Online Appendix

A.1 Package Generation

Because the total enumeration of all feasible packages for each surgeon results in an intractably large instance of the BSP, it is necessary to generate only a manageable subset of packages for each surgeon. At this point we can apply further criteria to prune the list of potential packages of block time, simultaneously reducing the computational difficulty of solving the BSP (by reducing the overall size of the formulation) and allowing us to better reflect surgeons' preferences by eliminating packages that wouldn't easily satisfy their needs. In this section we describe the few considerations that we used in our simulations to do just that, though we emphasize that this pruning step (i.e., only generating "good" packages from the domain of possible packages) may be modified by any hospital attempting to apply these techniques. Further pruning (more strict rules on package generation) could potentially increase surgeon satisfaction and further reduce computational run times, but could also lead to worse performance from the hospital's perspective.

The first step in our package generation procedure is to determine a reasonable range for the *total* amount (TB_i) of block time in a package, for a given surgeon over a two-week period. (In this subsection, we will focus on generating one package at a time for a single surgeon, so for simplicity our notation omits package and surgeon identifiers in the subscript where convenient.) As a first heuristic for this problem, we set $TB_i^- = \mu_i + 2\sigma_i$, where TB_i^- is the minimum amount of total block time for surgeon i, and μ_i and σ_i are the mean and standard deviation, respectively, of a_{it} over all past periods t. Similarly, we set $TB_i^+ = 1.1 \max_{t \in T}(a_{it})$, where TB_i^+ is an upper bound on TB_i. (Note that a given hospital is free to set the quantities TB_i^- and TB_i^+ in whatever manner they deem appropriate; the values used here reflect the preferences of HDH.)

For the second step we recognized that certain surgeons often require block time in consecutive morning and afternoon bins during the same day, in order to accommodate longer surgeries. (This was motivated by examining the historical data.) Thus we decided it was appropriate to generate packages with a varying number of guaranteed *long days*, defined as collections of primary and secondary time in the same day. At least one of the two bins of that day is completely designated as primary time, and the total amount of primary and secondary time in the day is at least as long as the longest single surgery observed from this surgeon.

Letting δ designate the total number of long days in a package, it was again appropriate to generate a heuristic minimum and maximum number of long days to consider in our package generation procedure, similarly denoted δ_i^- and δ_i^+, respectively. Thus, to generate packages with a reasonable number of long days, we let δ_i^- equal the historical average number of days in a two-week period for which surgeon i had a surgery of five or more hours, and δ_i^+ equal the historical maximum number of days in a two week period for which the surgeon had a surgery of five or more hours. Additional new notation is as follows:

Additional Parameters:
short: the length of the shortest observed surgery for this surgeon, rounded up to the nearest whole hour.
long: the length of the longest observed surgery for this surgeon, rounded up to the nearest whole hour.
Additional Decision Variables:
p_j: the amount of primary time awarded in bin j as part of a feasible solution.
s_j: the amount of primary time awarded in bin j as part of a feasible solution.
y_j: a binary variable equal to one if and only if the package includes positive primary time in bin j.
z_j: a binary variable equal to one if and only if the package includes positive primary or secondary time in bin j.
γ_j: a binary variable equal to one if and only if bin j is filled with primary time as part of a long day.

In order to generate a large but reasonable number of packages for each surgeon, we only considered those for which $TB_i \in [TB_i^-, TB_i^+]$, and $\delta \in [\delta_i^-, \delta_i^+]$. These restrictions allowed us to narrow our package generation considerably, reflecting a mild assumption that surgeons will find packages acceptable for only a few values of TB_i and δ. We first find a feasible package of block times that minimizes the total number of bins used for each pair $TB_i \in [TB_i^-, TB_i^+]$ and $\delta \in [\delta_i^-, \delta_i^+]$, by finding solutions to the following MIP called PG:

$$Z(TB_i, \delta) = \min \sum_{j \in B} z_j \tag{PG}$$

$$\text{s.t.} \quad \sum_{j \in B} (p_j + s_j) = TB_i \tag{1}$$

$$\sum_{j \in B} \gamma_j = \delta \tag{2}$$

$$p_j \le l\ y_j \qquad j \in B \tag{3}$$

$$p_j \ge short\ y_j \qquad j \in B \tag{4}$$

$$p_j + s_j \le l\ z_j \qquad j \in B \tag{5}$$

$$p_j \ge l\ \gamma_j \qquad j \in B \tag{6}$$

$$p_j + s_j + p_{j+1} + s_{j+1} \ge long\ \gamma_j \qquad j \in B,\ j \text{ odd} \tag{7}$$

$$p_{j-1} + s_{j-1} + p_j + s_j \ge long\ \gamma_j \qquad j \in B,\ j \text{ even} \tag{8}$$

$$\gamma_{j-1} + \gamma_j \le 1 \qquad j \in B,\ j \text{ even} \tag{9}$$

$$s_j,\ p_j \ge 0 \qquad j \in B \tag{10}$$

$$y_j, z_j, \gamma_j \in \{0, 1\} \qquad j \in B. \tag{11}$$

The objective function of (PG) represents the total number of bins for which this surgeon would receive a positive amount of block time, with minimization reflecting that for a fixed amount, TB_i, of block time a surgeon would prefer visiting the hospital in as few distinct trips as possible. Constraints (1) and

(2) fix the total amount of block time and the number of long workdays, thus ensuring that the parameters TB_i and δ are properly defined. Constraints (3) and (4) guarantee that if any primary time is awarded in a particular bin, it is at least enough time to perform a short procedure, but no longer than the length of a whole bin. The total amount of primary and secondary time in a bin is bound by (5) to be no longer than can be fit in a bin.

Constraints (6)-(9) govern the γ_j variables and the amount of time required to qualify as a long day. Constraint (6) ensures that a long day is only tallied if at least one "full" bin of primary time is awarded, while (7) and (8) ensure that whether an odd (morning) block or an even (afternoon) block serves as the "full" block of primary time, the total amount of block time in the day is long enough to perform the longest possible surgery. Constraint set (9) ensures that filling both the morning and afternoon blocks with primary time only counts as one long day, and not two.

We initially created schedules based on the formulation specified as problem (PG), using CPLEX 11.1's *solution pool feature* to generate a number of feasible solutions, but found that the solutions often included too many bins with positive secondary time. The result was that the assignment of secondary time was scattered throughout the period, and surgeons sharing these blocks often ended up with large gaps in their schedules. Working under the premise that surgeons would instead prefer to consolidate assigned secondary time as much as possible, our approach then became that of a two-stage optimization. Having found $Z(TB_i, \delta)$, we next added a constraint (12) to lock in this optimal objective value. Then we added a new set of constraints and a new objective to find (among the optimal solutions to problem (PG) a solution that minimizes the number of bins with positive secondary time. The result is the following second stage optimization:

Additional Parameter:

S: the total amount of secondary time in a package.

Additional Decision Variable:

w_j: a binary variable equal to one if and only if the package includes positive secondary time in bin j.

$$\min \sum_{j \in B} w_j \qquad \text{(PG}')$$

$$\text{s.t.} \quad (1) - (11)$$

$$\sum_{j \in B} z_j = Z(TB, \delta) \qquad (12)$$

$$\sum_{j \in B} s_j = S \qquad (13)$$

$$s_j \leq l \; w_j \qquad j \in B \qquad (14)$$

$$w_j \in \{0, 1\} \qquad j \in B \qquad (15)$$

where we allowed the additional parameter S to vary between zero and an

upper bound on secondary time, MS_i, more details of which are given §4.1. We then used CPLEX 11.1's solution pool feature to generate up to 10 solutions to (PG′) for each surgeon and (TB_i, δ, S) combination. The entire set of these solutions for a single surgeon was then used as the list of acceptable packages for that surgeon. Again, this two-stage optimization (first solving (PG) and then solving (PG′)) served as a simulation of surgeon preferences, where in real life, a surgeon could effectively use *any* criteria or technique for generating a set of acceptable packages. Eventually, we refined this package generation tool further, as described in §4.1, to limit the proportion of secondary time that could be awarded to higher volume surgeons, consistent with the premise that the surgeons prefer the convenience of primary time, and that the hospital is willing to use primary time as a reward for high usage.

A.2 The heuristic used for the Case Scheduling Problem in our simulations.

The case scheduling problem is to schedule cases for a specific day as the block schedule created in Stage 1 recurs over time. While we assume that cases must be scheduled within 10 days of the request, requests to schedule cases further out can be easily accommodated. Figure 9 provides a graphical representation of the online scheduling algorithm. The basic inputs to the algorithm are the block schedule and the amount of primary and secondary block time *remaining* for each surgeon in each bin. The scheduling algorithm is triggered by a request by a surgeon to schedule one or more cases, although cases are scheduled individually. The algorithm moves through a series of steps attempting to schedule the case within 10 days, with an emphasis on maximizing the utilization of primary time. When a case is scheduled the surgeon's available remaining time in that bin (or bins) is updated accordingly. In Figure 9, n is initially set to two days after the day the case request arrives (because a case typically require the patient to go through pre-operative test, not eat the night before, etc.), and c is a counter variable, initially set to one, telling how many days into the future the algorithm is attempting to schedule the case.

In essence, the algorithm makes two loops through the ten-day scheduling window, first attempting to schedule the case in a manner that makes use of the surgeon's primary time. The second loop attempts to schedule cases using only secondary time. Since it is possible that a surgeon has block time in both bins on a given day, the scheduling algorithm looks for opportunities to schedule cases in a manner that spans bins. If the algorithm terminates without scheduling a case, then the unscheduled case is handled by the OR manager or scheduler and scheduled for a date/time based on surgeon availability. These manual exceptions could occur as a result of unusually high demand from a surgeon, or an unusually large number of very long cases that exceed block-time duration. The output of Stage 2 is a schedule that is feasible using the expected number of open rooms given by the Stage 1 block schedule. However, if fewer cases than expected are scheduled on a given day, then the actual "day of" schedule could require fewer rooms than expected, allowing for extra rooms to be diverted to

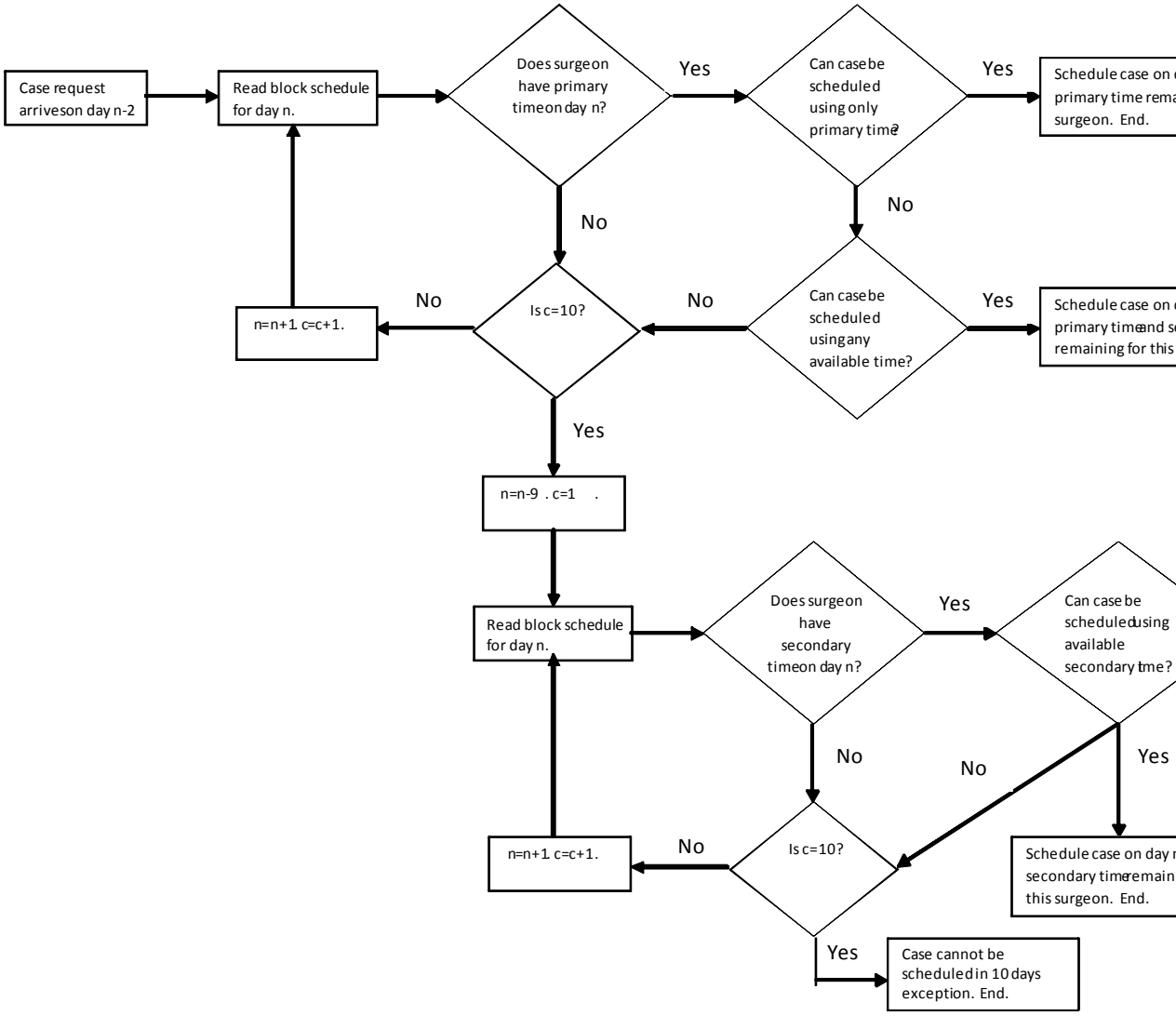

Figure 1: Online scheduling

other needs, for example as crash rooms for emergent cases.

A.3 Details of the Room Consolidation Problem

We first introduce some notation:
Parameters:
Ω_i: the set of cases scheduled on the day in question for surgeon i.
Ω: the set of all cases to be scheduled on the day in question, i.e., $\Omega = \bigcup_{i \in D} \Omega_i$.
Individual cases will be indexed by $\omega \in \Omega$.
P: the set of all rooms that could potentially be open during the day in question.
Individual rooms will be indexed by $\rho \in P$.
α_ω: the expected duration of case ω, inclusive of pre- and post-op times.
Decision Variables:

(Note: we re-use decision variables s, x, y, and z, but with greek subscripts in the current context. This additional usage should not cause confusion.)

$$q_{\omega_1\omega_2} = \begin{cases} 1 \text{ if case } \omega_1 \text{ precedes case } \omega_2 \\ 0 \text{ otherwise} \end{cases} \qquad (\omega_1, \omega_2) \in \Omega^2,\ \omega_1 \neq \omega_2$$

$$s_{\omega\rho} = \begin{cases} \text{starting time of case } \omega, \text{ if case } \omega \text{ is scheduled in room } \rho \\ 0 \text{ otherwise} \end{cases} \qquad (\omega, \rho) \in \Omega \times P$$

$$x_\rho^{am} = \begin{cases} 1 \text{ if room } \rho \text{ is opened in the morning} \\ 0 \text{ otherwise} \end{cases} \qquad \rho \in P$$

$$x_\rho^{pm} = \begin{cases} 1 \text{ if room } \rho \text{ is opened in the afternoon} \\ 0 \text{ otherwise} \end{cases} \qquad \rho \in P$$

$$y_{\omega\rho} = \begin{cases} 1 \text{ if case } \omega \text{ is scheduled in room } \rho \\ 0 \text{ otherwise} \end{cases} \qquad (\omega, \rho) \in \Omega \times P$$

$$z_{\omega_1\omega_2} = \begin{cases} 1 \text{ if cases } \omega_1 \text{ and } \omega_2 \text{ are scheduled in different rooms} \\ 0 \text{ otherwise} \end{cases} \qquad (\omega_1, \omega_2) \in \Omega^2,\ \omega_1 \neq \omega_2$$

The MIP formulation of the RCP is then:

$$\min \sum_{\rho \in P} (x_\rho^{am} + x_\rho^{pm}) \tag{16}$$

$$\text{s.t.} \quad \sum_{\omega \in \Omega} \alpha_\omega y_{\omega\rho} \leq l(x_\rho^{am} + x_\rho^{pm}) \qquad \rho \in P \tag{17}$$

$$\sum_{\rho \in P} y_{\omega\rho} = 1 \qquad \omega \in \Omega \tag{18}$$

$$s_{\omega\rho} \geq y_{\omega\rho} \qquad (\omega, \rho) \in \Omega \times P \tag{19}$$

$$2l y_{\omega\rho} \geq s_{\omega\rho} \qquad (\omega, \rho) \in \Omega \times P \tag{20}$$

$$s_{\omega\rho} \geq (l+1)(x_\rho^{pm} - x_\rho^{am} + y_{\omega\rho} - 1) \qquad (\omega, \rho) \in \Omega \times P \tag{21}$$

$$s_{\omega\rho} + (\alpha_\omega - 1) y_{\omega\rho} \leq l + l x_\rho^{pm} \qquad (\omega, \rho) \in \Omega \times P \tag{22}$$

$$s_{\omega_1\rho} - s_{\omega_2\rho} \geq \alpha_{\omega_2} - (2l+1)(q_{\omega_1\omega_2} + z_{\omega_1\omega_2}) \qquad (\omega_1, \omega_2, \rho) \in \Omega^2 \times P,\ \omega_1 \neq \omega_2 \tag{23}$$

$$s_{\omega_1\rho_1} - s_{\omega_2\rho_2} \geq \alpha_{\omega_2} - (2l+1) q_{\omega_1\omega_2} \qquad (\omega_1, \omega_2, \rho_1, \rho_2) \in \Omega_i^2 \times P^2,\ i \in D,\ \omega_1 \neq \omega_2 \tag{24}$$

$$q_{\omega_1\omega_2} + q_{\omega_2\omega_1} = 1 \qquad (\omega_1, \omega_2) \in \Omega^2,\ \omega_1 \neq \omega_2 \tag{25}$$

$$y_{\omega_1\rho} + y_{\omega_2\rho} + z_{\omega_1\omega_2} \leq 2 \qquad (\omega_1, \omega_2, \rho) \in \Omega^2 \times P,\ \omega_1 \neq \omega_2 \tag{26}$$

The objective (16) is to minimize the number of rooms opened for surgery. The constraints (17) ensure that the cases assigned to a given room do not exceed the time available in that room, depending on whether it is open for the morning, afternoon, or both. Constraints (18) indicate that all cases must be scheduled exactly once, i.e., to exactly one room. Constraints (19)-(20) force the start time of a case in a particular room to be zero when that case is not assigned to that room, and force the start time to be at least one but no more than $2l$ when the case is assigned to that room.

A constraint from the set (21) is nontrivial only when case ω is assigned to room ρ, and room ρ is open only in the afternoon, in which case it forces the start time of any case assigned to that room to the afternoon. (In all other cases these constraints say that a start time must be greater than zero or a negative number, which are redundant.) Similarly, constraints (22) are non-redundant only when case ω is assigned to room ρ, in which case they say that case ω must be completed before time l if the room is closed in the afternoon, or time $2l$ if the room is open in the afternoon.

Constraints (23) are active only when the two indexed cases ω_1 and ω_2 are assigned to the same room and when the q variables denote that ω_2 precedes ω_1, in which case they ensure that there is enough time to finish case ω_2 before ω_1 starts. Constraints (24) are similar, but only apply to cases performed by the same surgeon i, and are therefore designed to be active even when the cases are in different rooms. Constraints (25) state that either case ω_1 precedes case ω_2, or vice versa, although some feasible assignments of these variable values may be irrelevant when ω_1 and ω_2 are performed by different surgeons and are scheduled in different rooms.

Finally, constraints (26) define the function of the z variables, which relax constraints from set (23) when the corresponding cases are performed in different rooms. Constraints (26) ensure that the relevant constraint from set (23) is not relaxed when both case ω_1 and ω_2 are scheduled in the same room ρ. Additionally, this model can be easily extended to include surgeon specific bounds on start time variables if desired, or have $y_{\omega\rho}$ variables removed if a particular room ρ is deemed unsuitable for case ω.

A.4 Implementation at HDH

The simulation results were presented to Valerie Otey, OR Director at HDH and Donald Schildkamp, Director of Management Engineering for the Capital Division of HCA, inc. Based on the simulation results, HDH decided to implement a block schedule with package generation parameters of MS_{high} =$0.2TB_i$, MS_{med} =$0.5TB_i$. That is, to balance the needs of the hospital with those of the surgeons (by bounding the allotment of shared time for some surgeons,) HDH settled upon a schedule with around 80% utilization for the hospital, and drastically better scheduling consistency for surgeons. (Refer to §4.4 for these results.)

The implementation effort began in January, 2009 and started with an initial strategy meeting that included the OR director, OR manager, the Chief Operation Officer of HDH, OR nurses, schedulers, and the Chief of Anesthesia. The first two months were spent on training and education, so that everyone was fully aware of how and why the block time and scheduling were being changed.

Rather than implement a completely new block schedule, the revised schedule was rolled out in three phases over a three month period, with the first phase beginning on March 1, 2009 and the other phases in April and May. The phases, and surgeons that would be impacted, were determined based on surgical service lines, e.g., Phase 1 included all vascular, orthopedic, and ENT surgeons. HDH decided on this strategy because implementing the change based on service lines enabled them to train segments of the OR staff and surgeons over a period of time, rather than attempting to train everyone at once. At the same time, by training all staff in a given service line they were able to make sure that everyone impacted by the process change (surgeons, surgeon office staff, hospital schedulers, and OR staff) was fully aware of the changes that were taking place and the impact of these changes on their work flows.

During January and February the OR Director held individual meetings

with each surgeon to explain the new block schedule, how it would impact the surgeon, and the assumptions that were made when the block schedule was generated. The largest was that surgeons were generally flexible with respect to when they were available, provided they were given enough lead time to plan accordingly. In some cases, this assumption did not hold. Changes to the block schedule due to surgeon availability were easily corrected. For example, the block schedule might have given a surgeon four hours of primary time on Tuesday morning, but caused a conflict because the surgeon had ongoing clinic obligations that day and could not change them. In those cases, the OR Director would arrange for a "block swap" with another surgeon. Only "swaps" of equal amounts of primary or secondary time were considered, and thus these swaps had no effect on the objective value or constraints of the stage 1 optimization. However, if a situation had arisen in which no swap was possible, it would have been easy to run Stage 1 again with the additional availability constraints. Overall, the surgeons at HDH were very supportive of the effort to improve OR efficiency and access and, when their individual utilization data was provided to them, they felt their revised block time allocations were appropriate. As expected, high volume surgeons that needed, and received, more OR time were highly enthusiastic.

In order to ensure that block times were used effectively, the OR director held several meetings with the schedulers at HDH and with the surgeons' office staffs. In this case, it was necessary to change the scheduling habits of the physician office staff. In the past, as cases were planned, the office would hold that information until the block release date and then provide all case information to the hospital five days before the day of surgery. This relatively short lead time made it hard to perform all necessary pre-operative tests and screenings which, if not completed in advance, caused cases to start later than scheduled. Previously these late starts were an occasional frustration, but there was so much unused time during the day that the resulting late finish did not delay another surgeon or, if a delay occurred, it did not cascade through the rest of the OR schedule. However, under the new block schedule the cases are scheduled in a contiguous manner which makes starting on time very important. The new practice is for the physician offices to call or fax information to the schedulers at HDH as soon as the surgeon makes the decision to perform surgery so that pre-op testing can begin as soon as possible.

Another change dealt with the scheduled duration of each case. In the past the surgeons would typically tell the scheduler how much time each case would take. For example, if a surgeon had four hours of block time, she might schedule three cases and tell the scheduler that the first case would be two hours long and the other two cases would each be one hour long. HDH routinely tracks the duration of the surgical procedure (the so-called cut time), the total time the patient spends in the OR suite (which includes the induction of anesthesia, closing the surgical site, and recovery from anesthesia), and the clean up time needed before the next patient can enter the OR suite. We found that surgeons were not very good at estimating how much OR time is needed to complete a given case. While some surgeons were good at estimating the duration of the

surgical procedure, most were not good at estimating the total time needed. As a result, surgeons would sometimes schedule cases that, based on historical durations, would take more than five hours of OR time into a four hour block. As with late starts, this can create delays for other surgeons that will impact the new OR schedule more than in the past. The revised plan was to use the historical average duration of a given procedure for each surgeon. This capability had already been built into the scheduling module of the electronic medical record system used by HDH a few years prior to this study. By utilizing the software, the schedulers were able to use much more accurate estimates that prevented one surgeon from running late and delaying another surgeon.

As of late April, 2010, one year after implementation, the management team at HDH has been very satisfied with the impact of the new OR planning approach. HDH has realized impressive gains in two key areas. The most significant has been an increase in the number of cases that are being performed each day with the resulting increase in revenue. HDH has already been able to attract four new surgeons, bringing a total of 40 additional cases a month. These are surgeons that were not previously affiliated with HDH and performed surgery at a different hospital. The per-case contribution margin for the additional cases averages approximately $2500. The result is that HDH has already realized a yearly increase of $1,200,000 in operating income and is aggressively recruiting additional surgeons.

The second major benefit is that while HDH is performing more cases, it is staffing fewer rooms than it did using the prior approach. The presence of surplus staff has enabled HDH to change its plans for staffing the ambulatory surgical center due to open during the summer of 2010. Initially they thought they would have to hire and train additional OR staff. However, with the decreased need for rooms/staff in the main hospital OR, they have decided to use existing OR staff in the ambulatory surgical center. The ability to utilize existing staff reduces the anticipated cost of the ambulatory center and enables it to become fully operational more quickly because the staff will be experienced, trained, and familiar with the surgeons and the procedures they perform. A hospital in a different situation could have reduced costs by more than $1,000,000 per year by reducing OR staffing and eliminating the support cost associated with OR suites that are no longer needed to service existing volume.

Moving forward, HDH plans to track block utilization every month. If a surgeon has low utilization for two months, she will receive a letter making her aware that she is not utilizing her block time effectively. After three months of low utilization, the surgeon will receive a final letter that explains her block time will be changed if utilization is low during the fourth month. In this manner, HDH hopes to manage expectations so that any change in block time does not come as a surprise to the surgeon.