Operations Research

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

Integrated Anesthesiologist and Room Scheduling for Surgeries: Methodology and Application

Sandeep Rath, Kumar Rajaram, Aman Mahajan

To cite this article:
Sandeep Rath, Kumar Rajaram, Aman Mahajan (2017) Integrated Anesthesiologist and Room Scheduling for Surgeries: Methodology and Application. Operations Research 65(6):1460-1478. https://doi.org/10.1287/opre.2017.1634

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

OPERATIONS RESEARCH
Vol. 65, No. 6, November–December 2017, pp. 1460–1478
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Integrated Anesthesiologist and Room Scheduling for Surgeries: Methodology and Application

Sandeep Rath,[a] **Kumar Rajaram,**[b] **Aman Mahajan**[c]

[a] Kenan-Flagler Business School, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599; [b] University of California, Los Angeles (UCLA) Anderson School of Management, Los Angeles, California 90095; [c] Department of Anesthesiology and Perioperative Medicine, David Geffen School of Medicine at UCLA, Los Angeles, California 90095

Contact: sandeep@unc.edu (SR); kumar.rajaram@anderson.ucla.edu, http://orcid.org/0000-0002-6939-8219 (KR); amahajan@mednet.ucla.edu (AM)

Received: September 12, 2015
Revised: April 17, 2016; December 14, 2016
Accepted: March 31, 2017
Published Online in Articles in Advance: July 20, 2017

Subject Classifications: healthcare: hospitals; applications: scheduling; programming: integer; stochastic
Area of Review: OR Practice

https://doi.org/10.1287/opre.2017.1634

Abstract. We consider the problem of minimizing daily expected resource usage and overtime costs across multiple parallel resources such as anesthesiologists and operating rooms, which are used to conduct a variety of surgical procedures at large multispecialty hospitals. To address this problem, we develop a two-stage, mixed-integer stochastic dynamic programming model with recourse. The first stage allocates these resources across multiple surgeries with uncertain durations and prescribes the sequence of surgeries to these resources. The second stage determines actual start times to surgeries based on realized durations of preceding surgeries and assigns overtime to resources to ensure all surgeries are completed using the allocation and sequence determined in the first stage. We develop a data-driven robust optimization method that solves large-scale real-sized versions of this model close to optimality. We validate and implement this model as a decision support system at the UCLA Ronald Reagan Medical Center. This system effectively incorporates the flexibility in the resources and uncertainty in surgical durations, and explicitly trades off resource usage and overtime costs. This has increased the average daily utilization of the anesthesiologists by 3.5% and of the operating rooms by 3.8%. This has led to an average daily cost savings of around 7% or estimated to be $2.2 million on an annual basis. In addition, the insights based on this model have significantly influenced decision making at the operating services department at this hospital.

Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2017.1634.

Keywords: healthcare operations • mixed integer stochastic dynamic programming • robust optimization

1. Introduction

Surgical procedures are complex tasks requiring the use of several specialized and expensive resources. In a hospital, the operating services department is responsible for managing resources used in surgical procedures. Every day, this department assigns to each surgery an operating room (OR), an anesthesiologist, a nursing team, and the requisite surgical materials. The department also determines the sequence in which these surgeries will be performed and the scheduled start times. While performing these actions, the department ensures that the cost of the OR suite is minimized by reducing resource usage and overtime costs.

The operating services departments at large hospitals devote significant amount of time in making these resource management decisions. The complexity of these decisions is due to the following four primary reasons. First, OR resources are expensive (Macario 2010), and in short supply (Orkin et al. 2013), and thus surgeries are performed in highly resource-constrained environments. Second, surgical procedures are often very specialized. Therefore, equipment and facility requirements govern whether a procedure can be performed in a particular room. Anesthesiologist assignments too are dictated by specialty. Studies have demonstrated that not only do surgeons often prefer to have an anesthesiologist of the required subspecialty (Ghaly 2014), outcome indicators of surgical procedures are significantly better when anesthesia is delivered by an anesthesiologist with experience in that particular subspecialty (McNicol 1997). Pardo (2014) predicts that increasingly anesthesiologists will be assigned by their subspecialties. Third, the durations of surgical procedures are very difficult to predict (Kayis et al. 2012). This is partly because there is a large number of procedures, and newer procedures are constantly being developed (AMA 2016). Consequently, historical data on all these procedures is not available. Furthermore, a surgeon's estimates of durations are often unreliable. Studies have demonstrated systematic underestimation as well as overestimation of procedure times by surgeons while scheduling surgeries. Some surgeons overestimate the duration when they do not have enough cases to fill their scheduled

block time, while others may underestimate the time when they wish to fill in more cases (Laskin et al. 2013). Finally, the scale of large hospitals, in terms of the number of ORs, procedures conducted, the number and types of equipment and anesthesiologists used, makes the simultaneous scheduling of multiple resources a computationally challenging task.

We were exposed to these complexities at the operating services department of the UCLA Ronald Reagan Medical Center (RRMC), a large multispecialty hospital, which consistently ranks amongst the best five hospitals in the United States.[1] Management of this department felt that the daily resource allocation decision played a significant role in overall department cost, and in the service quality delivered to patients. They believed that these aspects can be significantly improved by developing an analytical model-based approach that considered the key complexities in this environment, and applied historical surgical data to decide resource assignment and scheduling. This paper describes the development, implementation, and evaluation of a model-based decision support system that uses a data-driven robust optimization procedure to determine the daily scheduling of anesthesiologists and rooms for elective surgeries at the UCLA RRMC.

There is a large body of literature on elective surgery scheduling. Min and Yih (2010) consider scheduling elective surgeries under uncertainty in surgery durations and downstream capacity constraints. Gupta (2007) discusses the broader issues of managing OR suites for elective surgeries. In this context, we study the problem of integrated scheduling of anesthesiologists and ORs to surgeries. The literature on scheduling anesthesiologists to surgeries includes Marcon and Dexter (2006), McIntosh et al. (2006), Dexter and Wachtel (2014), Dexter et al. (2016). However, these papers do not consider the joint scheduling with ORs and uncertainty in surgical durations, both of which were critical features in our application. Furthermore, the solution methods in these papers employ experience-based heuristics rather than optimization-based methods. The literature for scheduling ORs under uncertain surgery durations has primarily been focused on single resource type corresponding to the OR. Research in this area include Denton and Gupta (2003), Green and Savin (2008), Mancilla and Storer (2012), Mak et al. (2014a). In addition, Denton et al. (2010) solve the problem of assignment of surgeries to multiple parallel ORs under fixed costs of ORs and variable overtime costs. However, none of these papers consider multiple resources, the simultaneous sequencing and start times of surgeries or are tested with data in a large-scale application context. While a single resource type may be sufficient for specialized surgery suites, multispecialty hospitals like the UCLA RRMC require a holistic solution of surgery scheduling that simultaneously optimizes on all specialized parallel and multiple resources.

The literature on multiple resources can be classified in two broad categories: serial and parallel multiple resources. The research on serial multiple resources focuses on analyzing the impact of decisions made on an upstream resource such as the ORs on a downstream resource such as a postanesthesia care unit (Marcon and Dexter 2006, Augusto et al. 2010, Saadouli et al. 2015). Additional work in this area that considers other upstream and downstream resources include Cardoen et al. (2009a, b), and Gul et al. (2011). However, none of these papers consider parallel multiple resources (such as anesthesiologists and ORs), which are specialized and have to be scheduled simultaneously under uncertainty in surgical durations. These were important aspects in the application and significantly complicated the optimization model. The literature related to parallel multiple resource types is relatively scarce. Beliën and Demeulemeester (2008), Meskens et al. (2013) consider integrated OR scheduling with parallel multiple resources under deterministic surgery durations. Batun et al. (2011) consider scheduling of surgeries given two parallel resource types: ORs and surgeons under stochastic surgery durations. However, they do not consider specializations of rooms and anesthesiologists, and consider a problem significantly smaller in scale than in our application. For the scale of problem at the UCLA RRMC, sample average approximation (SAA) based stochastic optimization procedures as used in Denton and Gupta (2003), Min and Yih (2010) were intractable. This is due to the large number of possible integer assignments in the first stage, which increases the number of samples required to achieve convergence in objective value and solution. These difficulties in employing this method has been described in a more general context by Kleywegt et al. (2002). Furthermore, the overall complexity of the large-scale problem in our application precluded finding even feasible solutions using leading commercially available solvers such as `ddsip` (Märkert and Gollmer 2008) that employ the state-of-the-art procedures for solving stochastic programs such as dual-decomposition methods (Carøe and Schultz 1999). We describe this in our computational analysis. To circumvent these problems, we use a robust optimization procedure (Bertsimas and Thiele 2006, Bertsimas et al. 2013). While a similar approach has been used by Denton et al. (2010) and Mak et al. (2014b), our work extends theirs by considering multiple resource types. This extension requires significant modification to existing solution methods.

Our paper makes the following contributions. First, we consider two types of parallel resources, which are of critical importance to specialties: ORs and

anesthesiologists, and we simultaneously optimize their assignment and sequencing. Second, we develop an efficient solution method using robust optimization to provide effective solutions to large-scale problems. An important element when applying robust optimization is the estimation of an uncertainty set. We develop an estimation procedure to estimate the sets using historical data. This data-driven robust optimization approach was successful in solving the full-scale problem for the entire surgery suite at the RRMC within 25 minutes, with a performance gap within 5% from the lower bound. Third, our methodology significantly outperforms the best benchmark procedures in the literature. Fourth, we develop a model-based decision support system, which has been validated and implemented at the UCLA RRMC. To the best of our knowledge, this is the first real implementation of robust optimization in the healthcare industry. This system effectively incorporates the flexibility in the resources and uncertainty in surgical durations, and explicitly trades off costs. This has considerably improved upon current practice, and has resulted in average daily cost savings of around 7% or estimated to be $2.2 million on an annual basis. Further, the insights from our work has had a notable impact on decision making at the hospital.

The remainder of the paper is organized as follows. Section 2 provides a detailed problem description at the UCLA RRMC. Section 3 presents the model formulation, properties, and solution procedure. In Section 4, we describe the procedure for parameter estimation and model calibration. Section 5 provides the results of the computational analysis. Section 6 describes the implementation of this model at the UCLA RRMC, presents the financial benefits, provides managerial insights, and describes the organizational impact of this work.

2. Problem Description

Operating services is one of the largest departments at the UCLA RRMC with around $120 million in annual revenues representing about 10% of this hospital's revenues. This department serves around 27,000 patients annually by conducting around 2,700 *types of* elective and emergency surgical procedures across 12 specialties. Emergency surgeries are conducted in three dedicated ORs with a separate team of anesthesiologists. Since emergency surgeries are separated from elective surgeries and account for only about 15% of revenues, management of this department asked us to focus solely on elective surgeries. To perform these surgeries, the operating services department uses 23 ORs, which are further divided across these 12 specialties that require specific equipment. The details on the number of rooms that can perform each specialty is provided in Table 1. General surgery procedures can be performed in any of these 23 ORs dedicated for the exclusive use of elective surgeries. Add-on surgeries are not considered here as depending on availability, they are assigned to the ORs dedicated to emergency procedures. Surgeries are scheduled to start in ORs only between 7 A.M. and 3 P.M. Further, there are fixed costs for opening an OR each day. This consists of an initial cleaning and equipment setup costs along with daily nurse and technician staffing costs, whose assignments do not depend on specialty. In addition, overtime costs are incurred for nurses and technicians if the rooms are required to be open beyond 3 P.M. Finally, these ORs are scheduled and staffed simultaneously.

Table 1. Summary of Resource by Specialty

Surgery specialty	Number of ORs available	Number of anesthesiologists available
Vascular	1	9
Neuro	3	10
Plastics	23	NA
ENT	23	NA
Urology	23	NA
Liver	1	8
Thoracic	2	5
Cardiac	3	14
Trauma	1	NA
Pediatric	2	12
Eye surgery	23	NA
General	23	NA

There were 92 anesthesiologists at the UCLA RRMC divided across these 12 specialties. The number of anesthesiologists by specialty is also shown in Table 1. The assignment of anesthesiologists is according to the specialty required for the surgery. Anesthesia for surgeries in some specialties can be administered by any anesthesiologist. Such specialties are denoted by NA in this table. There are three shifts of equal duration for the anesthesiologists: day (7 A.M. till 3 P.M.), late (11 A.M. till 7 P.M.), and night (7 P.M. till 3 A.M.). Each anesthesiologist is preassigned to exactly one shift, and thus the regular working hours for each anesthesiologist is eight hours. In addition, anesthesiologists can only be assigned to surgeries that begin during their shift. Overtime costs for anesthesiologists are incurred if surgeries in progress exceed the duration of the shift. A certain number of anesthesiologists who are not scheduled to work on a given day are asked to be on standby or on call, so that they can be called to work if necessary. However, when anesthesiologists are assigned from on call, there are significant costs for using such an option. Anesthesiologists assigned from on call do not incur overtime costs. The anesthesiologists on call are informed of their status the previous day and assigned surgeries that day as required.

It is important to note that in the context of large multispecialty hospitals such as the UCLA RRMC,

surgeons are not part of the operating services department. They are usually from the independently administered specialty departments at this hospital, and on some occasions, can be from other hospitals. The surgeons bring their patients and use the operating services department as a service provider. Thus the operating services department does not have the option of assigning surgeons to patients. For this reason, we assume that each surgery-surgeon combination is already set and we consider them together. This ensures that each surgery has a clear and unchangeable link to the surgeon. This aspect is also consistent with the literature in this area (Dexter and Traub 2002, Marques et al. 2014).

Typically, a request to schedule a surgery is initiated by the surgeon on behalf of the patient with general admissions at the hospital. This request is assigned a date based on the earliest availability in the block reservations for the particular specialty. Once all the elective surgery requests have been received the day before the surgery, the operating services department decides which OR to open, finalizes assignment of these rooms, and anesthesiologist to surgeries, determines start times of surgeries, and effectively specifies the sequence of all the surgeries. These decisions are made in the previous day for all the surgeries that need to be conducted in the following day. Consequently, the planning horizon is a single day. The current planning process to make these decisions uses an experience-based practitioner's heuristic. Such types of heuristics have been reported in the literature (Dexter and Traub 2002, Cardoen et al. 2010). The practitioner's heuristic consists of the following steps:

Step 1. Assign surgeries to ORs in sequential fashion in order of start times requested by the surgeons, by surgery specialty, and duration estimates from surgeons, until the last surgery in the room can start before the end of the shift for the OR.

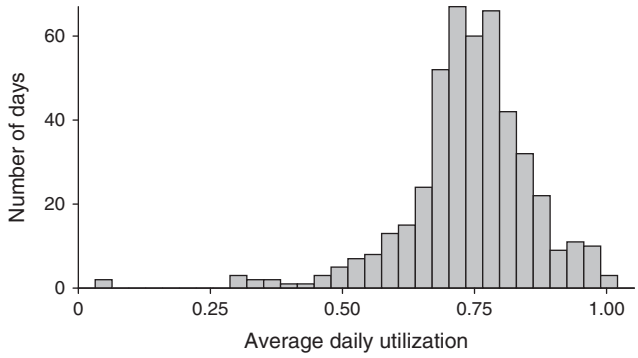
Step 2. Assign one anesthesiologist to each room so that the anesthesiologist can perform most of the surgeries in the room.

Step 3. A few anesthesiologists are assigned to surgeries across rooms to ensure all surgeries have been assigned an anesthesiologist by specialty.

Step 4. If above plan cannot be implemented by anesthesiologists on regular duty, assign anesthesiologists from on call.

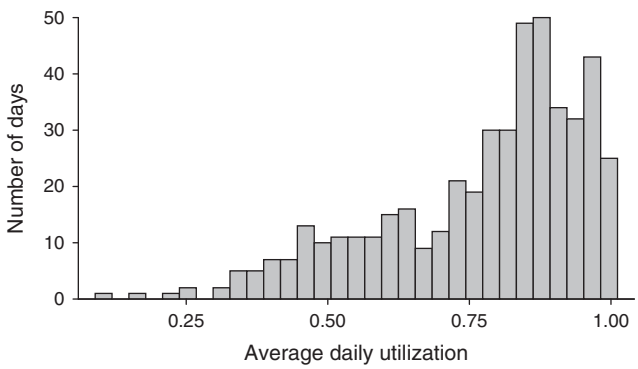
While this practitioners heuristic is easy to understand and implement, it does not consider two important aspects. First, it does not explicitly consider uncertainty in surgical durations. Second, it does not directly use the feature that most anesthesiologists and ORs can perform more than one specialty. Thus, it does not exploit the flexibility in these resources. Define utilization as the fraction of the available shift time that is used by a particular resource. Inefficient assignment and scheduling of anesthesiologists and ORs to surgeries leads to low utilization and overtime of these resources. As seen in Figure 1, average daily utilization across the anesthesiologists is close to 0.75, with around 25% of days having an average daily utilization of less than 0.70. However, despite these lower levels of utilization, the average number anesthesiologists on call is around six per day. Similarly, for ORs, the average daily utilization is close to 78% (Figure 2) but the average overtime per day is around 18 hours. Taken together average on call and overtime costs for anesthesiologists and rooms at this department are about 33% of revenues. A more effective optimization-based assignment and scheduling system that considers uncertainty in surgical durations and flexibility in the resources could potentially reduce overtime and on-call costs.

Figure 1. Histogram of Average Daily Utilization of Anesthesiologists



Assignment and scheduling decisions at this hospital are complicated by the large number of ORs and anesthesiologists, variety in surgical procedures, variability in anesthesiologist work load, and unpredictability in surgery durations. More details on these aspects are provided in the e-companion. Management of the operating services department felt that the current planning process did not adequately consider these complicating factors. Thus they believed that the daily expected resource usage and overtime

Figure 2. Histogram of Average Daily Utilization of ORs



costs across anesthesiologists and ORs could be considerably lowered by developing an optimization model, which led to our involvement. This model is formulated as a two-stage, mixed-integer stochastic dynamic program with recourse. The first stage of this model allocates these resources across multiple surgeries with uncertain durations, and prescribes the sequence of surgeries to these resources. Assuming that each surgery should be scheduled as early as possible, this, consequently, provides a scheduled start time for surgeries. The second stage determines the actual start times to surgeries based on realized durations of preceding surgeries, and assigns overtime to resources to ensure all surgeries are completed using the allocation and sequence determined in the first stage. The size and complexity of the problem precluded solution using conventional methods. Therefore we develop a data-driven robust optimization approach that solves large-scale real-sized versions of this model close to optimality. Next, we describe the model formulation, present its properties, and describe its solution techniques.

3. Model

We start by presenting a model formulation of the IARSP for surgeries. The planning horizon for this model is a single day. This was consistent with the requirements in the application context. Here, surgeries were finalized by the aggregate block scheduling system and released to the operating services department for detailed scheduling and staffing only the day before the surgery. Further, the availability of anesthesiologists on regular duty and on call was already determined by a longer range planning system at this department. Subsequently, the remaining decisions were the daily assignment of anesthesiologists and ORs to surgeries and determining the sequence of surgeries at these resources. The model makes these decisions in the previous day for the next day. The planning horizon of one day is also consistent with other research on OR models (Cardoen et al. 2009a and 2009b, Batun et al. 2011). To provide a precise definition of the model, let $h, i, j \in I$ index the set of surgeries, $a \in A$ index the set of anesthesiologists, and $r \in R$ index the set of ORs. We define the following variables that are optimized:

x_{ia}: 1 if anesthesiologist a is assigned to surgery i, 0 otherwise
y_a: 1 if anesthesiologist a is assigned from on call, 0 otherwise
z_{ir}: 1 if room r is assigned to surgery i, 0 otherwise
v_r: 1 if room r is assigned any surgery, 0 otherwise
u_{ij}: 1 if surgery i precedes surgery j, 0 otherwise
α_{ija}: 1 if surgery i and j are assigned to anesthesiologist a and i precedes j, 0 otherwise
β_{ijr}: 1 if surgery i and j are assigned to room r and i precedes j, 0 otherwise
s_i: Scheduled start time of surgery i (hours)
S_i: Actual start time of surgery i (hours)
$Over_a$: Overtime of anesthesiologist a (hours)
$Over_r$: Overtime of room r (hours)

In addition, let $\mathbf{x} = (x_{ia})\ \forall\, i \in I, a \in A$, $\mathbf{y} = (y_a)\ \forall\, a \in A$, $\mathbf{z} = (z_{ir})\ \forall\, i \in I, r \in R$, $\mathbf{u} = (u_{ij})\ \forall\, i, j \in I$, $\mathbf{s} = (s_i)\ \forall\, i \in I$ denote the vectors associated with these variables. Next, we define the following parameters or inputs:

κ^A_{ia}: 1 if anesthesiologist a can be assigned to surgery i, 0 otherwise
κ^R_{ir}: 1 if surgery i can be done in room r, 0 otherwise
g_a: 1 if anesthesiologist a is on regular duty, 0 otherwise
w_a: 1 if anesthesiologist a is on call, 0 otherwise
c_r: Fixed cost of opening OR r ($/day)
c_{oa}: Overtime cost of anesthesiologist a ($/hour)
c_{or}: Overtime cost of room r ($/hour)
c_q: Cost of assigning anesthesiologist from on call ($/day)
t^{start}_a: Start time of shift associated with anesthesiologist a (hour)
t^{end}_a: End time of shift associated with anesthesiologist a (hour)
T^{end}: End time of the day (hour)
$M, M_{\text{seq}}, M_{\text{anesth}}, M_{\text{room}}$: large positive numbers

The durations d_i of surgery i is uncertain $\forall\, i \in I$ and can be considered as a random variable. The vector of surgery durations for the day is represented by $\mathbf{d} = (d_i)$, $\forall\, i \in I$. We incorporate the uncertainty in surgery durations through a robust optimization approach, where we model d_i as an uncertain parameter that takes values in $[\bar{d}_i - \hat{d}_i, \bar{d}_i + \hat{d}_i]$. Here, $\bar{d}_i$ is the nominal duration for surgery i and $\hat{d}_i$ is one-sided maximum deviation for surgery i. We define the scaled deviations of d_i about its nominal value as $f_i = (d_i - \bar{d}_i)/\hat{d}_i$. Note that the scaled deviation f_i can take a value in $[-1, 1]$. Following the approach in Bertsimas and Sim (2004), Denton et al. (2010), we subject the scaled deviations to a constraint $\sum_{i \in I} |f_i| \leqslant \tau$ so that, the total deviation across all surgeries is less than a known threshold τ. Here, τ bounds the total maximum deviation of surgery duration from the nominal value across all surgeries. This threshold is called the budget of uncertainty and represents the level of pessimism on the number of surgeries deviating from their nominal value. If $\tau = 0$, it is equivalent to solving the nominal value problem with $d_i = \bar{d}_i$, $\forall\, i \in I$.

The IARSP consists of two stages. The first-stage problem assigns anesthesiologists and rooms to surgeries, and prescribes a sequence of surgeries to be performed in each room and by each anesthesiologist. The second-stage recourse function determines actual start times to surgeries based on realized durations of preceding surgeries, and assigns overtimes to

resources such that all the surgeries are completed in the assignment and sequence prescribed by the first-stage problem. Here, the two-stage approach assumes that all information about actual surgery durations is known early in the morning, which is, of course, not the case. However, this simplification has no impact on the solution since the only recourse action is to accumulate overtime without changing the sequences. Further, this simplification is consistent with the literature on surgery scheduling employing two-stage stochastic models with recourse (Denton et al. 2010, Batun et al. 2011, Mancilla and Storer 2012). The [IARSP] can be written as

[IARSP]

$$\mathcal{V}^*(\tau) = \min\left\{\sum_{r\in R} c_r v_r + \sum_{a\in A} c_q y_a + \boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})\right\} \quad (1)$$

subject to

$$\sum_{a\in A} x_{ia} = 1 \quad \forall\, i\in I \quad (2)$$

$$\sum_{r\in R} z_{ir} = 1 \quad \forall\, i\in I \quad (3)$$

$$z_{ir} \leqslant v_r \quad \forall\, i\in I, r\in R \quad (4)$$

$$x_{ia} \leqslant g_a + y_a \quad \forall\, i\in I, a\in A \quad (5)$$

$$y_a \leqslant w_a \quad \forall\, a\in A \quad (6)$$

$$s_i \geqslant t_a^{\text{start}} - M(1-x_{ia}) \quad \forall\, i\in I, a\in A \quad (7)$$

$$x_{ia} \leqslant \kappa_{ia}^A \quad \forall\, i\in I, a\in A \quad (8)$$

$$z_{ir} \leqslant \kappa_{ir}^R \quad \forall\, i\in I, r\in R \quad (9)$$

$$\alpha_{ija} \leqslant u_{ij} \quad \forall\, i,j\in I, a\in A \quad (10)$$

$$\beta_{ijr} \leqslant u_{ij} \quad \forall\, i,j\in I, r\in R \quad (11)$$

$$u_{ij} + u_{ji} \leqslant 1 \quad \forall\, i,j\in I \quad (12)$$

$$u_{ih} \geqslant u_{ij} + u_{jh} - 1 \quad \forall\, i,j,h\in I \quad (13)$$

$$\alpha_{ija} + \alpha_{jia} \leqslant x_{ia} \quad \forall\, i,j\in I, a\in A \quad (14)$$

$$\alpha_{ija} + \alpha_{jia} \leqslant x_{ja} \quad \forall\, i,j\in I, a\in A \quad (15)$$

$$\alpha_{ija} + \alpha_{jia} \geqslant x_{ia} + x_{ja} - 1 \quad \forall\, i,j\in I, a\in A \quad (16)$$

$$\beta_{ijr} + \beta_{jir} \leqslant z_{ir} \quad \forall\, i,j\in I, r\in R \quad (17)$$

$$\beta_{ijr} + \beta_{jir} \leqslant z_{jr} \quad \forall\, i,j\in I, r\in R \quad (18)$$

$$\beta_{ijr} + \beta_{jir} \geqslant z_{ir} + z_{jr} - 1 \quad \forall\, i,j\in I, r\in R \quad (19)$$

$$\alpha_{ija} \geqslant x_{ia} + x_{ja} + \beta_{ijr} - 2 \quad \forall\, i,j\in I, r\in R, a\in A \quad (20)$$

$$\beta_{ijr} \geqslant z_{ir} + z_{jr} + \alpha_{ija} - 2 \quad \forall\, i,j\in I, r\in R, a\in A \quad (21)$$

$$x_{ia},\, y_a,\, z_{ir},\, u_{ij},\, v_r,\, \alpha_{ija},\, \beta_{ijr} \in \{0,1\} \quad \forall\, i,j\in I, r\in R, a\in A \quad (22)$$

$$s_i \geqslant 0 \quad \forall\, i\in I. \quad (23)$$

Objective function (1) consists of three terms. The first term is the fixed cost for opening ORs each day. The second term is the cost of assigning anesthesiologists from on call. The third term $\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ represents the worst-case second-stage cost and is described in detail below. Constraints (2) and (3) assign each surgery exactly one anesthesiologist and one OR, respectively. Constraint (4) ensures that v_r is set to 1 whenever any surgery is assigned to OR r. Constraint (5) ensures that an anesthesiologist can be assigned to a surgery only if they are on regular duty or on call. Constraint (6) enforces that an anesthesiologist can be assigned from on call only if they are listed in the on-call list. Constraint (7) ensures that an anesthesiologist can be assigned a surgery only if the scheduled start time of the surgery is after the shift start time of the anesthesiologist. Constraints (8) and (9) ensure that surgeries are assigned rooms and anesthesiologists by specialty. Constraint (10) enforces the condition that if an anesthesiologist is used to conduct surgery i before surgery j, then surgery i has to precede surgery j or u_{ij} is set to 1. Constraint (11) imposes the similar condition and sets u_{ij} to 1 when surgery i precedes surgery j in an OR. Constraint (12) ensures that only one of u_{ij} or u_{ji} can be 1. Constraint (13) is required to maintain consistency of schedule between any three surgeries that follow each other, so that if i precedes j and j precedes h, then i should precede h. Constraints (14)–(15) restrict that only one of α_{ija} and α_{jia} can be 1 only if surgeries i and j are assigned to anesthesiologist a. Constraint (16) enforces either α_{ija} or α_{jia} is set to 1 if surgeries i and j are assigned to anesthesiologist a. In addition, constraint (16) ensures that the sequencing constraints for anesthesiologists α_{ija} is active only for those surgeries that are assigned to the same anesthesiologist. Constrains (17)–(19) are similar logical constraints corresponding to the sequencing of rooms. Constraints (20) and (21) maintain consistency of sequencing variables between OPs and anesthesiologists. Constraint (20) enforces that if anesthesiologist a and OR r is assigned to surgeries i and j and i precedes j in OR r, then i has to precede j in assignment to anesthesiologist a. Constraint (21) is a similar constraint that makes sure that if surgery i precedes surgery j with anesthesiologist a, then i has to precede j in the assignment of OR r. Constraints (22) and (23) represent variable domains.

The worst-case second-stage cost is given by

$$\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s}) = \max_{\mathbf{d}\in\mathcal{D}(\tau)} \boldsymbol{\mathcal{R}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{d}) \quad (24)$$

$$\boldsymbol{\mathcal{D}}(\tau) = \left\{\mathbf{d}\in\mathbb{R}^{|I|}\colon d_i = \bar{d}_i + f_i\hat{d}_i, i\in I, \mathbf{f}\in\boldsymbol{\mathcal{F}}(\tau)\right\} \quad (25)$$

$$\boldsymbol{\mathcal{F}}(\tau) = \left\{\mathbf{f}\in\mathbb{R}^{|I|}\colon \sum_{i\in I}|f_i| \leqslant \tau, -1 \leqslant f_i \leqslant 1\right\} \quad (26)$$

$\boldsymbol{\mathcal{R}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{d})$ is the total overtime cost across all resources, for a given assignment, sequence, scheduled surgery start times, and surgery durations. This is maximized over the vector of surgery durations $\mathbf{d}$ to determine the worst-case cost, where $\mathbf{d}$ is restricted to lie in the uncertainty set $\mathcal{D}(\tau)$ given by (25). This

equation restricts d_i, the duration of surgery i, to lie within a maximum deviation of $\hat{d}_i$ from the nominal value of the duration $\bar{d}_i$. The total extent of such deviations is specified by the set $\mathcal{F}(\tau)$, which is defined by (26) and is well suited to our problem context. In particular, the effective allocation of multiple parallel resources such as anesthesiologists and rooms, which are used repeatedly across the surgeries in a given day requires a specification of τ, an overall level or budget of uncertainty across surgical durations. This is enforced by (26), which specifies that the maximum deviation across all surgeries is at most τ. A schedule based on a large τ would be overly accommodating toward the second-stage cost, while a schedule corresponding to a small τ would not be accommodating enough. In Section 4, we present a methodology to determine $\bar{d}_i$, $\hat{d}_i$, and τ based on historical data.

In determining $\mathcal{R}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{d})$, it is important to note that the only decision variables at this stage are the actual start times of the surgeries and the overtime for the anesthesiologists and rooms. We pick these variables to minimize total overtime costs while ensuring that all the surgeries scheduled for the day are completed and there are no conflict in actual start times of surgeries assigned to the same resource. To compute $\mathcal{R}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{d})$, we formulate the linear program as follows:

$$\mathcal{R}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{d}) = \min\left\{\sum_{a\in A} c_{oa} Over_a + \sum_{r\in R} c_{or} Over_r\right\} \quad (27)$$

subject to

$$S_j \geqslant S_i + d_i - M_{\text{seq}}(1-u_{ij}) \quad \forall\, i,j \in I \quad (28)$$

$$S_i \geqslant s_i \quad \forall\, i \in I \quad (29)$$

$$Over_a \geqslant S_i + d_i - t_a^{\text{end}} - M_{anesth}(1 - x_{ia} + y_a) \quad \forall\, i \in I,\, a \in A \quad (30)$$

$$Over_r \geqslant S_i + d_i - T^{\text{end}} - M_{\text{room}}(1 - z_{ir}) \quad \forall\, i \in I,\, r \in R \quad (31)$$

$$S_i,\, Over_a,\, Over_r \geqslant 0 \quad \forall\, i \in I,\, a \in A,\, r \in R. \quad (32)$$

The objective function consists of the sum of overtime across all the resources. Constraint (28) ensures that the start time of the succeeding surgery is only after the end time of the preceding surgery. Constraint (29) ensures that the actual start time of the surgery can be no earlier than the scheduled start time. Constraints (30) and (31) define the overtime for anesthesiologists on regular duty and ORs, respectively, which is the time difference between the end time of the last surgery in that shift and the regular shift end time for the resource. Constraint (32) restricts start time and overtime variables to be nonnegative variables. The [IARSP] is a robust optimization model with the recourse represented by this linear program. We next develop some structural properties that are useful in constructing solution techniques for this model.

Proposition 1. *The* [IARSP] *has relatively complete recourse.*

All proofs are provided in the e-companion. Proposition 1 implies that for every feasible first-stage solution, there exists a feasible second-stage solution. This proposition allows us to evaluate second-stage costs for every feasible first-stage solution. This is important for the solution method for the [IARSP] described in Section 3.1.1. However, evaluating $\mathcal{Q}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ for any given first-stage solution requires one to solve the problem given in Equations (24)–(32), which is not easy due to the max-min operator in its objective. Let λ_{ij}, ϕ_i, μ_{ia}, $\theta_{i,r}$, $i,j \in I, a \in A, r \in R$ be dual variables corresponding to constraints (28)–(31), respectively. Further, define $\pi_i = \sum_{j\in I-\{i\}} \lambda_{ij} + \sum_{a\in A} \mu_{ia} + \sum_{r\in R} \theta_{ir}$, and $\xi_i = f_i \pi_i$. The following proposition simplifies the computation of $\mathcal{Q}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$, and, consequently, the [IARSP].

Proposition 2. *If parameter τ is chosen to be a positive integer, then $\mathcal{Q}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ can be reformulated as the following mixed-integer program (MIP):*

$$\begin{aligned}\mathcal{Q}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s}) = \max\Bigg\{ & \sum_{i\in I}(\bar{d}_i \pi_i + \xi_i \hat{d}_i) + \sum_{i\in I} s_i \phi_i - M_{\text{seq}} \sum_{\substack{i,j\in I \\ i\neq j}} \lambda_{ij}(1-u_{ij}) \\ & - M_{anesth} \sum_{\substack{i\in I \\ a\in A}} \mu_{ia}(1 - x_{ia} + y_a) - M_{\text{room}} \sum_{\substack{i\in I \\ r\in R}} \theta_{ir}(1 - z_{ir}) \\ & - \sum_{\substack{i\in I \\ r\in R}} \theta_{ir} T^{\text{end}} - \sum_{\substack{i\in I \\ a\in A}} \mu_{ia} t_a^{\text{end}} \Bigg\}\end{aligned}$$

subject to

$$\sum_{i\in I} \mu_{ia} \leqslant c_{oa} \quad \forall\, a \in A \quad (33)$$

$$\sum_{i\in I} \theta_{ir} \leqslant c_{or} \quad \forall\, r \in R \quad (34)$$

$$\sum_{\substack{j\in I \\ j\neq i}} \lambda_{ij} - \sum_{\substack{j\in I \\ j\neq i}} \lambda_{ji} + \sum_{a\in A} \mu_{ia} + \sum_{r\in R} \theta_{ir} - \phi_i \geqslant 0 \quad \forall\, i \in I \quad (35)$$

$$\sum_{j\in I-\{i\}} \lambda_{ij} + \sum_{a\in A} \mu_{ia} + \sum_{r\in R} \theta_{ir} = \pi_i \quad \forall\, i \in I \quad (36)$$

$$\sum_{i\in I} f_i \leqslant \tau \quad (37)$$

$$\xi_i \leqslant M_f f_i \quad \forall\, i \in I \quad (38)$$

$$\xi_i \leqslant \pi_i \quad \forall\, i \in I \quad (39)$$

$$\xi_i,\, \pi_i,\, \lambda_{ij},\, \theta_{ir},\, \mu_{ia},\, \phi_i \geqslant 0 \quad \forall\, i,j \in I,\, a \in A,\, r \in R \quad (40)$$

$$f_i \in \{0,1\} \quad \forall\, i \in I. \quad (41)$$

As described in the e-companion, the proof of this proposition follows from strong duality of the second-stage recourse problem for a given first-stage solution. Note that as a consequence of Proposition 2 in which τ is set to an integer, f_i, $i \in I$ are also now binary variables. Thus, in the worst case, surgeries are set to either their nominal value or maximum positive deviation.

In effect, τ can now be interpreted as the upper bound on the number of surgeries that reach their maximum deviation. Thus, restricting τ to be a positive integer, as in this proposition, allows for a more natural interpretation of τ, which was important in the application context. As discussed in Section 4, this interpretation drives our data-driven method in setting a parametric value for τ from historical data. Propositions 1 and 2 imply that $\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ can be evaluated for any given first-stage feasible solution by solving a MIP. In particular, let $(\boldsymbol{\pi}^l,\boldsymbol{\xi}^l,\boldsymbol{\lambda}^l,\boldsymbol{\phi}^l,\boldsymbol{\mu}^l,\boldsymbol{\theta}^l)$ be the solution to $\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ for some given $(\mathbf{x}^l,\mathbf{y}^l,\mathbf{z}^l,\mathbf{u}^l,\mathbf{s}^l)$. Then, the following propositions provide a lower bound and characterize the structure of $\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$. They will be used in the solution method provided in Section 3.1.1.

Proposition 3. *A lower bound on* $\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ *is provided by* $\sum_{i\in I}(\bar{d}_i\pi_i^l+\xi_i^l\hat{d}_i)+\sum_{i\in I}s_i\phi_i^l-M_{\text{seq}}\sum_{i,\,j\in I,\,i\neq j}\lambda_{ij}^l\cdot(1-u_{ij})-M_{\text{anesth}}\sum_{i\in I,a\in A}\mu_{ia}^l(1-x_{ia}+y_a)-M_{\text{room}}\cdot\sum_{i\in I,r\in R}\theta_{ir}^l(1-z_{ir})-\sum_{i\in I,r\in R}\theta_{ir}^lT^{\text{end}}-\sum_{i\in I,a\in A}\mu_{ia}^lt_a^{\text{end}}$.

Proposition 4. $\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ *is a piecewise-linear convex function in the first-stage decision variables* $\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s}$.

In light of Proposition 4, the [IARSP] now reduces to a piecewise-linear convex MIP in which $\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$, the convex part of the objective function can be evaluated by using Proposition 2 and solving a MIP. However, given this nonlinearity and the large number of integer variables in our application, the [IARSP] cannot be solved using powerful solvers for nonlinear programs such as BARON (Sahinidis 2014) and DICOPT (Viswanathan and Grossmann 1990). Consequently, we develop the following model-based heuristic procedure to solve this problem.

3.1. Solution Methods

We start by describing the model-based heuristic. We then discuss SAA-based techniques that are commonly used in literature, which we use to benchmark the model-based and practitioner's heuristics. The performance of these methods along with practitioner's heuristc (described in Section 2) will be discussed in Section 5.

3.1.1. Model-Based Heuristic.

This heuristic is based on Kelley's algorithm (Kelley 1960) as described in Thiele et al. (2009) to solve robust optimization problems with recourse. Here, we consider the [IARSP] and in light of Proposition 4, approximate $\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ by a piecewise-linear equation via successive linear cuts. We then use this approximation in constructing the master problem at the kth iteration of the heuristic, MP(k), defined as

$$[\text{MP}(k)]\quad \min_{v_r,\,y_a,\,\psi}\left\{\sum_{r\in R}c_rv_r+\sum_{a\in A}c_qy_a+\psi\right\} \tag{42}$$

subject to

$$(2)\text{–}(23)$$

$$\psi\geqslant\sum_{i\in I}(\bar{d}_i\pi_i^l+\xi_i\hat{d}_i)+\sum_{i\in I}s_i\phi_i^l-M_{\text{seq}}\sum_{\substack{i,\,j\in I\\ i\neq j}}\lambda_{ij}^l(1-u_{ij})-M_{anesth}\sum_{\substack{i\in I\\ a\in A}}\mu_{ia}^l(1-x_{ia}+y_a)-M_{\text{room}}\sum_{\substack{i\in I\\ r\in R}}\theta_{ir}^l(1-z_{ir})-\sum_{\substack{i\in I\\ r\in R}}\theta_{ir}^lT^{\text{end}}-\sum_{\substack{i\in I\\ a\in A}}\mu_{ia}^lt_a^{\text{end}}\qquad l=0,1,2,\ldots,k-1 \tag{43}$$

$$\psi\geqslant 0. \tag{44}$$

Observe that in this problem, we approximate the value of $\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ by a variable $\psi\geqslant 0$. To improve this approximation, in each iteration of the heuristic, we use (43) to enforce the condition that ψ is greater than or equal to the lower bound of $\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$, as established by Proposition 3. This results in constraints (43) in which ψ approximates $\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ by a piecewise-linear equation via successive linear cuts. This problem will be used in the model-based heuristic formalized by the following algorithm.

Algorithm (Model-Based Heuristic)

Step 1. Initialize $U\leftarrow\infty$, $L\leftarrow 0$, $k\leftarrow 0$, $l\leftarrow 0$. Set $\epsilon>0$ to be sufficiently small.

Step 2. Solve the [MP(k)] and let the solution be $\tilde{x}_{ia}^k$, $\tilde{y}_a^k$, $\tilde{z}_{ir}^k$, $\tilde{u}_{ij}^k$, $\tilde{s}_i^k$, $\tilde{v}_r^k$, $\tilde{\alpha}_{ija}^k$, $\tilde{\beta}_{ijr}^k$, $\tilde{\psi}^k$. Set $L\leftarrow\sum_{r\in R}c_r\tilde{v}_r^k+\sum_{a\in A}c_q\tilde{y}_a^k+\tilde{\psi}^k$.

Step 3. Compute $\boldsymbol{\mathcal{Q}}(\tilde{\mathbf{x}}^k,\tilde{\mathbf{y}}^k,\tilde{\mathbf{z}}^k,\tilde{\mathbf{u}}^k,\tilde{\mathbf{s}}^k)$ for given $\tilde{x}_{ia}^k$, $\tilde{v}_r^k$, $\tilde{z}_{ir}^k$, $\tilde{u}_{ij}^k$, $\tilde{s}_i^k$ obtained in **Step 2** by solving the mixed-integer programming formulation given in Proposition 2. Let the optimal solution be ξ_i^k, π_i^k, λ_{ij}^k, θ_{ir}^k, μ_{ia}^k, ϕ_i^k. $U\leftarrow\min\{U,\sum_{r\in R}c_r\tilde{v}_r^k+\sum_{a\in A}c_q\tilde{y}_a^k+\boldsymbol{\mathcal{Q}}(\tilde{\mathbf{x}}^k,\tilde{\mathbf{y}}^k,\tilde{\mathbf{z}}^k,\tilde{\mathbf{u}}^k,\tilde{\mathbf{s}}^k)\}$.

Step 4. If $U-L<\epsilon$, go to **Step 6**, **else** go to **Step 5**.

Step 5. $k\leftarrow k+1$. Add constraint $\psi\geqslant\sum_{i\in I}(\bar{d}_i\pi_i^k+\xi_i\hat{d}_i)+\sum_{i\in I}s_i\phi_i^k-M_{\text{seq}}\sum_{i,\,j\in I,\,i\neq j}\lambda_{ij}^k(1-u_{ij})-M_{\text{anesth}}\cdot\sum_{i\in I,a\in A}\mu_{ia}^k(1-x_{ia}+y_a)-M_{\text{room}}\sum_{i\in I,r\in R}\theta_{ir}^k(1-z_{ir})-\sum_{i\in I,r\in R}\theta_{ir}^kT^{\text{end}}-\sum_{i\in I,a\in A}\mu_{ia}^kt_a^{\text{end}}$ to the MP(k). Go to **Step 2**.

Step 6. $\tilde{x}_{ia}^k$, $\tilde{y}_a^k$, $\tilde{z}_{ir}^k$, $\tilde{u}_{ij}^k$, $\tilde{v}_r^k$, $\tilde{s}_i^k$, $\tilde{\alpha}_{ija}^k$, $\tilde{\beta}_{ijr}^k$ is the heuristic solution to [IARSP].

The above algorithm is well suited for the [IARSP] as from Proposition 4, its objective is piecewise-linear convex and cutting plane methods such as those used in Step 5 of this algorithm are finitely convergent for piecewise-linear functions (Ruszczyński 2006). Also, in this algorithm, in early iterations, the solution in Step 2 is obtained by employing the *user callbacks* feature of the solver used to solve MP(k). Here, instead of solving this problem to optimality in the initial iterations, we request the solver to return a feasible solution, which is then used to apply cuts in Step 5 and approximate the convex function $\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ at

each corresponding feasible solution. We do so because MP(k) is a problem with a large number of integer variables, and solving it to optimality can be computationally expensive with poor returns at early iterations when $\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ has not been approximated well enough by constraints (43). Results on the computational performance and the time required for this heuristic is provided in Section 5.

3.1.2. Benchmark Heuristic. Here, we use SAA-based methods similar to those provided in Denton et al. (2007) to benchmark the model-based heuristic. In SAA, instead of using the worst-case formulation, we solve with expected second-stage costs. The expectation is based on scenarios drawn from an estimated distribution of surgeries. The resultant two-stage stochastic optimization problem is solved by the *L-shaped* method (Birge and Louveaux 1988). We describe the model formulation of the SAA version and provide details on the estimation of the distribution in the e-companion. However, since the sample average-based method was unable to solve large-scale problems such as those found in the application, we use this method on smaller problems constructed from real data. We provide the results of the comparison between the SAA-based method and the robust optimization-based method in Section 5.

4. Parameter Estimation and Model Calibration

In this section, we use historical data of surgery durations to choose the uncertainty sets $\boldsymbol{\mathcal{D}}(\tau)$ and $\boldsymbol{\mathcal{F}}(\tau)$. The performance of robust optimization depends closely on the definition of these uncertainty sets. If the optimal values of d_i, $i \in I$ in the inner maximization problem ($\max_{\mathbf{d}\in\mathcal{D}(\tau)} \boldsymbol{\mathcal{R}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{d})$) are significantly larger than the corresponding $\bar{d}_i$, the resulting first-stage problem will be overly pessimistic toward the realization of surgery durations. This may lead to higher first-stage costs. Conversely, if the optimal values of d_i, $i \in I$ are too close to the corresponding $\bar{d}_i$, the uncertainty sets would not cover many cases of future realizations in which the surgery durations deviate significantly from the nominal value. This could result in higher second-stage costs. Thus we need to look at the combined first- and second-stage costs while designing the uncertainty sets. Designing the uncertainty sets involves setting the following parameters: the nominal surgery duration $\bar{d}_i$, $i \in I$, the maximum deviation $\hat{d}_i$, $i \in I$, and the robust optimization parameter τ.

There have been several approaches suggested for designing uncertainty sets. Ben-Tal et al. (2009) provide the theoretical background for deciding good uncertainty sets. Denton et al. (2010) use the 10th and 90th percentile width of historical surgery durations as the width $[\bar{d}_i - \hat{d}_i, \bar{d}_i + \hat{d}_i]$ in an OR assignment application. Subsequently, they perform sensitivity analysis and calibrate their model to an equivalent SAA-based solution to decide the robust optimization parameter τ. Bertsimas et al. (2013) propose using statistical hypothesis tests to construct uncertainty sets. Denton et al. (2010) and Bertsimas et al. (2013) model the uncertainty sets based on historically observed values of a *single* uncertain parameter. In our application with a wide variety of surgery specialties with considerable variability in surgery durations across specialties, a percentile width not conditional on surgery characteristics would be unnecessarily wide, leading to an overly pessimistic uncertainty set. Therefore we incorporate these characteristics and propose a joint estimation and calibration procedure to design the uncertainty set. Our procedure provides tight uncertainty sets that take into account observable surgery characteristics while making no assumptions on the probability distribution of surgery durations. We further calibrate the uncertainty set by evaluating the performance of the robust solution to empirical realizations.

There were two data sets available to us. The first data set $\Delta^E = \{\tilde{d}^{(m)}, \tilde{\mathbf{b}}^{(m)}\}_{m=1}^{M}$ consists of $M = 25{,}700$ samples of $\tilde{d}^{(m)}$ corresponding to the historical realization of durations of surgery m and $\tilde{\mathbf{b}}^{(m)}$, which represents the observed characteristics of surgery m. Table 2 provides details on the surgery characteristics included in $\tilde{\mathbf{b}}(m)$, $\forall m$, and the variable names used for the subsequent regression. The second data set Δ^C was partitioned into disjunctive training and testing sets, $\Delta^{\text{C-Train}} = \{\tilde{\mathbf{d}}^{(n)}, \tilde{\mathbf{b}}^{(n)}\}_{n=1}^{N_1}$ and $\Delta^{\text{C-Testing}} = \{\tilde{\mathbf{d}}^{(n)}, \tilde{\mathbf{b}}^{(n)}\}_{n=1}^{N_2}$. For these data sets, $N_1 = 120$ days and $N_2 = 60$ days. Both these data sets consist of $\tilde{\mathbf{d}}^{(n)}$ representing the vector of realized durations and $\tilde{\mathbf{b}}^{(n)}$ denoting the vector of surgery characteristics of all surgeries performed on day n.

We use Δ^E and $\Delta^{\text{C-Train}}$ to estimate $\boldsymbol{\mathcal{D}}(\tau)$ and $\boldsymbol{\mathcal{F}}(\tau)$. Most of the research in estimating uncertainty sets is for single-stage problems when feasibility is not guaranteed. Since we have two stages and second-stage feasibility is guaranteed by Proposition 1, we develop the following procedure that comprises of an estimation and a calibration step.

Step 1. Estimation

First, for a given parameter $\rho \in (0,1)$, we define conditional quantile functions $g_L(\mathbf{b};\rho)$ and $g_U(\mathbf{b};\rho)$ such that

$$\mathbf{P}[\tilde{d} \leqslant g_L(\tilde{\mathbf{b}};\rho)] = \frac{1-\rho}{2}, \quad \text{and} \quad \mathbf{P}[\tilde{d} \geqslant g_U(\tilde{\mathbf{b}};\rho)] = \frac{1-\rho}{2}.$$

Thus, given observed surgery characteristics $\tilde{\mathbf{b}}$, the future realization $\tilde{d}$ will lie in the set $[g_L(\tilde{\mathbf{b}};\rho), g_U(\tilde{\mathbf{b}};\rho)]$ with probability ρ. The true quantile functions are not known to us, therefore we obtain estimates of the quantile functions $\hat{g}_L(\tilde{\mathbf{b}};\rho)$ and $\hat{g}_U(\tilde{\mathbf{b}};\rho)$ through a conditional quantile regression method (Koenker 2005)

Table 2. Surgery Characteristics Provided in Data Sets

Surgery characteristics	Description	Variable name
Realized surgery duration	In hours	ACTUALHRS
Surgeon's estimate of surgery duration	In hours	BOOKEDHRS
Patient class	Inpatient, outpatient, or same-day admit	PATCLASS
Booked current procedural terminology (CPT) code	Medical code maintained by American Medical Association defines the services to be performed during surgery. A surgery may have multiple CPT codes. The surgeon provides a list of services that maybe performed as a part of the surgery. The realized CPT codes may and often do vary from the booked CPT code. Surgeries in our data set covered 2,700 unique CPT codes.	CPT
ASA score	A system for assessing fitness of patients before surgery, higher number signifies a less fit patient. Takes integer values between 1 and 6.	ASA
Patient age	In years	AGE
Surgery service	Cardiac surgery, neuro surgery, etc., full list as in Table EC.1	SERVICE
Surgeon's name	Names of 493 surgeons, unique surgeons (providers) who have performed surgeries in the period over which data was available	PROV
Number of CPT codes	Number of CPT codes associated with procedure	NUMCPT

applied on the data set Δ^E. The use of conditional quantile regression for estimating uncertainty sets has been recently proposed by Tulabandhula and Rudin (2014). Quantile regression estimates the quantiles of the response variable (i.e., the surgery durations), given certain values of the predictor variables. Quantile regression has several advantages over the commonly used ordinary least squares (OLS) regression. First, this approach suits our application better since our objective is to find upper and lower bounds on surgery durations, such that future realization would lie within this bound with a given probability. Conditional quantiles provide these bounds without making any assumption on the probability distribution of surgery durations. Second, quantile regression is more robust to outliers; and third, it does not assume the dispersion of the response variable to be independent of the predictor variables.

We perform quantile regression using the `quantreg` package available in `R` (Koenker 2013). The response variable is the realized surgery duration. The possible set of predictors are the surgeon's estimate of surgery duration, the fitness level of the patient prior to the surgery measured by the American Society of Anesthesiologist (ASA) score,[2] the age of the patient, whether the patient is an inpatient or outpatient, the specialty of the surgery, the surgeons name, the services provided indicated by the type of the Current Procedure Terminology (CPT)[3] codes used, and the number of CPT codes used by the surgeons. For the CPT codes and surgeon's names, we cluster the variables via a k-means clustering similar to He et al. (2012). This clustering was done to account for the large number of factors in these variables and to avoid overspecification of the model. The details of the clustering procedure followed is provided in the e-companion.

The selection of variables was done comparing the Akaike information Crietria (AIC) and the mean square prediction error (MSPE). The results of these tests are provided in the e-companion. We also checked for collinearity using variance inflation factors (VIF) following the criteria discussed in Hair et al. (2006, pp. 191–193). Highly collinear variables (i.e., with VIF $\geqslant 10$) were removed. For example, we found that the specialty of surgery had a VIF of 40.1 as it was collinear with CPT codes, as these codes were specific to a specialty. On performing these tests, we found that the ASA score, the surgeon's estimate of duration, patient class (inpatient, outpatient or same-day admit), clustered variables corresponding to surgeons, and CPT codes were significant. The ASA score is a strong indicator of increasing complexity of the surgical procedure since it is an indicator of the level of fitness of the patient before coming into the surgery. A patient with an ASA score of 1, implying the patient is a healthy person would be expected to demonstrate less complications during surgery, while a patient with an ASA score of 3 (with severe systemic disease) would be expected to have more complications during surgery. We found that the coefficient of ASA score was -0.023 at the 0.1 quantile and 0.122 at the 0.9 quantile. Thus, every increase in ASA score contributes to approximately 7.3 minutes (≈ 0.122 hours) of additional surgery time at the 0.9 quantile level, while the effect of increment in ASA score is negligible for very short surgeries. This is intuitive as the negative effects of patient fitness would be significant for longer surgeries and would not be as impactful for shorter surgeries. As expected, the surgeon's estimate of surgery duration would be strongly correlated with the actual duration, and would explain variance not captured by other variables, since there are several factors that the surgeon is aware of that are not captured by other available data. However, as explained previously, there is some error in surgeon's estimates

as well. We found that surgeon's estimate was, on average, 12 minutes higher than actual surgery durations. Also, the coefficient of surgeons' estimates in the quantile regression model was smaller for shorter surgeries than for longer surgeries. This is because surgeons tend to be more accurate in their estimates for longer surgeries than for shorter surgeries. One possible explanation is that it was observed that surgeons tend to round to the nearest quarter of an hour while providing their estimates. This leads to an error, which is more pronounced for shorter surgeries than for longer surgeries. We also found that clusters of surgeons are significant because, as described in the e-companion, these in effect represented the experience level of surgeons. Finally, as anticipated, the type of surgery itself with its associated CPT code affected surgical durations. However, the number of CPT codes was not significant as they could be associated with relatively simpler subprocedures common to all surgeries. Similarly, the age of the patient was not significant as it was captured in the surgeons estimate of duration.

Once we have obtained the estimated conditional quantile functions, for each surgery $i \in I$, we set $\bar{d}_i + \hat{d}_i = \hat{g}_U(\tilde{\mathbf{b}}_i; \rho)$, $\bar{d}_i - \hat{d}_i = \hat{g}_L(\tilde{\mathbf{b}}_i; \rho)$. This gives

$$\bar{d}_i = \frac{\hat{g}_U(\tilde{\mathbf{b}}_i; \rho) + \hat{g}_L(\tilde{\mathbf{b}}_i; \rho)}{2} \quad \text{and} \quad \hat{d}_i = \frac{\hat{g}_U(\tilde{\mathbf{b}}_i; \rho) - \hat{g}_L(\tilde{\mathbf{b}}_i; \rho)}{2}.$$

Define $\tau' \in [0,1]$ so that $\tau = \lfloor \tau' |I| \rfloor$. Here, τ' represents the fraction of total surgeries in a given day $|I|$, which have reached their maximum duration. We then substitute the above equations in (25) and (26) for observed surgery characteristics vector $\mathbf{b}_i$ and *given* parameters $\rho \in [0,1]$ and $\tau' \in (0,1)$. Then, the uncertainty sets are given by

$$\mathcal{D}(\tau) = \mathcal{D}(\rho, \tau') = \left\{ \mathbf{d} \in \mathbb{R}^{|I|} : d_i = \frac{\hat{g}_U(\tilde{\mathbf{b}}_i; \rho) + \hat{g}_L(\tilde{\mathbf{b}}_i; \rho)}{2} + f_i \left[\frac{\hat{g}_U(\tilde{\mathbf{b}}_i; \rho) - \hat{g}_L(\tilde{\mathbf{b}}_i; \rho)}{2} \right], \ i \in I, \ \mathbf{f} \in \mathcal{F}(\tau') \right\} \quad (45)$$

$$\mathcal{F}(\tau) = \mathcal{F}(\tau') = \left\{ \mathbf{f} \in \mathbb{R}^{|I|} : \sum_{i \in I} |f_i| \leqslant \lfloor \tau' |I| \rfloor, \ -1 \leqslant f_i \leqslant 1 \right\}. \quad (46)$$

Step 2. Calibration

If we had full information on the surgery durations (i.e., if they were observable ex ante), and a deterministic solution could be executed, the resulting cost obtained when there is full information would be a lower bound to any heuristic solution. In stochastic programming, this is referred to as the *wait-and-see* solution. The full information cost on day n is given as

$$\mathcal{W}^{\mathrm{FI}(n)} = \min \left\{ \sum_{r \in R} c_r v_r + \sum_{a \in A} c_q y_a + \sum_{a \in A} c_{oa} Over_a + \sum_{r \in R} c_{or} Over_r \right\} \quad (47)$$

$$\text{subject to} \quad (2)\text{–}(23) \quad (28)\text{–}(32).$$

We solve $\mathcal{W}^{\mathrm{FI}(n)}$ for each day in $\Delta^{\text{C-Train}}$ with $\mathbf{d} = \tilde{\mathbf{d}}^{(n)}$.

The first-stage variables for day n obtained by solving [IARSP] for day n using the model-based heuristic are $(\mathbf{x}^{*(n)}, \mathbf{y}^{*(n)}, \mathbf{z}^{*(n)}, \mathbf{u}^{*(n)}, \mathbf{s}^{*(n)})$. The cost of the model-based heuristic under a realized duration vector $\tilde{\mathbf{d}}^{(n)}$ is defined as

$$\mathcal{W}(\mathbf{x}^{*(n)}, \mathbf{y}^{*(n)}, \mathbf{z}^{*(n)}, \mathbf{u}^{*(n)}, \mathbf{s}^{*(n)}; \rho, \tau', \tilde{\mathbf{d}}^{(n)}) = \sum_{r \in R} c_r v_r + \sum_{a \in A} c_q y_a + \mathcal{R}(\mathbf{x}^{*(n)}, \mathbf{y}^{*(n)}, \mathbf{z}^{*(n)}, \mathbf{u}^{*(n)}, \mathbf{s}^{*(n)}, \tilde{\mathbf{d}}^{(n)}).$$

This represents the cost that would be realized at the end of day n if the model-based heuristic was implemented with uncertainty set $\mathcal{D}(\rho, \tau')$. The average performance of the model-based heuristic across N samples relative to the full information case is defined as follows:

$$\bar{\mathcal{W}}(\rho, \tau') = \frac{1}{N_1} \sum_{n=1}^{N_1} \frac{[\mathcal{W}(\mathbf{x}^{*(n)}, \mathbf{y}^{*(n)}, \mathbf{z}^{*(n)}, \mathbf{u}^{*(n)}, \mathbf{s}^{*(n)}; \rho, \tau', \tilde{\mathbf{d}}^{(n)}) - \mathcal{W}^{\mathrm{FI}(n)}]}{\mathcal{W}^{\mathrm{FI}(n)}}. \quad (48)$$

We calculate $\bar{\mathcal{W}}(\rho, \tau')$ for several values of $\rho \in (0,1)$ and $\tau' \in [0,1]$ and choose the pair that minimizes $\bar{\mathcal{W}}(\rho, \tau')$. This is summarized in Table 3.

From Table 3, we can see $\rho = 0.95$ and $\tau' = 0.2$ is optimal. This implies that at a 95% confidence level, we can set 20% of all surgeries to its maximum durations on any given day when we define the uncertainty sets $\mathcal{D}(\tau)$ and $\mathcal{F}(\tau)$ and solve the [IARSP]. At this value, the model-based heuristic solution was 24% more than the full information solution.

We used these values of ρ and τ' to evaluate the performance of the model-based heuristic relative to the full information case as defined in (48) for the testing data set $\Delta^{\text{C-Testing}}$. Here, we found that the model-based heuristic was 28% more than the full information solution. Thus the out-of-sample performance, using $\Delta^{\text{C-Testing}}$ was close to the in-sample performance using $\Delta^{\text{C-Train}}$. This provides validation to use these values of ρ and τ' in the computational analysis described next.

Table 3. Performance of Model-Based Heuristic ($\bar{\mathcal{W}}(\rho, \tau')$) across Budget of Uncertainty Parameter ($\tau = \lfloor \tau' |I| \rfloor$) and Conditional Quantile Parameter (ρ)

	τ'			
ρ	0.1	0.2	0.3	0.4
0.80	1.52	1.43	1.48	1.57
0.85	1.43	1.35	1.45	1.59
0.90	1.41	1.27	1.44	1.59
0.95	1.38	1.24	1.40	1.55
0.98	1.38	1.29	1.42	1.57

5. Computational Analysis

In this section, we conduct a computational analysis to evaluate our approach. To perform this analysis, we used data provided by the UCLA RRMC on all surgeries conducted in their OR suite over a 14-month period. This analysis was essential to provide confidence in our method. Our computational analysis is divided into two sections. In Section 5.1, we evaluate the performance of the practitioner's heuristic and the procedures described in Section 3.1. In Section 5.2 we compare the performance of the model-based heuristic with the actual resource assignment and scheduling decisions made at this hospital and estimate the cost savings.

5.1. Performance Evaluation

The size and scope of the scheduling activities during this time period demonstrated considerable variation as shown in Table 4. To ensure that our computational analysis captured this range of variation, we constructed five problems of varying sizes as shown in Table 5. For each of these five sets of problem instances, we considered different values for c_q, c_{or} and c_{oa}. The actual value of c_q, c_{oa}, and c_{or} at UCLA RRMC were \$1,000 per day, \$150 per hour, and \$450 per hour, respectively. In addition to these actual values, we considered values where we scaled one of these costs by a factor of 2 or 1/2 while keeping the other two at the current value. This led to 7 possible combinations of costs for each of the 5 problem instances and a total of $7 \times 5 = 35$ possible problems. We tried to solve the IARSP for these data sets using the leading commercial solver for stochastic programs such as `ddsip` (Märkert and Gollmer 2008). However, other than the smaller problem instances A and B, these solvers could not even generate feasible solutions after more than 24 hours of computation, and the runs were aborted. This provides validation for developing the model-based heuristic to solve the [IARSP].

The heuristic procedures were coded in Python programming language (van Rossum 2001). The computational analyses were run on a workstation with 3.8 GHz AMD A10 processor, 8 GB of RAM, and Linux Mint as the operating system. For the MIP subroutine calls, we used Gurobi 5.63 (Gurobi Optimization 2015) called from Python via the Gurobi Python Interface. In all the computations using the model-based heuristic, we set the gap $\epsilon = 5\%$. Thus all the solutions of the model-based heuristic were within a 5% gap from the lower bound and were solved within 25 minutes.

Table 5. Problems Used for Performance Analysis

Instance	Number of surgeries, $\|I\|$	Number of rooms, $\|R\|$	Number of anesthesiologists, $\|A\|$
A	10	3	5
B	15	5	8
C	25	7	10
D	40	10	25
E	65	23	40

Tables 6 and 7 summarize the results obtained for the computational analysis. In Table 6, the performance of the model-based heuristic and the practitioner's heuristic procedure is compared with the cost of the SAA-based solution for small-scale problems. This table shows that these procedures are all very close to the SAA method and this does not change with changes in the cost parameters. In Table 7, we consider the more realistic medium- and large-scale problems. Since SAA is unable to solve these problems, we provide the performance of the model-based heuristic with respect to the practitioner's heuristic. From Table 7, we note that for these problems, the model-based heuristic provides significant cost reductions over the practitioner's heuristic. In particular, the percentage cost reduction for these problems ranged from 2.26% to 7.56% averaging around 4.95%.

We can also observe from Table 7 that the gains of the model-based heuristic over the practitioner's heuristic improves as the size of the problem increases. This is because for small-sized problems, there are limited options and it is more likely the practitioner's heuristic achieves a solution that is close to optimal. Further, since in small-sized problems $|I|$ is low, the number of surgeries that reach its worst-case duration (i.e., $\tau = \tau'|I|$) is also low. In these circumstances, the solution of the model-based heuristic and the practitioner's heuristic are similar and close to the nominal value solution, where surgical durations are set to its nominal duration estimates. However, as the problem size increases, the number of surgeries reaching

Table 4. Data Sets for Performance Analysis

	Number of surgeries conducted per day		Number of anesthesiologists working per day		Number of ORs functioning per day	
	Weekdays	Weekends	Weekdays	Weekends	Weekdays	Weekends
Minimum	30	1	28	1	4	1
Maximum	62	15	38	14	23	11
Average	42	6	32	5	22	6
95 percentile	53	11	36	8	23	9

Table 6. Performance Evaluation of Heuristic Procedures for Small-Scale Problems

Instance	c_q	c_{oa}	c_{or}	% change in cost of model-based heuristic from SAA solution	% change in cost of practitioner's heuristic solution from SAA solution
A	1,000	150	450	0	2.97
	1,000	150	900	0	4.61
	1,000	150	225	0	1.15
	1,000	300	450	0	2.34
	1,000	75	450	0	0.75
	2,000	150	450	0	2.97
	500	150	450	0	2.97
B	1,000	150	450	−1.15	3.64
	1,000	150	900	−1.74	5.67
	1,000	150	225	0	1.35
	1,000	300	450	−0.74	2.76
	1,000	75	450	0	0.73
	2,000	150	450	−1.15	3.64
	500	150	450	−1.15	3.64

its worst-case duration increases. Under these circumstances, the practitioner's heuristic is outperformed by the model-based heuristic, as the optimization inherent to the model-based heuristic is more effective in utilizing resources that can be shared across multiple specialties and procedures. Finally, note that increasing the on-call costs leads to the practitioner's heuristic doing much worse as this heuristic opts for increasing the number of on-call anesthesiologists rather than trading off on-call costs against overtime costs.

Table 7. Performance Evaluation of Heuristic Procedures for Medium- and Large-Scale Problems

Instance	c_q	c_{oa}	c_{or}	% change in cost of model-based heuristic from practitioner's heuristic solution
C	1,000	150	450	−4.55
	1,000	150	900	−6.45
	1,000	150	225	−2.56
	1,000	300	450	−3.46
	1,000	75	450	−2.26
	2,000	150	450	−5.55
	500	150	450	−4.57
D	1,000	150	450	−6.64
	1,000	150	900	−4.45
	1,000	150	225	−3.37
	1,000	300	450	−5.52
	1,000	75	450	−2.65
	2,000	150	450	−7.55
	500	150	450	−4.75
E	1,000	150	450	−7.03
	1,000	150	900	−5.74
	1,000	150	225	−4.47
	1,000	300	450	−6.69
	1,000	75	450	−3.45
	2,000	150	450	−7.56
	500	150	450	−4.75

5.2. Model Validation

The objective of model validation was to demonstrate that the model-based heuristic provides tangible cost savings over current practice. This was an essential step in convincing management of the operating services department to implement our method. We performed model validation in two stages. In the first stage, the cost savings were computed using historical data. In the second stage, we conducted live validation, where we compared in real time our decisions with those made at this hospital. Note that while conducting these validations, the heuristic had precisely the same information the planners at the UCLA RRMC had at the point of planning.

In the historical validation, we took 80 sample days, such that we covered the range of problem sizes encountered. These 80 samples were divided into 5 sets as described in Table 8. We next calculated the average costs obtained by the model-based heuristic and the costs resulting from the actual assignment and sequencing that was done by the RRMC planners. This reduction in costs across the five problems is also reported in Table 8, and this shows that the benefits of using the model-based heuristic were significant and increasing in problem size.

The real-time live validation was conducted over a four-week period. The number of surgeries per day

Table 8. Results from Historical Validation

Surgeries per day	% of days	% reduction of cost of model-based heuristic from cost of actual plan
<10	28	0
10–30	4	2.4
30–40	18	3.3
40–50	41	7.2
50–65	9	8.9

Table 9. Results from Live Validation

Surgeries per day	% of days	% reduction of cost of model-based heuristic from cost of actual plan
<10	30	0
10–30	0	0
30–40	11	2.1
40–50	48	6.4
50–65	11	9.1

over this period was similar to the range of problem sizes observed historically as shown in Table 8. The results for the live validation are given in Table 9. This table shows that our heuristic reduced costs from current practice, on average, from 6.4% to 9.1% on 16 out of the 28 days corresponding to weekdays, which were not holidays. This implied an estimated annual cost savings between $2 million and $2.86 million. It is also important to note that the practitioner's heuristic and the model-based heuristic provided the same solution in the weekends, where the number of surgeries conducted are low, and both these methods provided solutions corresponding to the nominal value solution.

The model-based heuristic outperforms the current practice in historical and live validation for the following reasons. First, the nominal values obtained via quantile regression procedure provided better predictors for the realized surgical duration than the surgeon's estimates. Second, on average, around 50% of the surgeries exceeded the nominal value. This required an increase in realized work load from the nominal work load. We found that this increase can be effectively achieved by setting $\tau' = 20\%$. This led to the model-based heuristic operating with fewer ORs and fewer anesthesiologists than actually used at the hospital, since these resource assignments were based on trading off the fixed costs for these resources with the chance of incurring overtime. Third, on average, in 60% of the surgeries, the surgeon's estimate of surgery duration exceeded the realized duration. The planners chose additional resources and avoided overtime based on these quoted times. Thus the associated plans tended to incur more resource usage costs (comprising of the fixed cost of opening an OR and anesthesiologist on-call cost) rather than overtime costs in comparison to the model-based heuristic. However, since this decision to use more resources was made without explicit consideration of overtime costs and errors in the duration estimate provided by the surgeons, this often led to greater total costs. In sum, the model-based heuristic outperforms current practice due to better prediction and a more effective scheduling policy. The proportion of the gains due to each of these aspects are analyzed and summarized in the e-companion.

Generally, in a stochastic decision problem, it is not valid to judge the quality of a decision based on an outcome, as due to randomness, a good outcome does not necessarily imply a good decision. However, in this work, since the evaluation and validation of the model-based heuristic have been extensive, we were confident that they would perform well in the real application. In the final analysis, the real measure of performance of this heuristic is the quality of the decision based on its solution, a question we consider next in the application.

6. Application

6.1. Implementation

We have implemented the model-based heuristic as a decision support system at the operating services department of the UCLA RRMC. Details of this system are provided in the e-companion. The results before and after the implementation across key operational metrics and costs are summarized in Table 10. This table shows after implementation, the average number of anesthesiologists on call decreased by 6.7%, and average overtime hours for the anesthesiologists on regular duty reduced by 3.7%. This contributed to an increase of average daily utilization across the anesthesiologists by 3.5%. Similarly, the average number of ORs used decreased by 8.6%, and the average overtime hours at the OPs was reduced by 2.7%. This led to an increased average daily utilization across the OPs by 3.8%. The improvements in these operational

Table 10. Summary of Results Before and After Implementation of Decision Support System

Attributes	Before	After	% reduction
Average number of anesthesiologists on call per day	6.0	5.6	6.7
Average overtime per day for anesthesiologists (hours)	18.2	17.5	3.7
Average daily utilization of anesthesiologists (%)	75	77.6	−3.5
Average number of ORs used per day	20.4	18.6	8.6
Average overtime per day for ORs (hours)	18.5	18	2.7
Average utilization of ORs per day (%)	78	81	−3.8
Average daily OR costs ($)	57,350	52,417	8.6
Average daily overtime costs ($)	22,375	21,754	2.8
Average daily on-call costs ($)	7,145	6,527	8.5
Average total daily costs ($)	86,870	80,729	7.1

metrics reduced average daily OR costs by 8.6%, average daily overtime costs by 2.7%, and average daily on-call costs by 8.5%. This translates to an overall average daily cost savings of 7% or estimated to be $2.2 million on an annual basis.

The model-based heuristic improved upon decision making at operating services due to two main reasons. First, it was more effective at utilizing the flexibility in the resources. Most anesthesiologists and ORs can perform more than one specialty, typically a primary and a secondary specialty. The model identified these OR/anesthesiologist combinations and allocated surgeries across these different specialties to them. This led to better usage of resources than the previous approach, in which surgeries from a single specialty were assigned to an OR and anesthesiologist as much as possible. A surgery of a different specialty was assigned to an OR only when there was a high volume of surgeries in a particular day, and this was often done without explicit consideration of the allotted anesthesiologists' specialty. Thus, this often required a separate anesthesiologist to perform these surgeries, who were often assigned from on call and this was costlier. Second, the model-based heuristic explicitly considered uncertainty in surgical durations while determining the daily schedule of an OR. By using the estimation module, it determined which surgeries could be longer and more uncertain, and which surgeries could be shorter and more certain. It then combined long uncertain surgeries with short certain surgeries to effectively utilize gaps in the schedule in each OR. This, in turn, reduced the number of ORs each day with the resulting cost reduction being more than any potential increases in overtime costs, thus reducing total costs. In contrast, the previous approach used surgeons' predictions of surgery durations. To compensate for the errors in these predictions, planners often underutilized ORs by leaving sufficient gaps between surgeries. This was done as they did not want to create delays from scheduled start times of succeeding surgeries and incur overtime costs. However, this often led to a larger number of ORs being used each day, and, consequently, higher total costs.

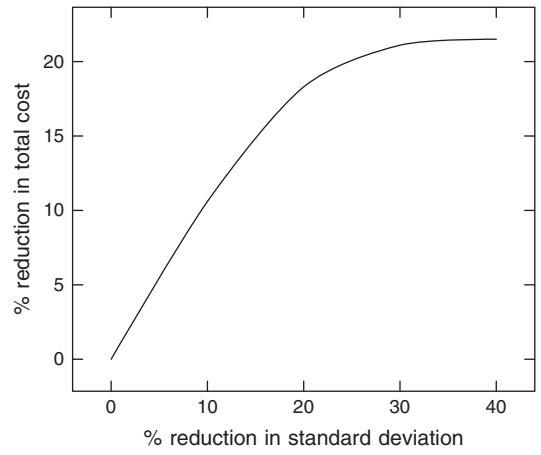
Finally, we considered the impact of the schedules generated by our approach on the surgeons. While surgeons are not part of the operating services department, they are a critical element in the system. First, we computed the average idle time between surgeries and found that it reduced by eight minutes after our work. The surgeons did not find this reduction significant enough to be disruptive, and, in fact, some of them preferred this as it made their schedule more efficient. Second, we calculated the average number of surgeons per OR per day. Prior to our work, on average, there were 1.54 surgeons per OR per day. After implementing the decision support system, there were 1.57 surgeons per OR per day. This marginal increase suggests that most of the benefits of our approach come from making the correct assignment of ORs and anesthesiologist to surgeries, and not from increasing the number of surgeons per OR per day. Both these aspects were important to verify that the surgeons were not inconvenienced by the model-based approach.

6.2. Managerial Insights

We used the model-based heuristic to generate several managerial insights. First, we considered the impact of reducing variability in surgical durations. In practice, this could be achieved by better procedures such as checklists, improved information technology, following the correct sequence in tasks and standardized operating protocols derived from best practices. These measures have been advocated by surgeons (Bates and Gawande 2003, Haynes et al. 2009, Gawande 2010). In addition, variance can be reduced by improving the prediction of surgical durations. This would require dividing the surgical process into a series of steps (such as time to incision, skin to skin, and closure to exit), and predicting each segment individually as different patient characteristics affect each segment differently (Hosseini et al. 2014). The accuracy of this prediction can be improved by collecting more data on patient characteristics and surgeon experience (Kougias et al. 2012). To consider the impact of variance reduction, we started with the current level of standard deviation in surgical durations, and systematically reduced the standard deviation of the distribution of surgical durations across all surgeries by a fixed value. We used these modified distributions to simulate realizations of surgical durations. We then used these data sets to solve the IARSP using the model-based heuristic and calculated the resulting total resource usage and overtime costs. These results are summarized by Figure 3. This figure shows that the benefits of further reduction in variability decreases, and that there are significant diminishing returns on reduction of variability. This suggests that rather than invest in capital-intensive medical equipment to achieve radical reductions in variability in surgical durations, the major cost benefits can be gained by focusing on incremental reduction in variability. This can be potentially achieved by better procedures and more detailed data collection for improved predictive analytics.

Second, we consider the impact of allowing surgeries to start in ORs after 3 P.M. but before the end of the late shift of the anesthesiologists at 7 P.M. This would require additional fixed technician and nurse staffing costs. Such extensions can be considered if surgical demand on any day is significantly larger than average daily surgical demand. To perform this analysis, we considered four levels of demand corresponding to increases from daily average demand in surgeries that could occur during the days of any given week. For

Figure 3. Effect of Reducing Variability of Surgical Durations on Total Resource Usage and Overtime Costs



these scenarios, we incrementally increased the number of ORs available after 3 P.M. by one unit. We then calculated the resulting change in total resource usage and overtime costs from the case when we do not start surgeries after 3 P.M., but only use the day shift with additional rooms to accommodate such increases in demand. These results summarized in Figure 4 suggest that it is beneficial to allow such extensions and the number of ORs used depends on the level of demand. This analysis helps management understand how best to react to different levels of daily surgical demand and estimate the corresponding changes in total costs.

Finally, we examined the benefit of increasing cross-functionality of the ORs. To do so, we considered the

Figure 4. Effect of Extending Shift Timings of ORs on Total Resource Usage and Overtime Costs

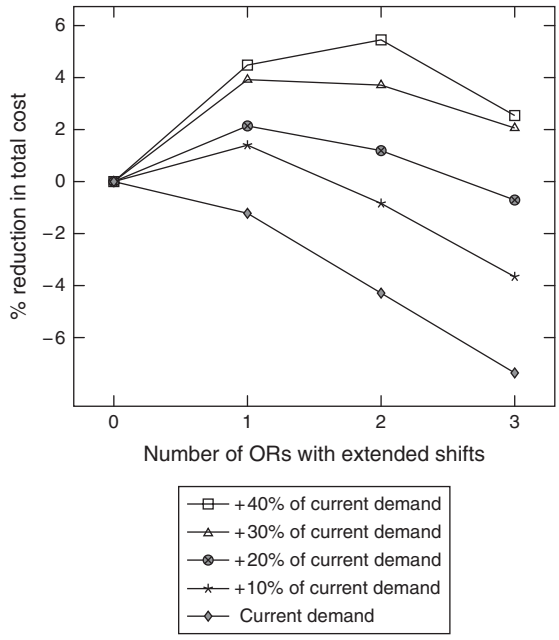
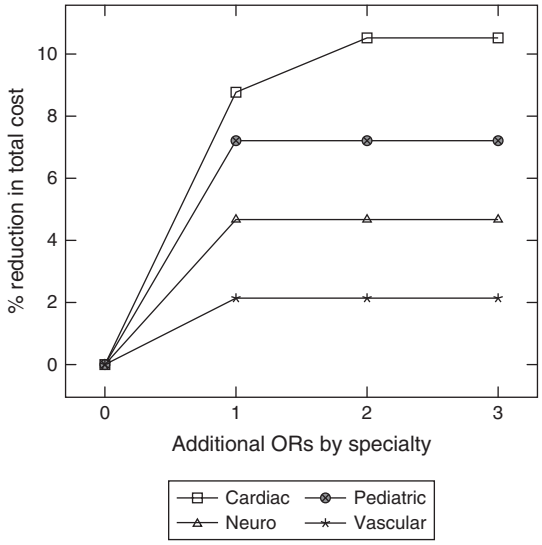


Figure 5. Effect of Increasing Number of ORs by Specialty on Total Resource Usage and Overtime Costs



various specialties and calculated the potential reduction in costs if the number of ORs available for each specialty described in Table 1 is increased. In practice, such increases can be achieved by investing in special equipment to convert general surgery ORs to have the cross-functionality to accommodate a particular specialty. These results described in Figure 5 show that as we increase the number of ORs that could be used for a particular specialty, this can lead to a significant reduction in total resource usage and overtime costs, and these benefits are often more pronounced in certain specialties. This analysis forms a basis to identify such specialties, and determine the priority in which these ORs should be made cross-functional to enable these additional rooms for the specialties. Further, an additional advantage of making ORs cross-functional was that a higher number of daily surgeries could be more effectively accommodated without conducting new surgeries after 3 P.M. In particular, we found this approach led to at least an additional 5% reduction from the lowest costs attainable for all the demand scenarios considered in Figure 4. This provides further justification for management to make the ORs more cross-functional.

6.3. Qualitative Impact

The organizational impact of our work has been significant. Prior to our work, simple rules were used to make important decisions on allocation of anesthesiologists and rooms to surgeries and determining surgery start times. These rules developed based on experience and anecdotal evidence worked well during holidays and weekends when the number of surgeries conducted were low. However, as shown in Section 5 and observed during the implementation, our model

significantly outperformed current practice during other days where the number of surgeries performed was high, and this resulted in considerable cost savings. Thus our work demonstrated the value of model-based approach and operations research methods in dealing with complexity. This has encouraged management to investigate other problems in this department using a structured and rigorous approach by employing operations research-based methodologies.

The managerial insights generated from our model have also contributed to the organizational impact. While the effect of variance reduction on improved clinical outcomes has been extensively documented (Neuhauser et al. 2011), our analysis showed that this could also reduce costs. This provided management with the further impetus to implement six sigma programs (Cima et al. 2011) to reduce variability at this department. In addition, our analysis provides management with clear guidance on when to start new surgeries after the day shift and in how many rooms. This provides them with a practical approach to mitigate the impact of varying levels of daily surgical demand on costs, and is currently under consideration for implementation in the short term. Finally, we showed the benefits of making some operational rooms cross-functional and how to prioritize implementation among the specialties. Furthermore, we demonstrate that this could potentially be a very effective way to accommodate changes in daily surgical demand. While management at the operating services department was intrigued by this analysis, they felt that there could be significant investments required, and this could also lead to disruptions in the schedule while some ORs were being reconfigured. Therefore they are considering this initiative as part of the next broader hospital renovation project.

6.4. Limitations

This work has the following limitations. First, the estimation of uncertainty sets can be improved with additional data on the duration of each step in a surgery. However, this data were not available in our application. Second, we do not explicitly consider requests from surgeons for particular start times on a given day and for specific anesthesiologists. While these aspects can be easily incorporated in our model, management felt that accommodating these requests explicitly can make the overall schedule inefficient and could create additional costs. Therefore they preferred to make changes to the output in the decision support system only in the most exceptional circumstances. Third, we assume that the overtime payment sufficiently compensates staff for extended shifts. However, in practice, such extensions are unpredictable and staff may not prefer such type of overtime. Thus, there is an implicit inconvenience cost associated with the overtime cost that is not considered in our work. Similarly, we do not consider the inconvenience costs associated with an anesthesiologist being on call, but not being asked to come in to work. While these aspects can be included in our model by suitably appending the overtime and on-call costs with the appropriate inconvenience costs, quantifying these costs would be challenging. In this regard, recent research in structural estimation (Olivares et al. 2008) could potentially be used to calculate these inconvenience costs and further enhance the outputs of the model. Finally, some anesthesiologists can be used across multiple specialties and this feature was incorporated in our model. However, we do not consider their preferences across specialties, as such data was unavailable to us. Future work could focus on all these aspects to improve the model and its ability to attend to the interests of the surgical teams.

In conclusion, the methodology described in this paper has had a major economic and organizational impact at the operating services department at the UCLA RRMC. This organization expects to maintain the described gains and to increase them continuously several years into the future.

Acknowledgments

The authors are indebted to Catherine Duda for initiating this project. The authors thank Dr. Ira Hofer and Dr. Michael J. Sopher for their generous assistance in several aspects of this work. The authors also gratefully acknowledge the excellent and constructive comments from Professor Andres Weintraub, an associate editor, and three anonymous reviewers.

Endnotes

[1] http://health.usnews.com/health-care/best-hospitals/articles/best-hospitals-honor-roll-and-overview.

[2] https://www.asahq.org/resources/clinical-information/asa-physical-status-classification-system.

[3] http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt.page.

References

AMA (2016) https://www.ama-assn.org/practice-management/newly-released-codes.

Augusto V, Xie X, Perdomo V (2010) Operating theatre scheduling with patient recovery in both operating rooms and recovery beds. *Comput. Indust. Engrg.* 58(2):231–238.

Bates DW, Gawande AA (2003) Improving safety with information technology. *New England J. Medicine* 348(25):2526–2534.

Batun S, Denton BT, Huschka TR, Schaefer AJ (2011) Operating room pooling and parallel surgery processing under uncertainty. *INFORMS J. Comput.* 23(2):220–237.

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

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust Optimization* (Princeton University Press, Princeton, NJ).

Bertsimas D, Sim M (2004) The price of robustness. *Oper. Res.* 52(1):35–53.

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

Bertsimas D, Gupta V, Kallus N (2013) Data-driven robust optimization. Preprint arXiv:1401.0212.

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

Cardoen B, Demeulemeester E, Beliën J (2009a) Optimizing a multiple objective surgical case sequencing problem. *Internat. J. Production Econom.* 119(2):354–366.

Cardoen B, Demeulemeester E, Beliën J (2009b) Sequencing surgical cases in a day-care environment: An exact branch-and-price approach. *Comput. Oper. Res.* 36(9):2660–2669.

Cardoen B, Demeulemeester E, Van der Hoeven J (2010) On the use of planning models in the operating theatre: Results of a survey in flanders. *Internat. J. Health Planning and Management* 25(4): 400–414.

Carøe CC, Schultz R (1999) Dual decomposition in stochastic integer programming. *Oper. Res. Lett.* 24(1):37–45.

Cima RR, Brown MJ, Hebl, JR, Moore R, Rogers JC, Kollengode A, Amstutz GJ, Weisbrod CA, Narr BJ, Deschamps C (2011) Use of lean and six sigma methodology to improve operating room efficiency in a high-volume tertiary-care academic medical center. *J. Amer. College of Surgeons* 213(1):83–92.

Denton B, Gupta D (2003) A sequential bounding approach for optimal appointment scheduling. *IIE Trans.* 35(11):1003–1016.

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

Denton BT, Miller AJ, Balasubramanian HJ, Huschka TR (2010) Optimal allocation of surgery blocks to operating rooms under uncertainty. *Oper. Res.* 58(4-part-1):802–816.

Dexter F, Traub RD (2002) How to schedule elective surgical cases into specific operating rooms to maximize the efficiency of use of operating room time. *Anesthesia and Analgesia* 94(4):933–942.

Dexter F, Wachtel RE (2014) Strategies for net cost reductions with the expanded role and expertise of anesthesiologists in the perioperative surgical home. *Anesthesia and Analgesia* 118(5):1062–1071.

Dexter F, Wachtel RE, Epstein RH (2016) Decreasing the hours that anesthesiologists and nurse anesthetists work late by making decisions to reduce the hours of over-utilized operating room time. *Anesthesia and Analgesia* 122(3):831–842.

Gawande A (2010) *Complications: A Surgeon's Notes on an Imperfect Science* (Profile Books, London).

Ghaly RF (2014) Do neurosurgeons need neuroanesthesiologists? Should every neurosurgical case be done by a neuroanesthesiologist? *Surgical Neurology Internat.* 5(1):365.

Green LV, Savin S (2008) Reducing delays for medical appointments: A queueing approach. *Oper. Res.* 56(6):1526–1538.

Gul S, Denton BT, Fowler JW, Huschka T (2011) Bi-criteria scheduling of surgical services for an outpatient procedure center. *Production Oper. Management* 20(3):406–417.

Gupta D (2007) Surgical suites' operations management. *Production Oper. Management* 16(6):689–700.

Gurobi Optimization I (2015) Gurobi optimizer reference manual. http://www.gurobi.com.

Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL (2006) *Multivariate Data Analysis*, Vol. 6 (Pearson Prentice-Hall, Upper Saddle River, NJ).

Haynes AB, Weiser TG, Berry WR, Lipsitz, SR, Breizat A-HS, Dellinger EP, Herbosa T, Joseph S, Kibatala PL, Lapitan MCM, Merry AF, Moorthy K, Reznick RK, Taylor B, Gawande AA, (2009) A surgical safety checklist to reduce morbidity and mortality in a global population. *New England J. Medicine* 360(5): 491–499.

He B, Dexter F, Macario A, Zenios S (2012) The timing of staffing decisions in hospital operating rooms: Incorporating workload heterogeneity into the newsvendor problem. *Manufacturing Service Oper. Management* 14(1):99–114.

Hosseini N, Hallbeck MS, Jankowski CJ, Huddleston JM, Kanwar A, Pasupathy KS (2014) Effect of obesity and clinical factors on preincision time: Study of operating room workflow. *Proc. AMIA Annual Sympos.*, Vol. 2014 (American Medical Informatics Association, Bethesda, MD), 691–699.

Kayis E, Wang H, Patel M, Gonzalez T, Jain, S, Ramamurthi RJ, Santos C, Singhal S, Suermondt J, Sylvester K, (2012) Improving prediction of surgery duration using operational and temporal factors. *AMIA Annual Sympos. Proc.*, Vol. 2012 (American Medical Informatics Association, Bethesda, MD), 456–462.

Kelley JE Jr (1960) The cutting-plane method for solving convex programs. *J. Soc. Indust. Appl. Math.* 8(4):703–712.

Kleywegt AJ, Shapiro A, Homem-de Mello T (2002) The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.* 12(2):479–502.

Koenker R (2005) *Quantile Regression* (Cambridge University Press, New York).

Koenker R (2013) Quantreg: Quantile regression. R Package Version 5.

Kougias P, Tiwari V, Orcutt S, Chen A, Pisimisis G, Barshes NR, Bechara CF, Berger DH (2012) Derivation and out-of-sample validation of a modeling system to predict length of surgery. *Amer. J. Surgery* 204(5):563–568.

Laskin DM, Abubaker AO, Strauss RA (2013) Accuracy of predicting the duration of a surgical operation. *J. Oral and Maxillofacial Surgery* 71(2):446–447.

Macario A (2010) What does one minute of operating room time cost? *J. Clinical Anesthesia* 22(4):233–236.

Mak HY, Rong Y, Zhang J (2014a) Appointment scheduling with limited distributional information. *Management Sci.* 16(2): 316–334.

Mak HY, Rong Y, Zhang J (2014b) Sequencing appointments for service systems using inventory approximations. *Manufacturing Service Oper. Management* 16(2):251–262.

Mancilla C, Storer R (2012) A sample average approximation approach to stochastic appointment sequencing and scheduling. *IIE Trans.* 44(8):655–670.

Marcon E, Dexter F (2006) Impact of surgical sequencing on post anesthesia care unit staffing. *Health Care Management Sci.* 9(1): 87–98.

Märkert A, Gollmer R (2008) User's guide to ddsip–AC package for the dual decomposition of two-stage stochastic programs with mixed-integer recourse.

Marques I, Captivo ME, Vaz Pato M (2014) Scheduling elective surgeries in a Portuguese hospital using a genetic heuristic. *Oper. Res. Health Care* 3(2):59–72.

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

McNicol R (1997) Paediatric anaesthesia–Who should do it? The view from the specialist hospital. *Anaesthesia* 52(6):513–515.

Meskens N, Duvivier D, Hanset A (2013) Multi-objective operating room scheduling considering desiderata of the surgical team. *Decision Support Systems* 55(2):650–659.

Min D, Yih Y (2010) Scheduling elective surgery under uncertainty and downstream capacity constraints. *Eur. J. Oper. Res.* 206(3):642–652.

Neuhauser D, Provost L, Bergman B (2011) The meaning of variation to healthcare managers, clinical and health-services researchers, and individual patients. *BMJ Quality and Safety* 20(Suppl 1): i36–i40.

Olivares M, Terwiesch C, Cassorla L (2008) Structural estimation of the newsvendor model: An application to reserving operating room time. *Management Sci.* 54(1):41–55.

Orkin FK, McGinnis SL, Forte GJ, Peterson, MD, Schubert A, Katz JD, Berry AJ, Cohen NA, Holzman RS, Jackson SH, Martin DE, Garfield JM (2013) United States anesthesiologists over 50: Retirement decision making and workforce implications. *Survey of Anesthesiology* 57(2):101–102.

Pardo M (2014) The development of education in anesthesia in the United States. Eger E, Saidman L, Westhorpe R, eds. *The Wondrous Story of Anesthesia* (Springer, New York), 483–496.

Ruszczyński AP (2006) *Nonlinear Optimization*, Vol. 13 (Princeton University Press, Princeton, NJ).
Saadouli H, Jerbi B, Dammak A, Masmoudi L, Bouaziz A (2015) A stochastic optimization and simulation approach for scheduling operating rooms and recovery beds in an orthopedic surgery department. *Comput. Indust. Engrg.* 80:72–79.
Sahinidis NV (2014) BARON 14.3.1: Global Optimization of Mixed-Integer Nonlinear Programs. User's Manual.
Thiele A, Terry T, Epelman M (2009) Robust linear optimization with recourse. *Rapport Technique* 4–37.
Tulabandhula T, Rudin C (2014) Robust optimization using machine learning for uncertainty sets. Preprint arXiv:1407.1097.
van Rossum FGD (2001) Python reference manual, Pythonlabs, VA. http://www.python.org.
Viswanathan J, Grossmann IE (1990) A combined penalty function and outer-approximation method for MINLP optimization. *Comput. Chemical Engrg.* 14(7):769–782.

Sandeep Rath is an assistant professor in operations at the Kenan Flager Business School at the University of North Carolina at Chapel Hill. His current research focuses on healthcare operations. His research interest is in applying analytical methods to complex business problems.

Kumar Rajaram is a professor of operations and technology management and the Ho Su Wu Chair in Management at the UCLA Anderson School of Management. His current research interests include improving operations in the healthcare, entertainment, and process manufacturing sectors, including food processing, pharmaceuticals, and the petrochemical industry. He has focused on developing analytical models of complicated systems with a strong emphasis on practical implementation. For his work in the process industry, he has been recognized as an Edelmann Laureate by INFORMS.

Aman Mahajan serves as Executive Chair of the Department of Anesthesiology and Perioperative Medicine at the UCLA David Geffen School of Medicine, and as the Ronald L. Katz Professor of Anesthesiology and Bioengineering. He is also the Executive Director of Operative Services at the UCLA Medical Center and a Co-Director of the UCLA Cardiac Arrhythmia Program. He is actively engaged in research that seeks a greater understanding of the mechanisms of sudden cardiac death.