The INFORMS Journal on the Practice of Operations Research

Interfaces

**Verizon Optimizes Work Center Locations to Reduce Installation and Repair Operations Costs**
J. David Allen, Roger L. Tobin, Anthony Calderan

**The Development of a Concurrent Spare-Parts Optimization Model for Weapon Systems in the South Korean Military Forces**
Seongmin Moon, Ui Jun Kim

**Sustaining the Drone Enterprise: How Manpower Analysis Engendered Policy Reform in the United States Air Force**
Kiel M. Martin, Daniel J. Richmond, John G. Swisher

**Snider Tire Optimizes Its Customers-Stores-Plants Transportation Network**
Sanjay L. Ahire, John B. Jensen

**Optimizing Student Team and Job Assignments for the Holy Family Academy**
Sharan Srinivas, Mohammadmahdi Alizadeh, Nathaniel D. Bastian

**A Review of Scheduling Problems and Research Opportunities in Motion Picture Exhibition**
Katherine Goff Inglis, Saeed Zolfaghari

**Book Reviews**
Wenjing Shen, ed.

Volume 47 • Number 2 • March–April 2017

Interfaces

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# Kaiser Permanente Oakland Medical Center Optimizes Operating Room Block Schedule for New Hospital

Brittney Benchoff, Candace Arai Yano, Alexandra Newman

informs.
http://pubsonline.informs.org/journal/inte/

INTERFACES
Vol. 47, No. 3, May–June 2017, pp. 214–229
ISSN 0092-2102 (print), ISSN 1526-551X (online)

# Kaiser Permanente Oakland Medical Center Optimizes Operating Room Block Schedule for New Hospital

**Brittney Benchoff,**[a] **Candace Arai Yano,**[b] **Alexandra Newman**[c]

[a] Alpine Data, San Francisco, California 94107; [b] Department of Industrial Engineering and Operations Research and the Haas School of Business, University of California, Berkeley, Berkeley, California 94720; [c] Doctoral Program in Operations Research with Engineering, Colorado School of Mines, Golden, Colorado 80401

**Contact:** bkbenchoff@gmail.com (BB); yano@ieor.berkeley.edu (CAY); newman@mines.edu (AN)

**Received:** November 6, 2015
**Revised:** May 14, 2016; October 1, 2016
**Accepted:** October 18, 2016
**Published Online in Articles in Advance:** April 12, 2017

https://doi.org/10.1287/inte.2017.0885

**Abstract.** In July 2014, Kaiser Permanente, a major integrated healthcare delivery consortium, opened a new hospital, replacing an existing hospital, adjacent to its headquarters in Oakland, California. Hospital staff needed to devise a new operating room schedule. In developing the schedule, the key decisions the staff would have to make were the type of block (i.e., a combination of surgery types that can be performed in the same operating room on the same day) to assign to each operating room on each day of the planning horizon. We report on the development and implementation of an integer programming model to generate a near-optimal block schedule. The approach differs from many in the literature because it considers both direct nursing costs and patient-related costs, and can accommodate a variety of practical constraints.

Kaiser Permanente Oakland implemented the proposed schedule and continues to use it with minor modifications in response to subsequent growth and changes in patient demand patterns. Three major benefits of the schedule are that it: (1) satisfies almost all of the monthly block requirements in only four weeks, thereby releasing capacity to reduce the surgical backlog; (2) eliminates days with excess admissions, which would have required additional nursing staff; and (3) reduces the number of surgeries canceled due to an insufficient number of available beds.

**History:** This paper was refereed.

**Keywords:** operating room scheduling • block scheduling • healthcare • integer programming

Kaiser Permanente (KP) is an integrated, managed-care consortium that comprises 38 medical centers, 618 outpatient facilities, over 17,000 physicians, 47,000 nurses, and 175,000 other employees, and provides services to 9.5 million healthcare members nationwide. One of its major medical centers is located in Oakland, California, where the firm is based. In July 2014, KP closed its existing hospital in Oakland and opened a new hospital with 15 operating rooms (ORs) for in-patient (nonambulatory) surgical procedures; of these ORs, 12 are fully equipped and three are available for future expansion.

In anticipation of the changeover, Dr. Thomas Barber, the Associate Physician-in-Chief, asked a team from University of California, Berkeley to develop a block schedule for the 12 ORs planned for the new hospital and a schedule that KP could use for the last six months at the existing hospital. As is common in the medical community and in the research literature, KP's management defines a block as a set of procedure types to be performed on the same day in a single OR. As examples, in the orthopedic specialty, one block type might consist of four simple bone-related or joint-related procedures (e.g., knee replacements), and another block type might consist of one so-called revision (i.e., repair or replacement of a prior joint replacement), which tends to be time consuming, plus one short joint-related procedure.

A block schedule specifies which block type is assigned to each OR on each day over a time horizon, such as a month, after which the schedule repeats. As a result of medical advancements, most surgeries are performed on an outpatient basis. These surgeries consume OR capacity that is shared with the inpatient surgical blocks; however, the outpatients are not expected to stay overnight in the hospital. In general, because bed capacity in hospital wards is often a constraining factor, from the standpoint of developing the block schedule, inpatient surgical blocks are planned first and outpatient blocks are scheduled around them.

After completing surgery, inpatients stay in the postanesthesia care unit (PACU) until the anesthesia has worn off and they are ready to move to a regular hospital ward. The vast majority of surgeries are scheduled in advance with durations that are relatively predictable (due to available historical data or the surgeon's own estimate of the duration). KP staffs the PACU accordingly; therefore, we did not have to consider PACU nurses when constructing the block schedule. The postsurgical patients, each of whom has an uncertain length of stay (LOS), utilize scarce bed capacity and require nursing resources, whose levels need to be aligned with the bed occupancy (i.e., number of occupied beds) in each ward.

Surgical and nonsurgical patients are intermixed in the hospital wards. Surgical patients use a substantial portion of the bed capacity in the orthopedic ward; conversely, patients who are in the hospital for reasons other than an elective surgery use more than half of the bed capacity in the nonorthopedic wards. The bed occupancy of nonsurgical patients is stable, fluctuating only about five percent over the course of the week, whereas the bed occupancy of surgical patients routinely fluctuates about plus or minus 20 percent from the (weekday) mean during the week. The bed occupancy on weekends is only a fraction of the weekday occupancy.

At the commencement of our project, Dr. Barber's immediate concern was that, as a result of the then-existing OR schedule, orthopedic-ward bed occupancy for postsurgical patients sometimes exceeded the ward's capacity. Although the excess patients could usually be accommodated in other wards, the nurses in the medically preferred ward have a better understanding of the needs of those patients and can therefore provide better care. Dr. Barber also anticipated difficulty in constructing a good, feasible schedule because 20 percent fewer beds would be allocated to orthopedic patients in the new hospital. In addition, the existing OR schedule caused a clustering of postsurgical hospital admissions, often more than the usual nursing team could handle, on specific days of the week.

Studies indicate that greater variability in patient loads leads to higher patient mortality (Aiken et al. 2002), that higher patient loads lead to greater risk of infection (Cimiotti et al. 2012), and that spikes in the number of patients admitted on the same day lead to a greater likelihood of readmission (Baker et al. 2009). Thus, achieving more stable bed occupancy levels that do not overload the nursing staff, and reducing the peak number of postsurgery admissions on any given day, can contribute to better medical outcomes for the patients and improve operational efficiency.

A team of doctors and nurses, known as the smoothing team because of its goal of smoothing bed occupancy in hospital wards, had been working on developing an improved schedule for several months using a trial-and-error approach; however, when we began the project, team members had not yet found a schedule that was satisfactory for the orthopedic ward in the existing hospital and had not made much progress on developing a schedule for the new hospital. To explore different scheduling options, they used a spreadsheet, seeking to minimize the mean absolute deviation (MAD) of the actual bed occupancy from the average occupancy. If $A_t$ is the actual bed occupancy in period $t$ and $\bar{A}$ is the average bed occupancy over a horizon of $T$ periods, then MAD is defined as

$$\mathrm{MAD} = \frac{\sum_t |A_t - \bar{A}|}{T}. \tag{1}$$

Although MAD is a useful metric if smoothing bed occupancy is inherently valuable, our team realized that minimizing MAD would not necessarily minimize the peak bed occupancy, which was clearly a consideration; therefore, we decided to delve further into the problem to better understand it.

We developed a block schedule that considers both nursing and patient-related costs, and a plethora of other practical constraints. The final schedule enables the orthopedics ward to stay within its reduced bed capacity and also allows KP to fit almost all of the monthly block requirements into a four-week time window, thereby freeing about six percent of the OR capacity for future growth or reducing the surgical backlog. We provide details in the *Implementation and Benefits* section.

**Table 1.** The Minimum Number of Blocks Required in a Month Differs Across the 15 Block Types. High = 13–30; Medium = 6–12; Low = 1–5

| Block type | Min. no. of blocks | Block type | Min. no. of blocks |
|---|---|---|---|
| Orthopedics (revision*) | Medium | Adult surgery (cancer) | Medium |
| Orthopedics (nonrevision) | High | Adult surgery (noncancer) | Medium |
| Spine (short surgeries) | Medium | Plastic surgery | Low |
| Spine (incl. one long surgery) | High | Pediatric spinal surgery | Low |
| Gynecology (cancer) | Low | Pediatric general surgery | Low |
| Gynecology (noncancer) | Low | Pediatric neurosurgery | Medium |
| Podiatry (inpatient) | Low | Pediatric surgery (other) | Medium |
| Urology | High | | |

*Note.* *Subsequent surgery for repair or replacement following a joint replacement or similar procedure.

## Kaiser Permanente's Block-Scheduling Problem

After several meetings with the smoothing team, we had a clearer understanding of KP's goals, both tangible and intangible, and information on the block-scheduling constraints, as we describe next.

KP utilizes 15 inpatient block types, which it differentiates based on surgical specialty or subspecialty, surgical procedure types, and sometimes the number of each type of surgical procedure, within a block. Table 1 lists the block types and the range corresponding to the minimum required number of blocks per month for each type. (For confidentiality reasons, we cannot disclose more detailed data.)

The three adult wards associated with the block schedule are orthopedics (ward A) with 24 beds and adult ward B (a combination of two 24-bed wards). The pediatric wards are smaller. The hospital also has intensive care units and wards that provide an intermediate level of care (i.e., more care than in general wards but less than in the intensive care unit); however, the utilization in these wards is purposely kept at a level that allows for unanticipated emergencies. Therefore, we did not need to consider them. The hospital also has other specialty wards for which the vast majority of the bed occupancy is *not* the result of elective surgery; therefore, we did not need to include them in our study.

### Elements of the Objective Function

Because the orthopedics ward typically had high bed utilization, minimizing MAD would have eliminated most of the unwanted peaks. However, as we inquired further, we discovered out-of-pocket costs that should legitimately be included in the objective function. As our study proceeded, we found that incorporating the peak expected bed occupancy in each ward with a sufficiently high weight in the objective function reliably reduced the peaks. We could achieve this while avoiding distortions that would have resulted from incorporating deviations below the mean (as were implicit in MAD), which had no adverse effects on costs or medical outcomes. For this reason, we only utilized MAD as a tiebreaker. Next, we list the considerations in our objective function and the rationale for each.

(1) Costs of nurses above a core (baseline) level in each ward on each day: KP prefers to maintain a core (constant minimum) staff of nurses in each ward; it maintains one level for weekdays and another level for weekends. It also schedules additional nurses as the bed occupancy necessitates; however, these nurses must be notified in advance (i.e., three days in advance when we started the project, but only one day as of this writing), and KP must pay them even if it ultimately does not use their services.

(2) Costs of nurses to handle excess admissions above a threshold in each ward on each day: Surgical patients who must stay overnight or longer are admitted to the hospital. The core nursing staff in each ward can handle the processing of a specified maximum number of new admissions each day. If admissions to the ward are expected to exceed that number, an additional nurse must be scheduled.

(3) Expected patient days in excess of the effective bed capacity set aside for surgical patients, assuming that each patient is assigned to the medically preferred ward: Excess patients can be assigned either to beds set aside for emergency patients in the same ward but not utilized, or to beds in other wards, an option that KP prefers to avoid.

(4) Reduction of backlogged blocks for each surgical specialty by an amount within a specified range: Because of small-to-modest fluctuations in the need for surgery among its members (health plan customers), KP's Oakland hospital has surgical backlogs that vary by specialty. Management wishes to reduce these backlogs by scheduling additional surgical blocks (within a specified range) above the steady state surgical demands reflected in the minimum number of blocks per month of each type. For some types of procedures (e.g., cancer surgery), it is desirable to keep the backlog close to zero. For other types of procedures, a small-to-modest backlog is desirable because the patient's condition may improve without surgical intervention while he or she is in the queue, or patients may need time to plan their personal and work affairs prior to undergoing major surgery (e.g., spinal surgery). We note that KP has three other hospitals within 15 miles of its Oakland hospital; therefore, it can manage surgical queues across a fairly large pool of ORs, if necessary.

(5) Peak expected bed occupancy in each hospital ward: When the wards approach 100 percent utilization of the beds, it is sometimes necessary to delay or cancel surgeries because of concerns about not having a suitable place for patients to recover after surgery. Thus, holding all else equal, minimizing the peak bed occupancy in each ward could contribute to reducing surgery cancelations and also leave slack for unanticipated events.

(6) MAD of bed occupancy in each hospital ward: As mentioned earlier, we use this smoothness metric as a tiebreaker.

In the remainder of the paper, we refer to the following combination of factors as patient-related costs: (a) medical consequences that patients incur when bed occupancy in the preferred ward exceeds the effective bed capacity (an aggregate measure that is reflected in Item 3); (b) inconvenience and potential medical costs borne by patients when the peak bed occupancy is high (an aggregate measure that is reflected in Item 5); and (c) the cost of waiting and a resulting worsening of the patient's medical condition because of a longer backlog, which we represent as the negative cost of a backlog reduction (reflected in Item 4).

We note that the trade-offs among the various costs are complex because of the highly interactive effects of the decisions on the various terms in the objective function. For example, switching the assignment of merely two blocks from different specialty areas may cause each of the terms in the objective function to increase, decrease, or remain the same, and the effects of an assignment change on the components of the objective function are not necessarily correlated as one might expect. The constraints further complicate this challenging problem. Next, we present specific examples.

KP staff articulated many requirements (i.e., hard constraints) and other considerations, which could be expressed either as hard constraints or as factors to be included in the objective function. These include resource constraints, demand satisfaction and backlog reduction, and smoothing and spacing considerations, as we describe next.

## Resource Constraints

- Constraints to ensure that at most one block is assigned to each eligible OR in eligible periods: ORs are rarely utilized on weekends except for emergency procedures, and some are set aside for specific purposes. For example, one OR is reserved for the entire day and another from early afternoon onward for emergency procedures. In addition, some ORs are dedicated to urology and podiatry outpatient procedures. (We do not need to consider the scheduling of rooms for outpatient procedures in our solution methodology.)
- Constraints due to surgeon availability: The surgeons based at KP's Oakland hospital also have regularly scheduled blocks at a smaller KP hospital located about 15 miles away in Richmond, California. These commitments must be considered in developing the Oakland hospital's schedule.
- Bounds on the maximum number of blocks of each type (or sum over a set of types) on a given day, or, in some cases, over a longer horizon such as a week: These constraints stem from the availability of surgeons qualified to perform the procedures within the specified block, or from the availability of ORs that are suitably equipped for the relevant procedures (e.g., spinal surgery).
- Constraints to ensure adequate nurse staffing according to specific ratios: California law allows a maximum patient-to-nurse ratio of 5:1 for regular adult wards; smaller ratios apply to adult intensive care and pediatric regular and intensive care wards. (We note

that KP was one of the first healthcare organizations to implement nursing ratio rules prior to the adoption of the related state legislation.)

- Constraints to ensure appropriate scheduling of teaching days for surgeons in each specialty: KP's Oakland hospital is a teaching hospital, which trains medical interns and residents. This involves periods of up to a full day of classes once a week, during which specific surgeons, or possibly all surgeons within a specialty, participate. For each specialty, there may be constraints specifying the allowable days, and the specific day can be selected when optimizing the block schedule.

### Demand Satisfaction and Backlog Reduction

- Minimum and maximum number of blocks of each type to be assigned during the scheduling horizon, where the minimum is the steady state requirement and the maximum is the steady state requirement plus the maximum allowable backlog reduction. There may be similar constraints on the minimum and maximum number of blocks for shorter time intervals, such as one week or two consecutive weeks.

### Smoothing and Spacing Considerations

- Limits on the number or proportion of a (sub)specialty's total block assignments over the horizon that is scheduled in any week, or the proportion of a (sub)specialty's total block assignments over a week that is scheduled on any day (and similar constraints for other time intervals).
- Limits on the change in the total number of assignments for a block type from one week to the next.

We have not recounted every constraint above; however, we provide more details in the appendix. In the next section, we briefly summarize related literature and explain how KP's problem differs.

## Literature Review

Planning and scheduling decisions for hospital ORs are usually made in a hierarchical fashion. Various researchers have partitioned the decision space in different ways; however, conceptually, the highest level of the hierarchy involves making aggregate capacity decisions, taking into account long-term considerations, such as profit or cost. The middle level involves allocating capacity at a finer level of detail, such as medical specialties or classes of patients; at this level, most models in the literature consider constraints pertaining to ORs, surgeons, and the resources required for preoperative and postoperative care, usually with a horizon that ranges from one week up to a few months. The resulting schedules are assumed to be cyclic (i.e., to repeat after the cycle has elapsed). The lowest level involves decisions regarding individual patients and (or) surgical procedures over a horizon of one day to one week. Santibáñez et al. (2007) and Testi et al. (2007), among others, provide overviews of three-level hierarchical planning for ORs. Each of these papers also reports on a three-level methodology developed for implementation in a large public hospital. The literature on various aspects of planning and scheduling hospital ORs is vast. We refer the reader to Cardoen et al. (2010), May et al. (2011), Vanberkel et al. (2009), Blake (2010), and Guerriero and Guido (2011) for surveys.

In this literature review, we focus on articles in which the primary decisions are the assignment of surgical specialties to OR time blocks (versus the assignment of a specific patient or patient type to a day or time slot in the schedule), and the primary objective is smoothing downstream resource utilization or matching a target utilization of the beds. Within this class of problems, virtually all models that represent the patient's LOS in a probabilistic fashion utilize expectations of the number of patients occupying beds; that is, if a patient will remain in the hospital until a given day with probability $p$, this generates a deterministic demand for a fraction $p$ of a bed, making it possible to formulate most versions of the block-scheduling problem as a deterministic mixed-integer program.

Several articles are closely related to our work. Belien and Demeulemeester (2007) develop an optimization model whose aim is to make block assignments to smooth the bed occupancy level throughout the planning horizon. They consider different ways in which to achieve smoothness, including minimizing variance and minimizing a metric corresponding to a specified percentile of the staffing or resource requirements. Belien et al. (2009) generalize the previous model to consider multiple wards, utilizing an objective that includes weighted values of the peaks and variances of the mean bed occupancy levels over the scheduling cycle. Santibáñez et al. (2007) present a block-scheduling model designed for a multihospital environment (although bed capacity at each hospital is analogous to bed capacity within a ward). They suggest several objectives; five pertain to weighted or unweighted throughput and one relates to minimizing peak usage of various hospital resources. Price et al. (2011) address this problem using the objective of reducing the sum (over days in the horizon) of the excess of admissions over discharges in the intensive care unit to reduce congestion in the (upstream) postanesthesia care unit.

Some researchers examine generalizations of the basic block-scheduling problem. For example, Belien et al. (2009) allow an OR to be divided between two surgical specialties on a given day, and Belien and Demeulemeester (2008) include nurse-scheduling decisions in their formulation. Researchers have specifically considered the broader consequences of block-scheduling decisions. For example, van Oostrum et al. (2008) consider a model in which blocks are constructed, taking into account the impact of uncertainty of procedure durations on overtime, and both maximization of OR utilization and leveling of bed occupancy are considered in the block-scheduling problem.

A few papers describe models for cyclic block scheduling with objectives that differ from ours. Minimizing deviations or shortfalls from target allocations of OR hours for various surgical groups or patient categories (e.g., Blake and Donald 2002 and Santibáñez et al. 2007) is among the more common objectives, and is often considered in highly constrained settings. This objective helps to limit waiting times for surgical procedures to be scheduled, and to achieve greater fairness in these waiting times among patient categories.

An emerging research stream considers uncertainty (e.g., in demand or LOS) in the context of block scheduling in which bed occupancy is a major consideration. Vanberkel et al. (2011) develop a method to calculate the distribution of bed occupancy for each day in the planning horizon for a given master surgery schedule. For examples of papers that address block scheduling under uncertainty, see Holte and Mannino (2013) and van Essen and Bosch (2013).

KP's problem is similar to others in the literature; however, it has many special features, including:

(1) Direct costs and patient-related costs as a result of either assigning a patient to a ward that is not ideal, or delaying a surgery due to bed unavailability;

(2) A threshold number of admissions in a ward above which an additional nurse is required;

(3) The concept of a core nursing staff;

(4) Discretionary scheduling of blocks above the steady state demand (to reduce the backlog);

(5) Scheduling of teaching (and, thus, nonsurgery) days for surgeons; and

(6) Numerous realistic constraints that are more complex than those described in the vast majority of the literature.

## Data Collection, Preparation, and Analysis

The staff and management at KP provided information on essential inputs for our model, including (1) the preferred hospital ward to which patients from each block type should be admitted; (2) the target number of blocks of each type to be scheduled during the horizon, the minimum and maximum for each week and (or) day, and analogous values for aggregations of block types associated with a subspecialty, if applicable; (3) the number of beds available in each hospital ward for patients undergoing elective surgery; (4) feasible OR assignments for each block type; (5) surgeon availability; (6) maximum patient-to-nurse ratios for the various wards; (7) current backlog of each block type; and (8) nursing costs.

KP also provided LOS data for the many thousands of procedures (on specific patients) that had been performed over an 18-month period, with each classified into one of a few thousand different types, as defined by KP's own classification of surgical procedures. For example, a total knee replacement and a laparoscopic appendectomy are different procedure types. As with most exercises involving raw data, our first challenge was to clean the data set to remove records with insufficient information and those with spurious entries. Our next challenge was to divide the procedure types into groups to determine a LOS distribution for each group. Groupings were tied to the definitions of the blocks. For example, one orthopedic block type consists of one long surgery (e.g., a hip replacement) and a short surgery (e.g., a knee replacement). Therefore, we needed to create a group of procedures that corresponds to the possible long surgeries and another that corresponds to possible short surgeries for that type of block. Dr. Barber provided guidelines for the initial grouping. For each group, we calculated the empirical probability mass functions representing the number of days (including the day of admission) that a patient stays in the hospital for a procedure within that group. We then checked that the LOS distributions (e.g., mean, 25th, 50th, 75th, and 90th percentiles) for procedure types within the same group were roughly similar to each other. This ensured that a weighted average LOS distribution (considering all procedure types in the group), which we use in our solution procedure, would be a reasonable representation of any procedure type within the group. To convey the diversity of LOS distributions, we present histograms for two groups of procedures with low and moderate lengths of stay in Figure 1.

Given the (group) LOS distributions, we calculated the expected bed occupancy that each block type would generate for the day of surgery and for each day until the maximum pertinent LOS, assuming that the LOS associated with the various procedures in a block are statistically independent; no evidence suggested that we should assume otherwise. As an example, if a block consists of two procedures from Group 1 and one procedure from Group 2 with the LOS distributions shown in Figure 1, then the expected bed occupancy vector would be as shown in Figure 2. In this example, all three surgery patients are occupying beds on Day 1; however, in expectation, 0.8 of a patient (40 percent of each of the two patients who received a procedure from Group 1) would have departed by Day 2, and by Day 3, only the patient who received a procedure from Group 2 would remain and only if his (her) LOS is three days or more, which occurs with a 75 percent probability. Due to the extensive statistical pooling resulting from having adult wards with 24 beds, we deemed that using expected bed occupancy vectors would be adequate. Our approach does, however, account for the right tails, some of which can be long for complicated procedures that require lengthy in-hospital recuperation. More generally, however, advances in medical technology have enabled less invasive surgical procedures for which LOS distributions have low variances (e.g., a range of plus or minus one day).

**Figure 1.** The LOS Distribution for Simple Procedures (Group 1) Differs from the Distribution for Moderately Complex Procedures (Group 2). (Probabilities Are Shown to the Nearest Five Percent)

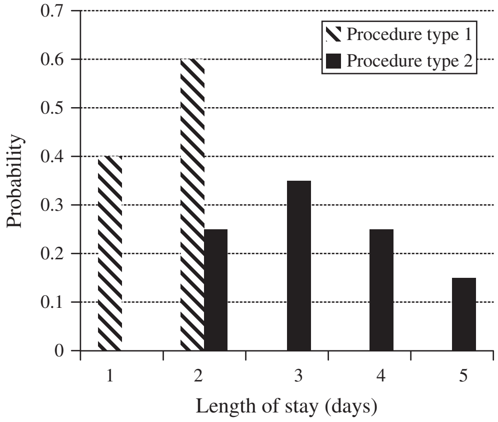
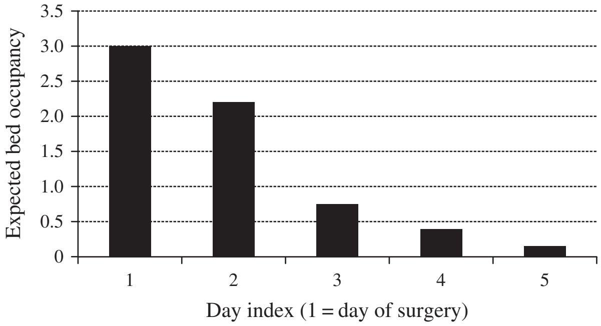


**Figure 2.** A Block Consisting of Two Procedures from Group 1 and One Procedure From Group 2 Generates the Pattern of Expected Bed Occupancy Levels Shown in the Graph, Where the LOS Distributions for the Procedures from Groups 1 and 2 Are Shown in Figure 1



## Solving the Scheduling Problems

As with most practical scheduling problems, the solution process was evolutionary. KP was using a monthly schedule and specified the number of blocks of each type that were required each month. It had no systematic procedure for making adjustments to account for the facts that months have different numbers of days and can start on different days of the week. We began with a formulation that would accommodate between 28 and 31 days in a month, but found it burdensome to address constraints that applied at the boundaries between adjacent months. We eventually discovered that scheduling virtually all of the required blocks for a month within a 28-day window was possible. With Dr. Barber's support, we convinced the smoothing team that we could move to a repeating four-week schedule.

Only the orthopedic bed capacity prevented the schedule from accommodating all of the baseline monthly block requirements. Orthopedic surgeons preferred to schedule patients who were expected to have a several-day LOS for surgery early in the week, creating peak bed requirements on Tuesday and Wednesday, which sometimes exceeded the ward capacity. These patients would often depart on Friday or Saturday, leaving the ward underutilized on weekends. Recognizing that the orthopedic ward would be smaller in the new hospital, we needed to assess the effects of various types of changes to identify a viable schedule. We used our methodology to create many schedules by adjusting the various constraints, and finally determined that it was essentially impossible to limit orthopedic surgeries to the weekdays without creating a serious overcapacity problem in the orthopedic ward in the new hospital. As a result of this analysis, Dr. Barber was able to convince two orthopedic surgeons to each work one Saturday per month, in lieu of a weekday, on a trial basis. This change would be enough to eliminate the overcapacity problem in the orthopedic ward in the new hospital.

Our final formulation is for a single 28-day horizon. We needed to adjust this basic 28-day formulation to consider that patients undergoing surgery near the end of the 28-day scheduling horizon might still be in the hospital at the beginning of the next scheduling horizon. We considered the option of a formulation with constraints to force the beginning and ending conditions—expressed as the expected bed occupancy in each of the five wards in our model from Day 1 through the day corresponding to the maximum LOS for any surgery type (10 days in our data)—for each scheduling horizon to be within some tolerance. Because all the primary decision variables in our problem are binary or integral, we found it difficult to adjust the tolerance so that the solution to the optimization problem would identify a schedule that minimized the objective function *and* resulted in nearly equivalent initial and terminal conditions. This was primarily due to including MAD in the objective function, which significantly increased computation times; these increases occurred because each new feasible schedule generated in the course of solving the optimization problem resulted in a different value of MAD, but not necessarily a change in the total nursing cost or peak bed occupancy levels. For this reason, we decided instead to repeatedly solve the problem, initially assuming no carryover patients at the beginning of the horizon, then utilizing resulting bed occupancy levels due to carryover patients as input to the next iteration. We repeated this process until the initial and terminal conditions were sufficiently close; three or four iterations were adequate for this purpose. This was computationally more efficient and allowed us to retain MAD in the objective function.

We also learned about new constraints whenever we generated a proposed schedule that violated an unmentioned constraint. One example is the disallowance of pediatric surgery on Mondays. If a patient became sick over the weekend (a situation that is more likely to occur with children than with adults because children who need surgery are generally in poorer health), that patient's surgery would have to be canceled; however, KP would have insufficient time to schedule another surgery in that time slot. We also learned about the importance of scheduling pediatric general-surgery blocks every few days so that relatively critical cases would not have to wait for more than that period of time. Numerous such issues arose and we gradually modified our formulation to account for them. Often, we had to decide whether to impose hard constraints or to modify penalties in the objective function to achieve the desired properties of the schedule, and we generally made these decisions pragmatically based on what was likely to be more effective.

We formulated the problem using AMPL and solved several variants of the problem (see the *Impact of Core Nurse Staffing Levels and Weights in the Objective Function* section) using CPLEX 12.6.0.1 on a Dell PowerEdge R410 with 16 processors (running at 2.72 GHz each) and 12 GB RAM. CPU times for our problem variants range from substantially less than one minute up to over one hour to achieve optimality gaps of 1.5 percent or less. The key differences in CPU time were due to the degree of flexibility in the range of values for the decision variables. In particular, when the core (minimum) nurse staffing levels are relatively high (e.g., one less than the maximum that could be required in that ward), near-optimal solutions are identified quickly because the practical range of decisions regarding staffing levels is small or zero. In addition to using a relatively efficient formulation of the problem, we adjusted CPLEX default parameter settings to reduce CPU times. We provide details in the appendix. Typically, block-scheduling decisions are made, at most, every few years; therefore, CPU times of a few hours or less for each problem variant are acceptable and would make it possible to solve several problem variants overnight. To utilize the optimization software in what-if mode, one could set optimality gaps of two to three percent to solve problem instances quickly (i.e., in a few minutes) in the exploratory phase, and then set smaller optimality gaps for a few problem instances when generating a final set of options from which to choose.

Recall that KP's initial objective was minimizing the mean absolute deviation of the expected number of patients. As we attempted to find good solutions, we discovered that the inclusion of MAD tended to lengthen solution times; however, it could also serve as a tiebreaker if the weight on MAD in the objective function is set to a very small value. As we explained earlier, reducing the peak value of the expected number of patients in each ward is useful because the resulting bed occupancy levels are more stable. This, in turn, reduces the need to cancel surgeries and makes the solution more robust to unanticipated fluctuations. MAD, however, accounts for shortfalls of the bed occupancy from the mean as well as deviations above the mean, but only the latter have the potential to incur costs or create practical challenges. The inclusion of MAD penalizes some outcomes that should not be penalized and may thereby distort the solution. As such, its inclusion may not be sensible in view of the increased solution times, unless a more fundamental reason for smoothness in the bed occupancy exists.

## Impact of Core Nurse Staffing Levels and Weights in the Objective Function

Our formulation provides flexibility in choosing certain parameters such as the core (baseline) nurse staffing levels in the various wards, which had been chosen based on tradition but were changeable. In solving a variety of problems with different objective function weights and different core nurse staffing combinations, we gained insights that would be useful to another hospital that is updating its block schedule. We found that objective function coefficients, if set within realistic ranges, have predictable impacts; therefore, we instead focus on the more interesting results.

Because our model pertains to inpatient surgical blocks, we report only corresponding bed occupancies; actual bed occupancies would include nonsurgical patients. We consider core nurse staffing levels that range from slightly below the levels necessary to support the average bed occupancy due to surgical patients to levels that are similar to those that would be needed to handle both surgical and nonsurgical patients. We focus here on results pertaining to the adult orthopedic ward, which we call Ward A, and the adult medical ward, which we call Ward B. (Ward B is a combination of two physical wards.) Occupancy in the other wards is also affected by the block schedule; however, KP experiences fewer capacity-related problems in those wards. We discuss several observations and insights later.

We found that the core nurse staffing levels have an unexpected effect on the smoothness of the bed occupancy levels in the resulting solutions. When the core nurse staffing levels are set at generous values, the total nursing cost becomes a sunk cost; therefore, the other factors, the bulk of which pertain to smoothness of the bed occupancy in the wards, carry more importance. Hence, even if two different combinations of core nursing levels lead to essentially the same total nurse staffing levels in the wards, the bed occupancy is often smoother when we start with larger core nurse staffing levels. This phenomenon is evident in Figure 3, which shows bed occupancy levels in Ward A for core nurse staffing combinations of $(4, 3)$ and $(2, 2)$, where the first (second) value in parentheses denotes the core nurse staffing level in Ward A (B). (Recall that the nurse staffing level in Ward B also affects the outcomes in Ward A because surgery blocks associated with the two wards share many of the same ORs.) Notice that the weekly patterns of expected bed occupancy for the $(4, 3)$ case—the case with the higher core nurse staffing levels—are similar, and midweek bed occupancy levels do not fluctuate much from day to day. To be more specific, the expected bed occupancy levels on Mondays (Days 1, 8, 15, and 22) are between 12.7 and 13.8; midweek levels are between 15.4 and 17.4; Friday levels are between 13.3 and 15.5; Saturday levels are between 8.1 and 10.0; and Sunday levels are between 5.6 and 6.0. By contrast, much greater variations can be seen

**Figure 3.** (Color online) Applying Our Optimization Procedure to Different Combinations of Core Nurse Staffing Levels Leads to Different Patterns of Expected Bed Occupancy Levels (Shown for Ward A). The Pair (4, 3) Denotes Four Nurses in Ward A and Three Nurses in Ward B in the Core Nurse Staffing Configuration, and (2, 2) Denotes Two Nurses in Each Ward in the Core Staffing Configuration

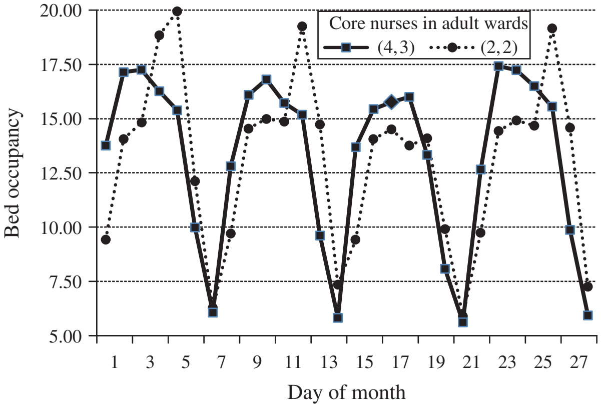
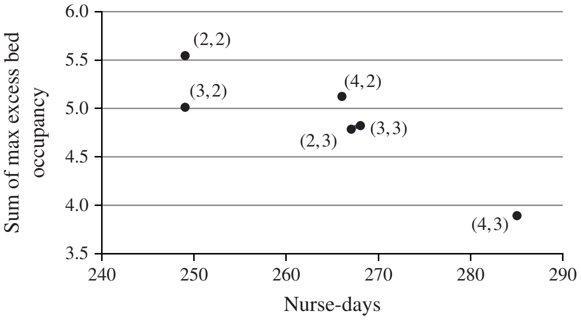


**Figure 4.** (Color online) Increasing Core Nurse Staffing Levels in the Two Adult Wards (and Thus Also Increasing the Total Number of Nurse-Days) Does Not Always Decrease the Sum of Maximum Excess Bed Occupancy Levels Across the Wards. (Marker Labels Show the Core Nurse Staffing Levels in the Two Adult Wards.) For Example, When Moving from (3, 2) to (4, 2), the Number of Nurse-Days Increases; However, the Sum of the Maximum Excess Bed Occupancy Levels Also Increases

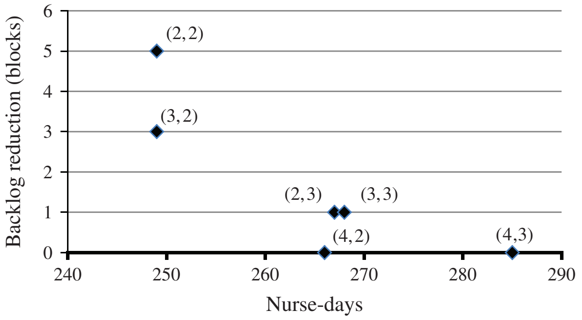


for the (2, 2) core nurse staffing combination. Although the Monday expected bed occupancy levels are stable (ranging from 9.4 to 9.8), the midweek range is 13.8 to 18.9; the range on Friday is 14.1 to 20; the range on Saturday is 9.9 to 14.7; and the range on Sunday is 5.9 to 7.4. As such, with the exception of Monday, the ranges under the (2, 2) core nurse staffing combination are approximately 2.5 times that of the (4, 3) core nurse staffing configuration. However, the former combination has lower nursing costs in Ward A because the expected bed occupancy is 15 or less (requiring only three nurses) on all but three days during the 28-day horizon, and the expected bed occupancy is roughly 19 (requiring four nurses) on the other days. On the other hand, the schedule with the (4, 3) core nurse staffing combination requires four nurses each day.

To further illustrate this point, for a representative problem instance, Figure 4 shows the values of nurse days (on the horizontal axis) and the sum of the maximum excess bed occupancy levels across the wards (on the vertical axis) for six combinations of core nurse staffing levels in the two adult wards. The maximum excess bed occupancy in each ward is the maximum, across the days in the scheduling horizon, of the number of patients in excess of the nominal number of beds allocated to surgical patients. In the figure, the markers are labeled with the corresponding pair of core nurse levels in the adult wards. Not surprisingly, using smaller values of the core nurse staffing levels generally leads to solutions with a smaller number of nurse days. As noted earlier, we found it surprising that the solution with the largest number of nurse days (which corresponds to core nurse staffing levels of four in Ward A and three in Ward B on the weekdays) leads to a solution with a *smaller* value of the maximum excess bed occupancy level, summed across the wards, which is an indicator of the smoothness of the bed occupancy levels. The relationship between the sum of the maximum excess bed occupancy levels and nursing cost is clearly not monotonic, partly because we are solving an integer program. Although we might expect that high values of the core nurse staffing levels lead to less stable bed occupancy levels, the opposite can occur for the reasons we previously explained.

We also found that the core nurse staffing levels have an unexpected effect on the number of additional blocks from the backlog that are accommodated in the schedule. Figure 5 shows that the solution with the largest number of nurse days (i.e., core nurse staffing levels of four and three in Wards A and B, respectively) on the weekdays leads to no reduction in the backlog, whereas four of the other five solutions do. High core nurse staffing levels lead to greater importance of smoothness of the bed occupancy and the resulting solutions may not leave sufficient slack to

**Figure 5.** (Color online) Increasing Core Nurse Staffing Levels in the Two Adult Wards (and Thus Also Increasing the Total Number of Nurse Days) Does Not Always Lead to a Larger Reduction in Backlogged Blocks. (Marker Labels Show the Core Nurse Staffing Levels in the Two Adult Wards.) For Example, Moving from (3, 2) to (4, 2), the Number of Nurse Days Increases; However, the Backlog Reduction Changes From Five to Zero



accommodate another block. However, when the core nurse staffing levels are low, if identifying a feasible OR schedule with only the core nursing staff is not possible, nurses must be added on a subset of the days. There may then be enough slack to allow additional blocks to be scheduled. That is, fluctuations in the nurse staffing level from day to day can facilitate accommodating additional blocks.

Although we considered all factors in the objective function in constructing the schedule that KP ultimately implemented, our formulation allows considerable flexibility in this regard, and other hospitals might choose different weights. For example, a public hospital with limited resources might consider only nursing costs. Conversely, a private hospital whose patients are wealthy or have generous insurance coverage might consider almost exclusively patient-related costs and pass on the additional nursing costs to patients. We found that calibrating weights on the various terms in the objective function requires some care. Although it is relatively easy to determine the direct cost of additional nurses, it can be difficult to specify the weight that should be associated with a patient who is assigned to a nonpreferred hospital ward because of insufficient space in the preferred ward, or to a surgery that—with some probability—will need to be delayed if the number of surgical patients in a ward exceeds the usual allocation *and* the number of nonsurgical patients pushes the total number of patients over the ward's capacity. In addition to these fundamental issues, we found that putting the entire weight on nursing costs did not necessarily lead to schedules that are bad for the patients, and that putting the entire weight on patient-related costs did not necessarily lead to solutions with high nursing costs. Overall, we found that—for our formulation—the impact of the core nurse staffing levels was much stronger than almost all other effects, because these levels indirectly change the *effective* weights on the terms in the objective function. Thus, we recommend that researchers and practitioners who wish to implement an approach similar to ours take advantage of the flexibility that such models afford and try a wide range of weights and values of controllable parameters to understand how they affect the characteristics of the solutions.

## Implementation and Benefits

As previously mentioned, our team implemented optimization code using AMPL and CPLEX. After several iterations to account for constraints and considerations that surfaced as we generated potential solutions, we were able to generate near-optimal schedules that satisfied KP's needs. KP first implemented our proposed new schedule designed for the existing hospital for its final six months of operation, and the proposed schedule for the new hospital when that hospital opened on July 1, 2014. The smoothing team had not developed a feasible schedule for the old hospital or a complete schedule for the new hospital; therefore, although we had no schedules for comparison, as we note previously, our solutions are within 1.5 percent of optimum.

KP staff does not include any members who are familiar with AMPL and CPLEX. As a result, the current organization does not have the technical skills to update the schedule. However, because the block plan needs to be revised only infrequently (no more frequently than every few years), KP could call upon technical assistance from outside consultants whenever it requires a new block schedule.

Here, we summarize several benefits KP obtained from this project.

- Virtually all (greater than 98 percent) of the previous monthly block requirements were compressed into four weeks, thereby releasing capacity for approximately six percent more surgeries (assuming the same mix of procedure types) and using *fewer* beds in the various wards. This allowed KP to reduce its backlog in the short run and to potentially handle a greater number of patients in the longer term. Because KP is a not-for-profit organization, it is difficult to put a dollar figure on this benefit; however, the flexibility to grow without additional infrastructure costs is valuable to any organization of this type.

- We were able to identify the infeasibility of weekday-only surgery schedules for the orthopedics department and also develop a schedule with minimal weekend blocks that would satisfy the ward bed capacity constraint in the old hospital; this would not have been possible without our optimization model. KP implemented a schedule with limited Saturday surgeries in the old hospital, and Dr. Barber reported that it worked well. However, it proved to be difficult to continue Saturday surgeries for a prolonged period due to the inconvenience to the surgeons and other auxiliary medical staff (e.g., physical therapists); therefore, it was necessary to revert to a weekday-only schedule. More beds were allocated to orthopedic patients to accommodate this change. Interestingly, the smoothing team applied this strategy with weekend surgeries in the maternity ward, which was not included in the optimization model because a great majority of its bed occupancy is not a consequence of surgical procedures. The strategy entailed scheduling induced births on weekends, which then also made it possible for all maternity-ward patients to have single (i.e., not shared) rooms because of the lower peak bed occupancy levels.
- Scheduling of teaching days for the surgeons (within the available options) was implemented as part of the block schedule optimization problem, rather than taking them as given.
- Days with excess admissions, which would have required additional nurses, were eliminated, as indicated by Dr. Barber.
- Surgery cancelations due to insufficient bed availability were eliminated. Dr. Barber emphasized the importance of reducing cancelations, especially for patients who would need to go to the intensive care unit, because these patients generally need more care.
- Nurse workloads were more balanced as a result of more stable bed occupancy levels. Our model predicted this improvement, and Dr. Barber confirmed it. Studies in our literature review show that more stable nurse workloads result in better patient care.

A schedule based on low core nurse staffing levels—and thus also more widely varying total nurse staffing levels across the days of the week—would have allowed a reduction in the number of nursing shifts, which could have produced an estimated savings of up to $1 million annually for each of the larger wards. Devising an acceptable assignment of nurses to days of the week proved to be difficult for the nurse supervisors, however. This difficulty, along with resistance from the nursing union, ultimately led KP to implement a schedule based on higher core nurse staffing levels, thus limiting the cost savings. In other settings, however, a greater reduction in nursing costs may be possible, even if—as in our implementation—both nursing costs and patient-related costs are included in the objective function.

Since KP implemented the initial schedule, it has made a few minor changes to the schedule. These include the addition of an OR, which had been held in reserve for future growth. This room is dedicated to emergent cases in the afternoon and to various subspecialties in the morning.

## Appendix. Problem Formulation

### Sets and Their Corresponding Elements

- $s \in \mathcal{S}$: set of subspecialties (i.e., orthopedics, spinal surgery, gynecology, urology, podiatry, plastic surgery, general surgery, pediatric neurology, pediatric spinal surgery, pediatric general surgery, other pediatric surgery);
- $b \in \mathcal{B}$: set of block types;
- $b \in \tilde{\mathcal{B}}_s$: set of block types corresponding to subspecialty type $s$;
- $r \in \mathcal{R}$: set of rooms;
- $d \in \mathcal{D}$: set of days;
- $w \in \mathcal{W}$: set of weeks;
- $c \in \mathcal{C}$: set of combinations of block types that may span multiple subspecialties and utilize common resources (e.g., specific operating rooms or surgeons);
- $d \in \bar{\mathcal{D}}_w$: set of days in week $w$;
- $d \in \tilde{\mathcal{D}}_b$: set of days on which block type $b$ can be scheduled;
- $b \in \bar{\mathcal{B}}_{cd}$: set of blocks from a combination of block types $c$ that can be scheduled on day $d$;
- $f \in \mathcal{F}$: set of floors, where each floor corresponds to a hospital ward;
- $b \in \hat{\mathcal{B}}_f$: set of blocks for which the surgical patients should be assigned to floor $f$ following surgery.

### Parameters

- $a_f$: patient admissions threshold on floor $f$ above which another nurse is required;

- $k_b$: baseline (steady state) scheduling requirement for block type $b$ over the scheduling horizon;
- $q_b$: number of procedures (surgeries) in block type $b$;
- $\underline{q}_b^w$: minimum number of blocks of type $b$ that must be assigned per week;
- $\bar{q}_b^w$: maximum number of blocks of type $b$ that can be assigned per week;
- $\underline{q}_{sd}$: minimum number of blocks in subspecialty $s$ that must be scheduled on day $d$;
- $\bar{q}_{sd}$: maximum number of blocks in subspecialty $s$ that can be scheduled on day $d$;
- $\underline{\underline{q}}_s^w$: minimum number of blocks in subspecialty $s$ that must be assigned per week;
- $\bar{\bar{q}}_s^w$: maximum number of blocks in subspecialty $s$ that can be assigned per week;
- $\bar{q}_b^h$: maximum number of backlogged blocks of type $b$ that can be assigned over the horizon;
- $g_c$: maximum number of block types from a combination of block types $c$ that can be scheduled on a given day;
- $\alpha_b$: upper bound on the fraction of the total number of blocks of type $b$ scheduled over the horizon that can be assigned within the same week;
- $\tilde{\alpha}_s$: upper bound on the fraction of the total number of blocks for subspecialty $s$ scheduled over the horizon that can be assigned within the same week;
- $i_{fd}$: expected number of surgery patients from the prior scheduling horizon who will remain on floor $f$ on day $d$;
- $p_{bdd'}$: expected number of patients from block type $b$ remaining in the hospital on day $d'$ who had surgery on day $d$;
- $n_{fd}$: baseline (core) number of nurses preassigned to floor $f$ on day $d$;
- $\rho_f$: maximum number of patients each nurse working on floor $f$ can handle (i.e., patient-to-nurse ratio);
- $e_f$: effective bed capacity on floor $f$ for postsurgical patients;
- $\gamma_1, \gamma_2, \ldots, \gamma_6$: coefficients for objective function terms.

## Decision Variables

- Primary decision variables
  - ∗ $X_{brd}$: 1 if block type $b$ is assigned to (operating) room $r$ on day $d$;
  - ∗ $N_{fd}$: number of nurses working on floor $f$ on day $d$.
- Auxiliary variables
  - ∗ $Y_{fd}$: 1 if an additional nurse is needed to handle excess admissions (above the threshold of $a_f$ for floor $f$) on day $d$; 0 otherwise;
  - ∗ $M_{fd}$: expected number of patients occupying beds on floor $f$ on day $d$;
  - ∗ $\hat{M}_f$: peak value of the expected number of patients on floor $f$ during the scheduling horizon;
  - ∗ $M_f^{\max}$: positive part of the maximum difference (across the days) between the expected number of patients and the effective bed capacity on floor $f$ over the horizon.

We minimize the following objective function:

$$\gamma_1 \cdot \sum_{f\in\mathscr{F}, d\in\mathscr{D}} Y_{fd} + \gamma_2 \cdot \sum_{f\in\mathscr{F}, d\in\mathscr{D}} N_{fd} + \gamma_3 \cdot \sum_{f\in\mathscr{F}, d\in\mathscr{D}} \{M_{fd} - e_f\}$$
$$- \gamma_4 \cdot \sum_{b\in\mathscr{B}} \left\{ \sum_{r\in\mathscr{R}, d\in\mathscr{D}} X_{brd} - k_b \right\} + \gamma_5 \cdot \sum_{f\in\mathscr{F}} M_f^{\max}$$
$$+ \gamma_6 \sum_{f\in\mathscr{F}} \sum_{d\in\mathscr{D}} \left| \frac{M_{fd} - \bar{M}_f}{\mathrm{card}(\mathscr{D})} \right|, \tag{A.1}$$

where $\bar{M}_f = \sum_{d\in\mathscr{D}} M_{fd}/\mathrm{card}(\mathscr{D})$ (i.e., the average number of patients on floor $f$ during the horizon) and card$(\cdot)$ denotes the cardinality of the set.

The objective is to minimize a weighted sum of six factors, where the weights may correspond to true costs, or may be selected in a way that helps to smooth the schedule and (or) break ties among solutions that are effectively identical. The first term represents the costs for nurses required to handle above-threshold admissions, summed across all floors and days. The second term represents the cost for nurses to satisfy the patient-to-nurse ratios, summed across all floors and days. The third term corresponds to the penalty for patient days in excess of the effective bed capacity, summed across all floors and days. This term is intended to capture the degradation in care and consequent medical outcomes when a patient cannot be assigned to the most suitable floor. The fourth term reflects the total bonus (negative cost) for all backlog blocks scheduled. The fifth term represents a penalty for the peak bed occupancy for each floor, assuming that all patients are assigned to the most suitable floor, summed across floors. This term helps to smooth the bed occupancy over time and to break ties among solutions whose objectives would otherwise be equal. Finally, the last term is a penalty for the mean absolute deviations (MAD) of the expected number of patients, summed across the various floors.

The constraints fall into several categories: (1) nurse staffing requirements; (2) feasible assignments of blocks to ORs; (3) backlog balance equations and bounds; (4) smoothing and spacing constraints; (5) bounds on certain aggregates of assignments over applicable periods; and (6) patient inventory balance equations. Next, we present the constraints and provide associated explanations.

## Nurse Requirements

$$M_{fd} \le \rho_f \cdot N_{fd} \quad \forall\, f \in \mathscr{F}, d \in \mathscr{D}. \tag{A.2}$$

$$N_{fd} \ge n_{fd} \quad \forall\, f \in \mathscr{F}, d \in \mathscr{D}. \tag{A.3}$$

$$\sum_{r\in\mathscr{R}} \sum_{b\in\hat{\mathscr{B}}_f} q_b X_{brd} - a_f \le \mathscr{M} \cdot Y_{fd} \quad \forall\, f \in \mathscr{F}, d \in \mathscr{D}, \tag{A.4}$$

where $\mathscr{M}$ is a sufficiently large number, which may vary with $f$ and $d$, if appropriate.

Constraints (A.2) state that the number of patients on a floor on a given day is restricted by the nurse staffing level, while constraints (A.3) require that the number of nurses for each day and floor meet or exceed the minimum (i.e., core) staffing level. Constraints (A.4) ensure that an extra nurse is assigned to floor $f$ on any day $d$ for which the number of admissions exceeds the threshold $a_f$. We note that the number of excess admissions is bounded by the difference between the effective bed capacity and $a_f$; therefore, $\mathscr{M}$ can be set accordingly.

### Operating Room and Associated Capacities

$$\sum_{b\in\mathscr{B}} X_{brd} \le 1 \quad \forall\, r \in \mathscr{R},\, d \in \mathscr{D}. \tag{A.5}$$

Constraints (A.5) allow at most one block type to be scheduled in a room on each day.

### Backlog

$$k_b \le \sum_{r\in\mathscr{R},\, d\in\mathscr{D}} X_{brd} \le k_b + \bar{q}_b^h \quad \forall\, s \in \mathscr{S},\, b \in \tilde{\mathscr{B}}_s. \tag{A.6}$$

Each constraint in (A.6) ensures that the number of blocks scheduled over the horizon for each subspeciality and corresponding block types within the subspecialty lies between the baseline number and the baseline plus the maximum number of backlogged blocks that can be scheduled. This ensures that the baseline demand is satisfied and the backlog is not drawn down too quickly.

### Smoothing and Spacing Constraints

$$\underline{\underline{q}}_s^w \le \sum_{b\in\tilde{\mathscr{B}}_s,\, r\in\mathscr{R},\, d\in\bar{\mathscr{D}}_w\cap\tilde{\mathscr{D}}_b} X_{brd} \le \bar{\bar{q}}_s^w \quad \forall\, s \in \mathscr{S},\, w \in \mathscr{W}. \tag{A.7}$$

Constraints (A.7) restrict the number of blocks scheduled for each subspecialty to lie between certain minimum and maximum numbers each week. Constraints of this form also can be used to limit the assignments of specific subspecialties or block types to acceptable combinations of days.

$$\sum_{r\in\mathscr{R},\, d\in\bar{\mathscr{D}}_w} X_{brd} \le \alpha_b \cdot \sum_{r\in\mathscr{R},\, d\in\mathscr{D}} X_{brd} \quad \forall\, b \in \mathscr{B},\, w \in \mathscr{W}. \tag{A.8}$$

$$\sum_{b\in\tilde{\mathscr{B}}_s,\, r\in\mathscr{R},\, d\in\bar{\mathscr{D}}_w} X_{brd} \le \tilde{\alpha}_s \cdot \sum_{b\in\tilde{\mathscr{B}}_s,\, r\in\mathscr{R},\, d\in\mathscr{D}} X_{brd} \quad \forall\, s \in \mathscr{S},\, w \in \mathscr{W}. \tag{A.9}$$

Constraints (A.8) and (A.9) help to smooth the number of blocks of a given block type or subspecialty over the planning horizon by permitting no more than a certain proportion of these blocks to be scheduled during each week. A more restrictive version of constraints (A.8) ensures that the number of block type $b$ scheduled on a given day does not exceed some fraction of the number of blocks of type $b$ scheduled during that week.

$$\sum_{r\in\mathscr{R},\, d\in\bar{\mathscr{D}}_{w+1}} X_{brd} - 1 \le \sum_{r\in\mathscr{R},\, d\in\bar{\mathscr{D}}_w} X_{brd} \le \sum_{r\in\mathscr{R},\, d\in\bar{\mathscr{D}}_{w+1}} X_{brd} + 1$$
$$\forall\, b \in \mathscr{B},\, w \in \mathscr{W} \text{ for } w < \text{card}\,(\mathscr{W}). \tag{A.10}$$

Constraints (A.10) ensure that the total number of blocks of a given type differs by no more than one from one week to the next.

$$\sum_{r\in\mathscr{R}} X_{br,d+1} - 1 \le \sum_{r\in\mathscr{R}} X_{brd} \le \sum_{r\in\mathscr{R}} X_{br,d+1} + 1$$
$$\forall\, b \in \mathscr{B},\;\; d \in \mathscr{D} \text{ for } d < \text{card}\,(\mathscr{D}). \tag{A.11}$$

Constraints (A.11) prevent the total number of blocks of a given type on a given day from differing by more than one between that day and the next.

$$M_{fd} \le \hat{M}_f \quad \forall\, f \in \mathscr{F},\, d \in \mathscr{D}. \tag{A.12}$$

Constraints (A.12) define the peak value of the expected number of patients on a floor during the scheduling horizon by ensuring that it is greater than or equal to the expected number of patients on that floor on all days in the horizon.

$$M_f^{\max} \ge \hat{M}_f - e_f \quad \forall\, f \in \mathscr{F}. \tag{A.13}$$

Constraints (A.13) define the maximum value (across the days in the horizon) of the expected number of patients exceeding the effective bed capacity on each floor, assuming that each patient is assigned to the medically preferred floor.

### Additional Bounds on Certain Aggregates of Assignments

$$\underline{q}_{sd} \le \sum_{b\in\tilde{\mathscr{B}}_s,\, r\in\mathscr{R}} X_{brd} \le \bar{q}_{sd} \quad \forall\, s \in \mathscr{S},\, d \in \mathscr{D}. \tag{A.14}$$

Constraints (A.14) ensure that the number of scheduled blocks for each subspecialty is within a specified range for each day within the time horizon; these constraints often arise due to surgeon availability.

$$\sum_{b\in\bar{\mathscr{B}}_{cd},\, r\in\mathscr{R}} X_{brd} \le g_c \quad \forall\, c \in \mathscr{C},\, d \in \mathscr{D}. \tag{A.15}$$

Constraints (A.15) prevent the total number of scheduled blocks within a given combination of block types from exceeding an upper limit on each day. One example is a constraint on the total number of urology and spinal surgery blocks because of the availability of rooms with the necessary equipment. Another example is an upper limit on the number of general surgery blocks because of teaching days on which surgeons are not available. More general versions of this constraint allow block types to be scheduled either on one set of days or on another set of days, but not both.

$$\underline{q}_b^w \le \sum_{r\in\mathscr{R},\, d\in\bar{\mathscr{D}}_w\cap\tilde{\mathscr{D}}_b} X_{brd} \le \bar{q}_b^w \quad \forall\, b \in \mathscr{B},\, w \in \mathscr{W}. \tag{A.16}$$

Constraints (A.16) place bounds on the total number of a given block type scheduled during a specified set of days within a given week. Such constraints can account for resource limitations or scheduling rules. As one example, pediatric general surgery blocks must be scheduled on one and only one of the following pairs of days within a given week: Tuesday–Thursday, Tuesday–Friday, or Wednesday–Friday.

### Patient Inventory

$$M_{fd} = i_{fd} + \sum_{b\in\hat{\mathscr{B}}_f}\sum_{r\in\mathscr{R}}\sum_{d'\in\mathscr{D}} p_{bdd'}\cdot X_{brd'} \quad \forall\, f\in\mathscr{F},\, d\in\mathscr{D}. \quad \text{(A.17)}$$

Constraints (A.17) define the expected value of the patient inventory on floor $f$ and day $d$ as the sum of: (1) the expected number of patients who had surgery during the previous scheduling horizon, but will remain until (at least) day $d$ in the current horizon; and (2) those who were added based on the date of their surgery within the current horizon and the distributions of their lengths of stay in the hospital.

To account for patients who have surgery near the end of one scheduling horizon and remain in the hospital past the end of that horizon, we utilize a set of auxiliary variables $i_{fd}$ representing the expected number of patients who will still be occupying a bed past the end of the scheduling horizon (by day and floor). The values of these auxiliary variables are computed at the end of each horizon as input to the problem for the next horizon when we iteratively solve the problem to achieve a schedule that reflects (near-) steady state conditions. These auxiliary variables are defined as

$$i_{f,d'-28} = \sum_{b\in\hat{\mathscr{B}}_f}\sum_{r\in\mathscr{R}}\sum_{d\in\mathscr{D}} p_{bdd'}\cdot X_{brd} \quad \forall\, f\in\mathscr{F},\, d'\in\mathscr{D}.$$

### Nonnegativity and Integrality

$$X_{brd}\text{ binary; } Y_{fd}\text{ binary; } N_{fd}\text{ integer; } M_{fd}, \hat{M}_f, M_f^{\max}\geq 0$$
$$\forall\, b\in\mathscr{B},\, r\in\mathscr{R},\, d\in\mathscr{D},\, f\in\mathscr{F}. \quad \text{(A.18)}$$

In addition to these constraints, we impose restrictions that arise as a result of scheduling decisions at a sister hospital, such as unavailability of surgeons when they are working at that hospital.

### CPLEX Settings and Solution Details

We solve the problem using CPLEX 12.6.0.1, as formulated in the appendix, without any special procedures to accelerate run times. We did, however, take care to formulate the problem instances as efficiently as possible. For example, we define variables only for the pertinent decisions and we restrict variable ranges to sensible values. As an example, for regular adult hospital wards with 24 beds, the maximum required number of nurses is five; therefore, we set the upper bound on the corresponding integer variable to five.

Problem instances with low core nurse staffing levels, not surprisingly, are more difficult to solve because the number of feasible nurse staffing profiles (i.e., the number of nurses in each hospital ward on each day) is much larger than for instances with high core nurse staffing levels. We observe that problems with moderate core nurse staffing levels are not necessarily easy to solve because the decisions ultimately entail determining which (few) days of the scheduling horizon should be allocated an extra nurse and (or) whether it is advantageous to schedule two extra nurses on one or possibly two days of the horizon. Of course, for each of these possible nurse staffing profiles for each ward, the (still difficult) problem of identifying the best assignment of blocks to operating room days exists.

Default settings in CPLEX produce solutions within one percent of optimality in a matter of seconds for the more tractable instances; however, the less tractable instances require much more time. Ultimately, we decided to solve each instance to a 1.5 percent optimality gap using settings conducive to the more difficult instances. Although we set an optimality gap of 1.5 percent, in many cases, CPLEX achieves a smaller gap than this; in a few instances, the gap amounts to tenths or hundredths of a percent. We used the following settings: (1) `nodefile 3` and (2) `memoryemphasis 1`, both of which encourage CPLEX to conserve memory; (3) `dgradient 5`, which employs "Devex" pricing for the node LPs; (4) `mipemphasis 4`, which induces a search for "hidden" integer-feasible solutions; (5) `probe 2`, which fixes variable values a priori based on a presolve procedure; (6) `rinsheur 200`, which activates CPLEX's proprietary "relaxation induced neighborhood search heuristic" every 200 nodes within the branch-and-bound tree; and (7) `mipcuts 2`, which aggressively implements cuts. In summary, the most difficult instances require conservation of computer memory, and methods to improve both the best lower bound and the best integer solution.

With these settings, CPLEX requires between a few seconds and two hours to solve the series of two to four optimization problems (i.e., iterations) required to reach steady state, except for the case with (2, 2) as the core nurse staffing level. Because three nurses (and sometimes four) are typically needed in Ward A, starting with a core staffing level of two nurses gives rise to an extremely difficult combinatorial problem, and the four iterations took over 30 hours to solve. This points to the need to judiciously set the core nurse staffing level when there is flexibility to do so. Setting the level below, but not too far below, the average requirement leaves room for the solver to optimize, while keeping computing times low to moderate.

### Acknowledgments

The authors thank Dr. Thomas Barber for introducing them to this problem and to the numerous members of Kaiser Permanente's smoothing team for helping them understand the broader issues surrounding the problem and in obtaining data. They also acknowledge the assistance of former UC Berkeley students Paul Kim, Erik Bertelli, Rebecca Tanuwidjaja, and Michael Signorotti, who participated in aspects of this project, including data analysis and parameter testing, and Professor Asoo Vakharia who participated in the early stages of the problem definition. The authors thank Dr. Ed Klotz of IBM for his helpful suggestions on CPLEX solver parameter settings. The first author worked on this project while she was a student at UC Berkeley and as an intern at Kaiser Permanente during summer 2013. The opinions expressed in this article are solely those of the authors and do not necessarily represent the views of Kaiser Permanente or its staff.

## References

Aiken LH, Clarke SP, Sloane DM, Sochalski J, Silber JH (2002) Hospital nurse staffing and patient mortality, nurse burnout and job dissatisfaction. *J. Amer. Medical Assoc.* 288(16):1987–1993.

Baker DR, Pronovost PJ, Morlock LL, Geocadin RG, Holzmueller CG (2009) Patient flow variability and unplanned readmissions to an intensive care unit. *Critical Care Medicine* 37(11):2882–2887.

Belien J, Demeulemeester E (2007) Building cyclic master surgery schedules with leveled resulting bed occupancy. *Eur. J. Oper. Res.* 176(2):1185–1204.

Belien J, Demeulemeester E (2008) A branch-and-price approach for integrated nurse and surgery scheduling. *Eur. J. Oper. Res.* 189(3):652–668.

Belien J, Demeulemeester E, Cardoen B (2009) A decision support system for cyclic master surgery scheduling with multiple objectives. *J. Scheduling* 12(2):147–161.

Blake JT (2010) Capacity planning in operating rooms. Yih Y, ed. *Handbook of Healthcare Delivery Systems* (CRC Press, Boca Raton, FL), 34-1–34-12.

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

Cardoen B, Demeulemeester E, Belien J (2010) Operating room planning and scheduling: A literature review. *Eur. J. Oper. Res.* 201(3): 921–932.

Cimiotti JP, Aiken LH, Sloane DM, Wu ES (2012) Nurse staffing, burnout and health care-associated infection. *Amer. J. Infection Control* 40(7):486–490.

Guerriero F, Guido R (2011) Operational research in the management of the operating theatre: A survey. *Health Care Management Sci.* 14(1):9–114.

Holte M, Mannino C (2013) The implementor/adversary algorithm for the cyclic and robust scheduling problem in health-care. *Eur. J. Oper. Res.* 226(3):551–559.

May JH, Spangler WE, Strum DP, Vargas LG (2011) The surgical scheduling problem: Current research and future opportunities. *Production Oper. Management* 20(3):392–405.

Price C, Golden B, Harrington M, Konewko R, Wasil E, Herring W (2011) Reducing boarding in a post-anesthesia care unit. *Production Oper. Management* 20(3):431–441.

Santibáñez P, Begen M, Atkins D (2007) Surgical block scheduling in a system of hospitals: An application to resource and wait list management in a British Columbia health authority. *Health Care Management Sci.* 10(3):269–282.

Testi A, Tanfani E, Torre G (2007) A three-phase approach for operating theatre schedules. *Health Care Management Sci.* 10(2):163–172.

Vanberkel PT, Boucherie RJ, Hans EW, Hurink JL, Litvak N (2009) A survey of health care models that encompass multiple departments. *Internat. J. Health Management Inform.* 1(1):37–69.

Vanberkel PT, Boucherie RJ, Hans EW, Hurink JL, van Lent WAM, van Harten WH (2011) An exact approach for relating recovering surgical patient workload to the master surgical schedule. *J. Oper. Res. Soc.* 62(10):1851–1860.

van Essen JT, Bosch JM (2013) Reducing the number of required beds by rearranging the OR-schedule. *OR Spectrum* 36(3):585–605.

van Oostrum JM, Van Houdenhoven M, Hurink JL, Hans EW, Wullink G, Kazemier G (2008) A master surgical scheduling approach for cyclic scheduling in operating room departments. *OR Spectrum* 30(2):355–374.

## Verification Letter

Thomas C Barber, MD, The Permanente Medical Group, 1 Kaiser Plaza, Oakland, CA 94612, writes:

"I am the Associate Physician in Chief for the Kaiser East Bay. My responsibilities include overseeing perioperative services, tertiary care, and graduate medical education.

"I had the pleasure of working with Candace Yano and Brittney Benchoff on a project around operating room scheduling. The model developed for surgical block scheduling was insightful, and useful. When we moved to our new hospital in July of 2014 our surgical block scheduling was based on the model developed. Implementation was successful and decreased variability in hospital census on each surgical floor.

"The article "*Kaiser Permanente Oakland Medical Center Improves Operating Room Schedule Planning*" is accurate, and reflects the work done in this project."

**Brittney Benchoff** received a bachelor's degree in industrial engineering from West Virginia University and a master's degree in operations research from University of California, Berkeley. Her research centered on the application of optimization in the healthcare industry. Brittney works as a product manager at Alpine Data, a machine learning software company in San Francisco.

**Candace Arai Yano** is a professor of industrial engineering and operations research as well as the Kalbach Chair of Business Administration and professor of operations and information technology management at the Haas School of Business at the University of California, Berkeley. Her primary research areas are supply chain management and operations-marketing interface issues. She holds an AB in economics, an MS in operations research, and an MS and PhD in industrial engineering from Stanford University. Professor Yano is a Fellow of the Institute for Operations Research and the Management Sciences (INFORMS) as well as the Institute of Industrial and Systems Engineers (IISE; formerly Institute of Industrial Engineers).

**Alexandra Newman** is a professor in the Mechanical Engineering Department at the Colorado School of Mines (CSM). Prior to joining CSM, she was a research assistant professor at the Naval Postgraduate School in the Operations Research Department. She obtained her BS in applied mathematics at the University of Chicago and her PhD in industrial engineering and operations research at the University of California, Berkeley. Professor Newman specializes in deterministic optimization modeling, especially as it applies to energy and mining systems, and to logistics, transportation and routing. She received a Fulbright Fellowship to work with industrial engineers on mining problems at the University of Chile in 2010, and was awarded the INFORMS Prize for the Teaching of Operations Research and Management Science Practice in 2013.