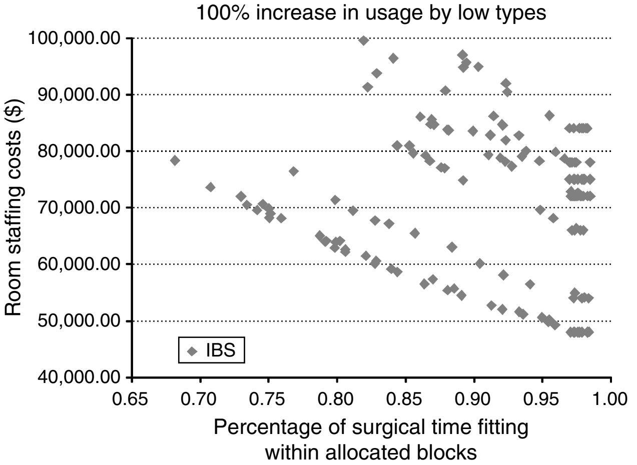


Figure 8 **One Hundred Fifty Simulated Days of Surgery Under an Unanticipated 100% Increase in Case Arrival Rates for Low-Usage Surgeons**



It is only when we push the low-volume arrival rate to double its historical average (depicted in Figure 8, which shows 64% overtime days) that we see any days with costs exceeding the cost of the highest planned day, including all of our previously described simulations of IBS. This suggests that IBS is robust (in terms of maximum daily cost) to an unexpected increase in arrival rates as long as the increase is not too drastic. Therefore in practice, the hospital administration should make efforts to determine how much business low-usage surgeons may transfer from other hospitals if given block time before running the BSP, or perhaps bound the number of cases they are allowed to bring in above their historical rate. But given these results and the profitability of performing more cases (these figures reflect only costs), the hospital should be willing to squeeze a few extra cases into overtime for surgeons exceeding their historical rates. In equilibrium over the long term, however, new arrivals will be assimilated into the planning phase, when the BSP is periodically re-solved with updated data, and new block schedules reflecting increased volume will eliminate these short disruptions of unexpected demand.

5. Conclusions and Future Research

In this paper we frame the problem of providing surgeons with predictable, reliable access to the OR while maintaining high levels of OR utilization as a three-stage process, based on the three distinct information epochs that appear to be essential parts of the decision-making process. In the first stage, surgeons are assigned regularly recurring blocks of time in the OR, time that may be exclusively for a given surgeon or shared among a small group of surgeons. The second stage involves the real-time scheduling of cases into the blocks of time the surgeons have been granted over a rolling two-week scheduling horizon and identifying a schedule that is feasible in an expected sense. Finally, in the third stage, the cases for each day are assigned to specific rooms and given specific start times. A comprehensive simulation was conducted to evaluate the performance of the three-stage mechanism in comparison against actual historical performance data provided by HDH, showing drastically less scheduling variability and more efficient use of hospital resources.

A second set of simulations showed that there are advantages to the IBS approach compared to the related MCA approach that has been suggested in the literature. The most drastic difference is in the amount of surgery that can be performed within block time, indicating that the biggest winners would be the surgeons—in particular the low-usage surgeons, who could count more dependably on having access to the hospital at regular predefined times. The hospital also experienced a 35% reduction in the cost of poor utilization under IBS relative to MCA, indicating a Pareto improvement from this benchmark. Observing that the cost difference comes primarily by reducing overtime cost, the primary advantage of IBS is its ability to facilitate load balancing over an entire two-week scheduling horizon, in contrast to previous practice in which the block release policy and lack of block time for all surgeons made this type of load balancing impossible.

Thus the main contribution of this paper is the development of a practically viable system for providing all surgeons at least some recurring time on the hospital schedule, in such a way that the system can be expected to perform without conflict in expectation, captured here as the BSP. Other stages of the problem are discussed for completeness and for the sake of evaluating the entire OR scheduling system, but this portion of the problem proved difficult, both theoretically and computationally, and required pushing CPLEX fairly hard to get nearly optimal solutions. Further research may improve the estimation of cost parameters to obtain better solutions, and it is possible that the introduction of more sophisticated package-generation routines may also improve overall performance. In particular, the bounds on secondary time awards for each surgeon class were selected in an ad hoc manner through simulation trial and error and consultation with a real hospital; a more systematic approach to measure the value to a surgeon and to the hospital of changes in these parameters is open to further study. Still, we have demonstrated excellent performance of the IBS system with these heuristic design choices, and a more systematic method for setting these parameters can only improve performance. The flexibility of the BSP to accept *any* method of package generation makes it particularly promising that these techniques can be adapted to nearly any hospital environment.

A subsequent implementation of IBS at HDH resulted in improved OR suite utilization, and this in turn enabled the hospital to recruit new surgeons. A more comprehensive description of this implementation is provided in Online Appendix §A.4, but we note here that so far, the newly recruited surgeons are generating $1,200,000 in yearly operating income for the hospital, and HDH continues to consider the recruitment of new surgeons. The decrease in the number of rooms needed to provide surgical services, even with higher volume, has also enabled HDH to shift staff and reduce the expected increase in staffing costs associated with their plan to open an outpatient surgical center.

In addition, the package generation and simulation techniques described in this work can be further developed into decision support tools that provide hospital administrators with the ability to evaluate the impact of capacity expansion initiatives and physician recruitment efforts and the accessibility and performance of the OR. In the meantime, we continue to explore the application of the ideas presented here in the real world, with goals to expand their application among the subsidiaries of HCA and at other peer institutions.

Electronic Companion

An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/msom.1110.0372.

References

Dexter, F., A. Macario. 2002. Changing allocations of operating room time from a system based on historical utilization to one where the aim is to schedule as many surgical cases as possible. *Anesthesia Analgesia* **94**(5) 1272–1279.

Dexter, F., J. Ledolter, R. Wachtel. 2005. Tactical decision making for selective expansion of operation room resources incorporating financial criteria and uncertainty in subspecialties' future workloads. *Anesthesia Analgesia* **100**(5) 1424–1432.

Dexter, F., A. Macario, R. D. Traub, D. A. Lubarsky. 2003. Operating room utilization alone is not an accurate metric for the allocation of operating room block time to individual surgeons with low caseloads. *Anesthesiology* **98**(5) 1243–1249.

Houdenhaven, M. V., J. M. van Oostrum, E. W. Hans, G. Wullink, G. Kazemier. 2007. Improving operating room efficiency by applying bin-packing and portfolio techniques to surgical case scheduling. *Anesthesia Analgesia* **105**(3) 707–714.

Jackson, R. L. 2002. The business of surgery: Managing the OR as a profit center requires more than just IT. It requires a profit-making mindset, too. *Health Management Tech.* **23**(7) 20–22.

McKesson Corporation. 2002. Achieving operating room efficiency through process integration. Online report for the Healthcare Financial Management Association (HFMA), San Francisco. Accessed January 5, 2009, http://www.allbusiness.com/health-care-social-assistance/491677-1.html.

Spangler, W. E., D. P. Strum, L. G. Vargas, J. H. May. 2000. A minimal cost analysis model for utilization and capacity planning in surgical services. *Internat. J. Healthcare Tech. Management* **2**(1–4) 56–70.

Strum, D. P., L. V. Vargas, J. H. May. 1999. Surgical subspecialty block utilization and capacity planning: A minimal cost analysis model. *Anesthesiology* **90**(4) 1176–1185.