POMS PRODUCTION AND OPERATIONS MANAGEMENT

PRODUCTION AND OPERATIONS MANAGEMENT

Vol. 31, No. 3, March 2022, pp. 1271–1289
ISSN 1059-1478 | EISSN 1937-5956 | 22 | 3103 | 1271
DOI 10.1111/poms.13610

Staff Planning for Hospitals with Implicit Cost Estimation and Stochastic Optimization

Sandeep Rath*
Kenan-Flagler Business School, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27517, USA, sandeep_rath@kenan-flagler.unc.edu

Kumar Rajaram
UCLA Anderson School of Management, Los Angeles, California 90095, USA, kumar.rajaram@anderson.ucla.edu

We consider the anesthesiologist staff planning problem for operating services departments in large multi-specialty hospitals without limit on anesthesiologist supply, where the planner makes monthly and daily decisions to minimize total costs. Each month the staff planner decides the number of anesthesiologists on regular duty and an on-call consideration list for each day of the following month. In addition, each day, the staff planner decides how many on-call anesthesiologists to call for the following day. Total costs consist of explicit and implicit costs. Explicit costs include the costs of calling an anesthesiologist and overtime costs. These costs are specified by the organization. Implicit costs encompass costs of not calling an on-call anesthesiologist and under-utilizing an anesthesiologist, and these have to be deduced from past decisions. We model the staff planning problem as a two-stage integer stochastic dynamic program. We develop structural properties of this model and use them in a sample average approximation algorithm constructed to solve this problem. We also develop a procedure to estimate the implicit costs, which are included in this model. Using data from the operating services department at the UCLA Ronald Reagan Medical Center, our model shows the potential to reduce overall costs by 16%. We provide managerial insights related to the relative scale of these costs, hiring decisions by service, sensitivity to cost parameters, and improvements in the prediction of the booked time durations.

Key words: healthcare stochastic models; workforce planning; operating room
History: Received: January 2020; Accepted: October 2021 by Chelliah Sriskandarajah, after 4 revisions.
*Corresponding author.

1. Introduction

Healthcare expenditures in the United States are expected to rise to 20% of GDP by 2027 (Sisko et al. 2019). Evidence suggests that a significant portion of this expenditure is wasted because of operational inefficiencies at healthcare sites such as hospitals, which constitute around 32% of healthcare expenditures in the United States (Smith et al. 2012). In hospitals, the total labor expenditure can exceed 50% of operating costs and may be up to 90% of variable costs (Healthcare Insights 2014). Thus, efficient deployment of labor becomes one of the primary methods of cost control at hospitals.

There are several challenges in managing labor at a hospital. First, there is uncertainty in the demand for services. Second, the skill set of staff is often specialized and not easily substitutable. Finally, because of the characteristics of health services, tactics such as production smoothing cannot be employed effectively. Hospitals make efficient use of labor through staffing that can be made flexible in volume by calling additional employees, use of floating resources, and overtime (Kesavan et al. 2014). Such volume flexibility can help reduce costs at hospitals by reacting to changes as information about the future workload become available (Bard and Purnomo 2005). Volume flexibility in hospitals has been used in staff planning for nurses and physicians (Brunner et al. 2009).

Overtime is a key feature in achieving volume flexibility. However, some researchers associate excessive overtime of clinical staff with lower patient safety (Rogers et al. 2004), higher employee burnout (Stimpfel et al. 2012), and deteriorating employee health (Trinkoff et al. 2006). Thus, to reduce reliance on overtime, staff planners often use additional employees who can be called on short notice. The use of this contingency labor supply reduces the number of overtime hours. However, depending on the staffing policy, this may give rise to additional administrative costs. These consist of both explicit and implicit costs. Explicit costs represent the actual monetary payment made and recorded for an activity. Such costs could include overtime compensation and extra payments made to staff who report for work on short notice. In contrast, an implicit cost is not recorded but instead

implied. Implicit costs could include the opportunity cost to the organization associated with staff idle time and the inconvenience to employees whose schedules change on short notice.

Traditionally, staff planning at hospitals has been a manual process. While evidence suggests that the use of analytic, data-driven, model-based systems would be beneficial from a cost perspective (Healthcare Insights 2014), implementing such systems for labor scheduling has been challenging. Some staff planning systems have not been successful at large retail organizations like Starbucks (Kantor 2015). The principal challenge in implementing model-based staff planning systems is minimizing overall costs by incorporating the explicit and implicit human costs of the employees. Not incorporating all the human costs would likely lead to failure in acceptance and implementation of these systems (Bernstein et al. 2014).

In this study, we provide an approach to estimate the implicit costs in staff planning. Subsequently, we use explicit and implicit costs in an optimization model for anesthesiologist staff planning at the UCLA Ronald Reagan Medical Center (RRMC).

1.1. Problem Description

The UCLA RRMC is a large multi-specialty hospital that consistently ranks among the best five hospitals in the United States (http://health.usnews.com/health-care/best-hospitals/articles/best-hospitals-honor-roll-and-overview). The operating services department of the UCLA RRMC is responsible for staffing physician anesthesiologists to surgical services at the hospital. The focus of our work is the staff planning of physician anesthesiologists at this department of the UCLA RRMC.

The operating services department manages the surgery suite at the UCLA RRMC. Surgeons across all services in this hospital perform around 27,000 surgeries annually across 2700 unique surgery types. The UCLA RRMC classifies the anesthesia required for these surgeries into four services: Cardiothoracic, General, Neuro, and Pediatric. The staff planning for anesthesiologists consists of two stages: monthly and daily decisions. We give details of these decisions below.

- *Monthly decisions*: By the 20th of each month, depending on the teaching and vacation commitments of anesthesiologists, the staff planner knows the availability of anesthesiologists for each day of the upcoming month. Once anesthesiologists have provided their availability, they can be scheduled across all these days. Based on this availability and the historical data of surgical workload, the planner prepares the staffing plan for each service for each day of the following month. This plan consists of dividing the anesthesiologists available each day of the following month into two groups: those who would be available on regular duty and those on a reserve list, called the *on-call* consideration list. Anesthesiologists on the on-call consideration list are informed the day before the surgery if their services are required the next day. In this case, they are paid an additional $1000 for the entire day. However, if they are not needed, they are not paid this additional amount. Thus, being on the on-call consideration list and not being called is not desirable for the employees. The planner manages the number of employees on the on-call consideration list so that this does not occur frequently.
- *Daily decisions*: The day before the surgery, the planner schedules the total number of elective procedures to be performed the next day and finalizes the anesthesiologists' booked hours. Based on this information, a certain number of anesthesiologists of each service from the on-call consideration list are informed that they would be working the next day. The number of anesthesiologists actually called and the number of anesthesiologists on regular duty determines the total available work hours. When the actual surgical hours are realized, the costs of overtime or idle time are realized.

The staff planner has to balance four costs when making the monthly and daily decisions involved in the staffing plan. These include:

1. The explicit cost of calling anesthesiologists from the on-call consideration list. This is the additional payment made to the anesthesiologists for coming on short notice. At the UCLA RRMC, this was $1000 per day.
2. The implicit cost of having anesthesiologists on the on-call consideration list but not calling them. This is the inconvenience cost of keeping an anesthesiologist on hold for a day and not compensating him or her. Anesthesiologists on the on-call consideration list have to alter their schedule outside work such that they have to stay within an acceptable distance from the hospital. Therefore, there is an inconvenience in being placed on the on-call list (Olmstead et al. 2014). At UCLA RRMC, the physicians are only compensated if they are called. Therefore, the inconvenience cost of being on the on-call list is implicit.

However, physicians expect that sometimes they might be placed on-call and not get called.

Therefore, we assume that being on-call but not getting called has a cost only after a threshold.

3. On the day of the surgery, each anesthesiologist on regular duty works an eight-hour shift. If a surgery in progress is incomplete at the end of the shift, there are no hand-offs, and the anesthesiologist continues to accrue overtime. At the UCLA RRMC, such overtime is compensated at $180 per hour.
4. If the total number of work hours of available anesthesiologists is greater than the total realized hours of surgery, there will be idle time. The operating services department seeks to keep idle time low, and thus, there is an implicit cost of idle time.

In Table 1, we present the summary statistics of the number of anesthesiologists on regular duty, those on the on-call consideration list, and those who actually get called. This table shows that, on average, 17.48 anesthesiologists work on regular duty; 6.89 are on the on-call list, out of which 2.77 are called. Furthermore, there is considerable variation in staffing levels across services. This is primarily because of the demand characteristics of the services. General anesthesia services require a greater proportion of on-call anesthesiologists than other services. This is because the coefficient of variation of daily demand for general anesthesia services is larger than that of other services.

In 2014, the UCLA RRMC instituted an electronic health system (http://careconnect.uclahealth.org/about-careconnect). The management at the operating services department was keen on using the data from this system to develop an analytical model-based approach to staff planning that incorporated all the relevant costs. Implementing such an analytical model to address staff planning could pose similar challenges, as described in Kantor (2015), and Bernstein et al. (2014) if UCLA RRMC does not incorporate implicit human costs of staffing. Therefore, we take a two-part approach to staff planning at this hospital. In the first part, we model the staff planning as a two-stage integer stochastic dynamic program. The first stage captures the monthly decisions, while the second stage includes the daily decisions involved in staff planning. We then develop an algorithm to solve this model to provide the monthly and daily anesthesiologist staffing plan across each service for given cost parameters. In the second part, we develop a procedure to estimate the implicit costs. These include the inconvenience costs of scheduling anesthesiologists on the on-call consideration list but not calling them and the implicit cost of idle time. Subsequently, we use these estimated costs to demonstrate the total cost savings from using the optimization model.

Table 1 Summary Statistics for Historical Anesthesiologist Planning by Service

Service	Staff type	Average	Max	Min	SD
Cardiothoracic	Regular	4.93	10	0	1.63
	On-call consideration	1.18	6	0	1.04
	On-call actually called	0.45	5	0	0.74
General	Regular	8.65	16	0	2.58
	On-call consideration	4.61	7	0	1.76
	On-call actually called	1.85	10	0	1.85
Neuro	Regular	2.72	6	0	0.85
	On-call consideration	0.72	4	0	0.77
	On-call actually called	0.30	3	0	0.56
Pediatric	Regular	1.69	6	0	0.76
	On-call consideration	0.53	4	0	0.73
	On-call actually called	0.24	4	0	0.47
Total	Regular	17.48	26	0	3.12
	On-call consideration	6.89	11	0	2.07
	On-call actually called	2.77	9	0	1.32

1.2. Literature Review

The staff planning problem we consider in this study is related to three streams of literature. The first is in staff planning for services, particularly for operating rooms. The second stream is based on two-stage stochastic dynamic programming models. The third is associated with the estimation of operational parameters.

Several papers model the stages of staff planning at service organizations as a dynamic optimization problem. Wild and Schneewei (1993) provide a model for staff planning for the long term, medium term, and short term when volume flexibility is available in the form of contingent workers. Pinker and Larson (2003) provide a model for flexible workforce management in environments with uncertainty in the demand for labor. In the context of staff planning at hospitals, Dexter et al. (2005) provide a framework for tactical decision-making when allocating operating room time approximately one year in advance. The decisions that are a part of this time frame include hiring additional staff and building new operating rooms. Slaugh et al. (2018) provide results around managing on-call pool of nursing staff to provide last minute staffing for nurse absences. He et al. (2012) analyze decision-making for nurse staffing as more information becomes available about the workload on the day of the surgery. Through numerical analysis, they identify that deferring staffing decisions until the time procedure type information is available could help hospitals save up to 49% of staffing costs. While hospitals would like to defer staffing decisions as late as possible, this often leaves staff without final schedules until shortly before the day of the surgery. This uncertainty in schedules is not

desirable from a staff perspective. Thus, the UCLA RRMC, like several other service organizations, mitigates this problem by using a base level of staff who know they will be required on a given day and a reserve (on-call) list. Anesthesiologists on the on-call list will know if they need to come in only the previous day. McIntosh et al. (2006) state that this refinement of service-specific staffing, months before the day of the surgery, has a high degree of influence on staff satisfaction at hospitals. Xie and Zenios (2015) analyze the nursing staff planning problem within a time frame of a few months and propose a dynamic staffing policy, with adjustments to staffing levels as information on different types of surgeries arrives sequentially. They find that a threshold policy (analogous to a base stock policy) is optimal.

The staff planning problem at the UCLA RRMC is a two-stage integer stochastic dynamic program. When we remove the integrality requirement, this problem reduces to a two-stage stochastic dynamic program. Such problems have been extensively studied (Birge 1985). When applied in the retail context, this is known as a two-stage newsvendor problem. Gurnani and Tang (1999) characterize the optimal solution to this problem at a retailer that has two instances to order a seasonal product. Fisher et al. (2001) propose a heuristic solution to solve the two-stage newsvendor problem in an application at a catalog retailer. Recently, such two-stage models have also been used in agro-business (Bansal and Nagarajan 2017). In contrast, integrality requirements in our problem are essential because we consider staff planning and, as shown in Table 1, the average number of anesthesiologists deployed in each service on a given day is small. Recent theoretical work on integer stochastic dynamic programs includes Kong et al. (2013) and Sun et al. (2015). Easton (2014) considers a two-stage problem to manage workforce allocation by incorporating the joint variability of attendance and demand. Kim and Mehrotra (2015) consider a two-stage nurse staffing and scheduling problem. In the first stage, they find the initial staffing levels and schedules, and in the second stage, they adjust staffing levels after demand is observed. To solve this problem, they employ a two-stage stochastic integer program. Their modeling approach differs from our work in two key aspects. First, they assume no uncertainty in the second stage, so staffing adjustments are made in this period under known demand. Second, the feasible set of second-stage staffing patterns is constant and does not depend on the first-stage decision. In contrast, our application context required that we consider uncertainty in both stages, and the second-stage problem depends on the first-stage decision. These aspects significantly complicate the solution method. In addition, in their numerical analysis, they assume that the cost of idle time was zero. This was an important parameter in our setting, and we apply a data-driven approach to estimate the idle time cost of the anesthesiologists.

Literature related to dynamic optimization-based staff planning assumes that all the appropriate costs are known. As we described before, this is often not the case since there are several implicit costs in staff planning. Dexter and O'Neill (2001) discuss the importance of implicit costs in creating a staffing plan for anesthesiologists. Therefore, for an optimization model to be useful, these implicit costs must be estimated and included. In the econometric literature, Rust (1987) and Aguirregabiria (1999) discuss structural estimation of the costs involved in dynamic problems.

In the operations management literature, Allon et al. (2011) use a structural estimation approach to estimate the impact of waiting time performance on market share in the fast-food industry. Deshpande and Arikan (2012) estimate how airline schedules affect flight delays. Structural estimation of operational parameters has also been used in the call center industry by Aksin et al. (2017) to estimate customer preferences. In terms of the application context, our paper is closest to Olivares et al. (2008), who model the operating room time allocation problem as a newsvendor problem. They then employ a structural estimation approach to assess the relative costs of idle time and overtime for operating rooms. However, all these papers use the estimates created from structural estimation primarily for descriptive purposes, and they are not linked with an optimization model. This link is of significant importance in our application context. Furthermore, structural estimation assumes that the decision-maker makes optimal decisions and therefore does not capture the errors made by the decision-maker in the decision process. To overcome this, in our estimation procedure, we use an approach similar to Su (2008) ho assume that the decision-maker is bounded rational. This implies that decision-makers are not perfect optimizers and make errors resulting from both insufficient information and cognitive limitations.

1.3. Contributions and Managerial Insights

Our paper makes the following contributions. First, we develop a two-stage integer stochastic dynamic programming model for medium- and short-term planning for anesthesiologists while incorporating implicit costs, demand uncertainty, and multiple services. To the best of our knowledge, this is the first paper to consider this approach in the healthcare industry. Second, this study develops a procedure to estimate implicit cost parameters used in the model. This provides a framework for creating staff planning models that overcome the shortcomings of dynamic

optimization models in situations where some cost parameters may be implicit, as is often the case in service organizations. Third, we provide structural results and develop a general method for solving two-stage integer stochastic dynamic programs, which can also be used in other applications. Fourth, we test our model with real data at the operating services department at the UCLA RRMC and demonstrate cost savings from such an estimation and optimization approach.

We draw several managerial insights from this work. These include:

1. The implicit cost of not calling an anesthesiologist on the call list is significantly more expensive than actually calling the anesthesiologist.
2. The implicit costs of idle time are substantially higher than the costs of overtime. This suggests that it is important to have a data-based understanding of implicit costs to make effective staff planning decisions.
3. It may be efficient to have more anesthesiologists on the on-call consideration list, as long as we carefully choose the days that require an on-call list.
4. A good understanding of demand variability and differences in costs can reduce overall staffing costs across specialties.

The remainder of the study is organized as follows. In section 2, we provide the formulation of the model and describe the variables, parameters, objectives, and constraints. We also provide the structural properties of the model and describe its solution method. In section 3, we describe the data and methodology for the estimation of demand for anesthesia services based on historical data. In section 4, we present the procedure to estimate the implicit cost parameters. In section 5, we describe the results of the computational analysis. In section 6, we summarize our work, provide managerial insights, describe the limitations of our study, and suggest future research directions.

2. Model

We start by presenting a model formulation of the staff planning problem. To provide a precise definition of the model, let S be the set of services {*Cardiothoracic*, *General*, *Neuro*, *Pediatric*}, and let T be the set of days in a given month. We define the following variables, which are optimized:

x_{st}: Number of anesthesiologists of service $s \in S$ placed on regular duty on day $t \in T$.
y_{st}: Number of anesthesiologists of service $s \in S$ placed on the on-call consideration list on day $t \in T$.
z_{st}: Number of anesthesiologists of service $s \in S$ *called* from the on-call list for day $t \in T$.

Next, we define the following parameters or inputs:

n_{st}: The number of anesthesiologists of service s available for day $t \in T$.
h: The regular hours of work done per day by an anesthesiologist (hours).
c_o: Overtime cost of anesthesiologists ($/hour).
c_u: Idle time cost of anesthesiologists ($/hour).
c_q: Cost of calling an anesthesiologist from the on-call list ($/day).
c'_q: Cost of keeping an anesthesiologist on the on-call list but not calling ($/day).
τ: Threshold parameter (anesthesiologist per day).
B_{st}: The distribution of anesthesia hours booked for service $s \in S$ for day t.
$\tilde{B}_{st}$: Realization of B_{st}.
D_{st}: The distribution of anesthesia hours used for service $s \in S$ at the end of day t
$\tilde{D}_{st}$: Realization of D_{st}.
$f(D_{st}|B_{st})$, $F(D_{st}|B_{st})$: the marginal density and distribution of D_{st} given B_{st} respectively.

Furthermore, for conciseness, let:

$a^+ = \max(0, a)$.
$\lceil a \rceil = \min\{n \in \mathbb{Z} | n \geq a\}$.
$\lfloor a \rfloor = \max\{n \in \mathbb{Z} | n \leq a\}$.
$\mathbf{c} = (c_o, c_u, c_q, c'_q)$.

The staff planning model is a two-stage, integer stochastic dynamic program. The first stage consists of the Monthly Staff Planning Problem (*MSPP*), which determines the number of anesthesiologists on regular duty and the on-call list for each day of the given month across each service. The second stage consists of the Daily Staffing Planning Problem for service s in time period t ($DSPP_{st}$). This determines how many anesthesiologists to call from the on-call list for service s for day t. We next describe each of these problems in detail.

In the *MSPP*, the planner makes staffing decisions before the beginning of the given month. Thus, at this point, the planners are only aware of the historical distribution of B_{st} and the total number of anesthesiologists available for each day of this month (n_{st}). For each service on each day of the upcoming month, the planners decide the number of anesthesiologists who should be present for regular duty (x_{st}) and the number of anesthesiologists who should be a part of the on-call consideration list (y_{st}). The *MSPP* is formulated as:

$$(\textbf{MSPP}) \quad \mathcal{V}(\mathbf{n}, \mathbf{c}) = \min \sum_{s\in S, t\in T} \{\mathbf{E}_{\mathbf{B_{st}}}[\mathcal{W}_{st}(x_{st}, y_{st}; \mathbf{c}, B_{st}, n_{st})]\}, \quad (1)$$

$$x_{st} + y_{st} \le n_{st} \quad \forall s \in S, t \in T, \quad (2)$$

$$x_{st}, y_{st} \in \mathbb{Z}_0^+ \quad \forall s \in S, t \in T. \quad (3)$$

The objective (1) represents the total expected monthly costs. This is the sum of expectation of $\mathcal{W}_{st}(x_{st}, y_{st}; \mathbf{c}, B_{st}, n_{st})$ over B_{st}, where the total expectation of the future cost is carried over to the beginning of the horizon when the decision is made. Here, $\mathcal{W}_{st}(x_{st}, y_{st}; \mathbf{c}, B_{st}, n_{st})$ represents the cost of service s on day t and depends on the decisions x_{st} and y_{st}, cost parameters $\mathbf{c}$, the number of available anesthesiologists n_{st}, and the booked time B_{st}. The exact form of $\mathcal{W}_{st}(x_{st}, y_{st}; \mathbf{c}, B_{st}, n_{st})$ will be defined in the $DSPP_{st}$. Constraint (2) enforces the total allocation of anesthesiologists for each service, and each time period cannot be greater than the total availability of anesthesiologists on that day and for that service. Constraint (3) ensures that the decision variables are non-negative integers.

Next, we describe the second-stage problem, $DSPP_{st}$, which considers the daily decision of calling in additional anesthesiologists from the on-call consideration list to support the surgical schedule for the next day. At this point, the planner is aware of the total booked hours of surgeries for each service (B_{st}). Using this information and knowledge of the conditional distribution of the actual realization of surgery duration ($f[D_{st}|B_{st}]$), the planner decides to call in a certain number of additional anesthesiologists from the on-call consideration list (z_{st}). Each of these anesthesiologists will be paid an additional amount (c_q). On the day of surgery, the actual surgical duration of each surgery is realized, which determines the total workload for each service (D_{st}). Depending on the total available labor hours of each service ($h(x_{st} + y_{st})$), the overtime and idle time costs will be realized. The $DSPP_{st}$ is formulated as:

$\textbf{DSPP}_{st}$

$$\mathcal{W}(x_{st}, y_{st}; \mathbf{c}, B_{st}, n_{st}) = \min\Big\{\Big[c_q z_{st} + c'_q(y_{st} - z_{st} - \tau)^+\Big] + \mathbf{E}_{D_{st}|B_{st}}\Big[c_o\big(\tilde{D}_{st} - h(x_{st} + z_{st})\big)^+ + c_u\big(h(x_{st} + z_{st}) - \tilde{D}_{st}\big)^+\Big]\Big\}, \quad (4)$$

$$z_{st} \le y_{st}, \quad (5)$$

$$z_{st} \in \mathbb{Z}_0^+. \quad (6)$$

The objective (4) of the $DSPP_{st}$ consists of four terms. The first term, $c_q z_{st}$, is the cost of extra payments made to the anesthesiologists who are called from the on-call consideration list. The second term, $c'_q(y_{st} - z_{st} - \tau)^+$, is the inconvenience cost of not calling anesthesiologists from the on-call consideration list. As we described in section 1.1, these costs are incurred only if the total number of anesthesiologists not called from the on-call list $(y_{st} - z_{st})$ is greater than threshold τ. The third term $c_o(\tilde{D}_{st} - h(x_{st} + z_{st}))^+$ is the overtime pay when the demand realized is greater than the total workload available for service s. The fourth term, $c_u(h(x_{st} + z_{st}) - \tilde{D}_{st})^+$ is the cost of idle time when the demand falls short of total available work hours. For these costs, the expectation is taken over the conditional distribution of D_{st}. Note that the third and fourth terms together are the expected costs of the day of surgery and are similar to the well-known newsvendor cost (Nahmias and Cheng 2009). Constraint (5) restricts the additional number of anesthesiologists who can be called to those who are on the on-call consideration list, which is set in the first stage. Constraint (6) restricts the decision variable z_{st} to be a positive integer.

It is important to note that the staff planning model (consisting of the *MSPP* and the $DSPP_{st}$) is an aggregate planning model over a monthly horizon. Thus, we consider overtime from an aggregate perspective and ignore the bin-packing problem of scheduling surgeries and the problem of assigning anesthesiologists to individual surgeries. Olivares et al. (2008) and He et al. (2012) used similar approaches in aggregating workload by services in an operating room context. In addition, we assume overtime costs are computed on a daily basis when the workload exceeds 8 hours in a day. This was the case in our application and was also consistent with California law (https://www.shrm.org/ResourcesAndTools/tools-and-samples/how-to-guides/Pages/californiahowtocalculatedailyandweeklyovertimeincalifornia.aspx). Alternatively, even in states where daily overtime is not mandated by state law, daily overtime for physicians is often covered by employment contracts, and individual anesthesia group practices may have contracts that cover daily overtime costs (Dexter and O'Neill 2001). Furthermore, computing overtime on a daily basis is common in the literature (Dexter et al. 1999, Dexter and Traub 2002, Olivares et al. 2008). However, our approach is general and can be easily extended to other settings. For example, if overtime is calculated on a weekly basis when the weekly workload exceeds 40 hours, we would solve the second-stage problem for a week instead of a day.

Finally, we assume that the available pool of anesthesiologists is so large that individuals can reliably set up appointments far in advance and confidently know that there will be enough total people available that those appointments can be made. In addition, we assume that there are a large number of anesthesiologists who are willing to be available on-call and paid only if needed. These assumptions seem reasonable in our setting, where we consider a hospital in a large urban environment. This is analogous to models to even workload on surgical wards by adjusting the master surgical schedule. Such models generally assume an unlimited number of hospital beds (Fügener et al. 2016). However, if we need to configure our model to a finite number of anesthesiologists, we would appropriately reduce the value of the parameter n_{st} representing the number of anesthesiologists available for service s on day t. This, in turn, would require adjustment of the higher-level block schedule specifying the number of surgeries that can be performed for each specialty on a given day.

2.1. Structural Properties

In this section, we derive structural properties of the model that can be used to develop its solution method. Let $\mathcal{U}(z_{st})$ denote the objective function of the $DSPP_{st}$, where $\mathcal{U}(z_{st})$ is given as:

$$\begin{aligned}\mathcal{U}(z_{st}) = \Big\{ \Big[& c_q z_{st} + c'_q (y_{st} - z_{st} - \tau)^+ \Big] \\ & + \mathbf{E}_{D_{st}|B_{st}} \Big[c_u \big(\tilde{D}_{st} - h(x_{st} + z_{st})\big)^+ \\ & + c_o \big(h(x_{st} + z_{st}) - \tilde{D}_{st}\big)^+ \Big] \Big\}. \end{aligned} \tag{7}$$

The first proposition provides the optimal solution for the daily staff planning problem ($DSPP_{st}$).

Proposition 1. *If the distribution of $D_{st}|B_{st}$ is stochastically increasing in B_{st}, then the optimal solution for $DSPP_{st}$ is given by $z^*_{st}(x_{st}, y_{st}; B_{st})$:*

$$z^*_{st}(x_{st}, y_{st}; B_{st}) = \begin{cases} \lceil \hat{z}_{st} \rceil & \text{if } \mathcal{U}(\lceil \hat{z}_{st} \rceil) \le \mathcal{U}(\lfloor \hat{z}_{st} \rfloor) \\ \lfloor \hat{z}_{st} \rfloor & \text{otherwise,} \end{cases} \tag{8}$$

where,

$$\hat{z}_{st} = \begin{cases} 0 & \text{if } B_{st} \le B^L_{st}(x_{st}, \kappa(\mathbf{c})) \\ \frac{1}{h} F^{-1}_{D_{st}|B_{st}} \left[\frac{c_o h + c'_q - c_q}{h(c_u + c_o)} \right] - x_{st} & \text{if } B^L_{st}(x_{st}, \kappa(\mathbf{c})) \le B_{st} \le B^U_{st}(x_{st}, y_{st} - \tau, \kappa(\mathbf{c})) \\ y_{st} - \tau & \text{if } B^U_{st}(x_{st}, y_{st} - \tau, \kappa(\mathbf{c})) < B_{st} \le B^L_{st}(x_{st}, \kappa_1(\mathbf{c})), \\ \frac{1}{h} F^{-1}_{D_{st}|B_{st}} \left[\frac{c_o h - c_q}{h(c_u + c_o)} \right] - x_{st} & \text{if } B^L_{st}(x_{st}, \kappa_1(\mathbf{c})) \le B_{st} \le B^U_{st}(x_{st}, y_{st} - \tau, \kappa_1(\mathbf{c})) \\ y_{st} & \text{if } B_{st} > B^U_{st}(x_{st}, y_{st} - \tau, \kappa_1(\mathbf{c})), \end{cases} \tag{9}$$

where, $\kappa(\mathbf{c}) = \frac{c_o h + c'_q - c_q}{h(c_u + c_o)}$ *and* $\kappa_1(\mathbf{c}) = \frac{c_o h - c_q}{h(c_u + c_o)}$.

All proofs are provided in the Electronic Companion (EC.1). We describe the expressions for threshold values for the lognormal distribution (used to fit the data in the demand estimation procedure in section 3.2) in the proof of Proposition 1. This proposition implies that the number of anesthesiologists who should be called from the on-call list can be described as a threshold policy depending on the booked time information $\tilde{B}_{st}$ that is available the day before surgery. If the booked time is below $B^L_{st}(x_{st}, \kappa(\mathbf{c}))$, then the number of anesthesiologists available on regular duty (x_{st}) would be sufficient. If the booked time is above $B^U_{st}(x_{st})$, then all the anesthesiologists on the on-call consideration list would be required. For intermediate values of $\tilde{B}_{st}$, the proposition above provides for the optimal number of anesthesiologists who should be called from the on-call list.

Let $\mathcal{W}^{LP}(x_{st}, y_{st}; \mathbf{c}, B_{st}, n_{st})$ be the linear programming relaxation of $\mathbf{DSPP}_{st}$ with the integrality constraint (6) relaxed. Then we define the **MSPP′** as:

$$\begin{aligned}(\mathbf{MSPP'}) \quad & \mathcal{V}'(\mathbf{n}, \mathbf{c}) \\ & = \min \sum_{s \in S, t \in T} \Big\{ \mathbf{E}_{\mathbf{B}_{bfst}} \big[\mathcal{W}^{LP}_{st}(x_{st}, y_{st}; \mathbf{c}, B_{st}, n_{st}) \big] \Big\}, \end{aligned} \tag{10}$$

subject to,

$$(2), (3) \tag{11}$$

Since $\mathcal{W}^{LP}(x_{st}, y_{st}; \mathbf{c}, B_{st}, n_{st}) \le \mathcal{W}(x_{st}, y_{st}; \mathbf{c}, B_{st}, n_{st})$, $\mathcal{V}'(\mathbf{n}, \mathbf{c}) \le \mathcal{V}(\mathbf{n}, \mathbf{c})$. Thus, the *MSPP′* is a lower bound to the *MSPP*. The next proposition provides a property of *MSPP′* that will be used in constructing its solution method.

Proposition 2. *The MSPP′ is discretely convex in (x_{st}, y_{st}).*

2.2. Solution Method

Next, we utilize Propositions 1 and 2 to develop a computationally tractable algorithm to solve the *MSPP*. First, we solve the integer convex program *MSPP′*. To do so, we approximate the expectation in *MSPP′* by its sample average approximation (SAA) as:

$$\mathcal{V}'(\mathbf{n},\mathbf{c}) \approx \hat{\mathcal{V}}'(\mathbf{n},\mathbf{c}) = \min \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{s\in S,t\in T}\mathcal{W}_{st}^{LP}(x_{st},y_{st};\mathbf{c},b_{st}^{m},n_{st})\right] \quad (12)$$

subject to,

$$(2),\ (3). \quad (13)$$

As shown in Proposition 2, $\mathcal{W}^{LP}(x_{st},y_{st};\mathbf{c},b_{st}^{m},n_{st})$ is a discretely convex function. Therefore, the sample average approximation $\hat{\mathcal{V}}'(\mathbf{n},\mathbf{c})$ is also a discretely convex problem. We solve $\hat{\mathcal{V}}'(\mathbf{n},\mathbf{c})$ by first solving its integer relaxation, employing the subgradient method for constrained problems (Boyd and Vandenberghe 2004). As $\hat{z}_{st}=0$ is always a feasible solution to $\mathcal{W}^{LP}(x_{st},y_{st};\mathbf{c},b_{st}^{m},n_{st})$, there will always be a solution to $\mathcal{W}^{LP}(x_{st},y_{st};\mathbf{c},b_{st}^{m},n_{st})$ for every feasible (x_{st},y_{st}) at each iteration of the subgradient method. Furthermore, we stop the subgradient method when the current solution does not improve the previous best solution by a pre-specified tolerance. Let this current solution be (x_{st}^{*},y_{st}^{*}) with a corresponding objective value of $(1/M)\sum_{m=1}^{M}\sum_{s\in S,t\in T}\mathcal{W}^{LP}(x_{st}^{*},y_{st}^{*};\mathbf{c},b_{st}^{m},n_{st})$. This value is a lower bound to the *MSPP*. Then, we find the best nearest feasible integer solution $(\hat{x}_{st},\hat{y}_{st})$, and its corresponding objective value $(1/M)\sum_{m=1}^{M}\sum_{s\in S,t\in T}\mathcal{W}(\hat{x}_{st},\hat{y}_{st};\mathbf{c},b_{st}^{m},n_{st})$. This provides a heuristic solution to the *MSPP*.

Define $\hat{g}(\hat{x}_{st},\hat{y}_{st})$, an estimate of the integrality gap at $(\hat{x},\hat{y}_{st})$, as:

$$\hat{g}(\hat{x}_{st},\hat{y}_{st}) = \frac{1}{M}\sum_{m=1}^{M}\sum_{s\in S,t\in T}\mathcal{W}(\hat{x}_{st},\hat{y}_{st};\mathbf{c},b_{st}^{m},n_{st}) - \frac{1}{M}\sum_{m=1}^{M}\sum_{s\in S,t\in T}\mathcal{W}^{LP}(x_{st}^{*},y_{st}^{*};\mathbf{c},b_{st}^{m},n_{st}). \quad (14)$$

The above equation defines the integrality gap as the difference between the cost of the nearest feasible integer solution from its optimal continuous solution, averaged across M realizations of anesthesia hours by service and day. Below, we formalize the heuristic algorithm based on SAA to solve the *MSPP*.

It is apparent from the above algorithm that the MSPP is decomposable by both service and days. Thus, we could potentially solve this problem by more direct methods, such as complete enumeration of the first-stage variables. However, as discussed in the literature, these complete enumeration methods for two-stage stochastic integer programs could be computationally challenging (Schultz et al. 1998). To

Heuristic Algorithm to solve $MSPP$

1. Set $\epsilon>0$ to be sufficiently small and M to be sufficiently large.
2. For a given (s,t), draw M samples of B_{st}, represented by b_{st}^{k}, $k=1,2,\ldots,M$, from the distribution of B_{st}.
3. Solve the convex program $\hat{\mathcal{V}}'(\mathbf{n},\mathbf{c})$ by employing Proposition 2 and the subgradient method. Use Proposition 1 to compute $\mathcal{W}^{LP}(x_{st},y_{st};\mathbf{c},b_{st}^{k},n_{st})$ at each iteration of the subgradient method. Let the subgradient solution be (x_{st}^{*},y_{st}^{*}).
4. Find the best nearest feasible integer solution $(\hat{x}_{st},\hat{y}_{st})$ corresponding to the subgradient solution (x_{st}^{*},y_{st}^{*}).
5. Compute the estimate of the integrality gap $\hat{g}(\hat{x}_{st},\hat{y}_{st})$ using (14)
6. If the integrality gap, $\hat{g}(\hat{x}_{st},\hat{y}_{st})>\epsilon$, increase sample size M and go to Step 2. Otherwise, $(\hat{x}_{st},\hat{y}_{st})$ is the heuristic solution for the given (s,t). Go to Step 7.
7. Repeat Step 2 to Step 6 $\forall(s,t)$.
8. End.

test this approach in our context, we decomposed the problems by service and found that solving this problem across the four services for a given day took about an hour, and for the whole month, it took about 98 hours or more than 4 days. This seemed computationally intensive from a practical standpoint. Furthermore, the analysis in sections 5.2 through 5.5 required solving several instances of the MSPP. Thus, such complete enumeration-based methods preclude these types of analysis, which were important from a practical standpoint. In contrast, our algorithm described above, where $\varepsilon = 0.05$ and $M = 500$, solved the entire problem in less than 10 minutes in all the considered test problem instances and was within 2% of the costs of the solution obtained by the enumeration-based approach. We provide more details in the Electronic Companion (EC.3). Therefore, it seemed reasonable to employ our solution method to solve this problem and conduct the associated analysis.

Finally, it is important to note that the value of the solution using this method would naturally depend on the reliability of the cost parameters c_q, c'_q, c_o, c_u. While c_q and c_o are known, as these are actual dollar payments, the hospital makes to the anesthesiologists, c'_q and c_u are implicit. Therefore, we develop an estimation procedure to determine these costs. This procedure first requires estimating the demand distributions for each anesthesia service. Thus, in the next section, we describe our methodology to specify and estimate these distributions.

3. Estimation of Demand Distributions

Estimation of demand distribution for anesthesia services consists of two stages. First, we estimate the distribution for the booked hours for service s and day t (B_{st}). The realization $\tilde{B}_{st}$ of the distribution are the booked hours and is known the day before t. To incorporate add-ons and cancellations that may incur after the booked hours are determined and before the end of day t, we estimate $D_{st}|B_{st}$. This represents the distribution of daily anesthesia hours used for service s at the end of day t, conditional on the distribution of booked hours B_{st}.

3.1. Estimating Distribution of Booked Hours (B_{st})

Surgery requests start coming in sequentially about six months before the day of surgery. Subsequently, requests for cancellations and add-on cases keep coming in until one day before the day of surgery. While these advance bookings might be informative about the actual realization of B_{st}, other hospital departments do not pass on the information to the operating services department, as it is subject to change. Only the final booked hours for each department are sent by admissions to operating services the day before the scheduled surgeries. This implies that no advance information from early bookings is available when the *MSPP* is being solved. The information available is restricted to the day of the week, the month, and whether an upcoming day is a holiday. Therefore, we use only these variables to estimate the distribution of B_{st}. We plot the empirical distribution of the booked hours for each of the services (B_{st}) in the Electronic Companion (Figure EC.1).

From Figure EC.1, we can see that for Cardiothoracic, Neuro, and Pediatric anesthesia services, there is a concentration of data at zero. This is because these services are specialized, and they are not performed every day of the week. Meanwhile, general anesthesia service is performed almost every day, and we do not see such a concentration of data at zero. Therefore, we used a separate procedure to estimate the anesthesia required for specialized and general surgeries. We refer to these as specialized services and general services. Next, we describe the procedure to estimate the demand for these services.

Estimation of B_{st} for Specialized Services. We use a two-step estimation method to estimate the distribution of booked anesthesia hours for services such as Cardiothoracic, Neuro, and Pediatric surgeries. Duan et al. (1983) and Min and Agresti (2002) provide a more detailed description. Here, in the first step, the dependent variable is a binary outcome variable with $B_{st} = 0$, indicating there is no demand for service s on day t. Conditional on this first-stage binary variable being false (i.e., $B_{st} > 0$), we then estimate the magnitude of B_{st}.

More specifically, in the first step, the binary outcome variable B_{st} is modeled by logistic regression. The specification of this logistic regression is:

$$\begin{aligned}\text{logit}[P(B_{st}) = 0] = \alpha_{s,0} + \alpha_{s,1} \times Day\ of\ Week_t \\ + \alpha_{s2} \times Month_t + \alpha_{s3} \times Holiday_t.\end{aligned} \tag{15}$$

This can be written concisely as:

$$\text{logit}[P(B_{st}) = 0] = \boldsymbol{\alpha}'_s \mathbf{h}_t. \tag{16}$$

In the second part of the estimation procedure, we estimate the distribution of the magnitude of B_{st}, conditional on it being positive. Although the empirical distribution was the best fit, we elected to use a lognormal specification of the magnitude of B_{st} to effectively model conditional distributions. In

addition, the lognormal distribution was a better fit in comparison to other distributions such as the Weibull. Duan et al. (1983), May et al. (2000), and He et al. (2012) used a lognormal distribution for surgical services demand. This specification is:

$$\log(B_{st}|B_{st}>0) = \beta_{s0} \times Day\ of\ Week + \beta_{s1} \times Month + \beta_{s2} \times Holiday + \varepsilon_{st} \quad (17)$$

We simplify the above as:

$$\log(B_{st}|B_{st}>0) = \boldsymbol{\beta}_s'\mathbf{h}_t + \varepsilon_{st}, \quad (18)$$

where $\varepsilon_{st} \sim \mathcal{N}(0, \sigma_s^2)$. Following Duan et al. (1983) and Min and Agresti (2002), the maximum likelihood of the two-part model is given by:

$$\ell(\boldsymbol{\alpha}_s, \boldsymbol{\beta}_s, \sigma) = \ell_1(\boldsymbol{\alpha}_s)\ell(\boldsymbol{\beta}_s, \sigma), \quad (19)$$

and

$$\ell_2(\boldsymbol{\beta}_s, \sigma_s) = \prod_{B_{st}>0} \sigma_s^{-1}\phi\left(\frac{\log(B_{st}) - \boldsymbol{\beta}_s'\mathbf{h}_t}{\sigma_s}\right). \quad (20)$$

As the likelihood function is separable in the parameters, we can estimate $\boldsymbol{\alpha}_s$, $\boldsymbol{\beta}_s$, and σ by independently solving the maximum of the two likelihood functions, $\ell_1(\boldsymbol{\alpha}_s)$ and $\ell_2(\boldsymbol{\beta}_s, \sigma_s)$.

We summarize the results of the estimation procedure in the Electronic Companion (EC.4). From these results, we can conclude that the procedure is very effective in estimating B_{st} for specialized anesthesia services at the UCLA RRMC.

Estimation of $B_{s,t}$ for General Service. We can observe from Figure EC.1 that the distribution of booked anesthesia hours for general surgeries is bimodal. This is because, while general surgeries are performed on most days, there is a lower demand on weekends and holidays, while there is higher demand on regular days. Therefore, we model the distribution of anesthesia booked for general surgeries as a mixture of two Gaussian distributions. This approach for modeling bimodal distributions has been suggested by Allenby et al. (1998) for capturing a wide variety of heterogeneity in demand distributions. In Gaussian mixture models, the distribution of the mixture is given by the weighted sum of the two Gaussian distributions. Thus, the conditional distribution $g(B_{st}|\mathbf{h}_t)$ is given by:

$$g(B_{st}|\mathbf{h}_t) = \sum_{k\in\{1,2\}} \pi_k \phi_k(B_{st}|\mathbf{h}_{tk}; \boldsymbol{\beta}_k), \quad (21)$$

where π_k are weights assigned to the two-component distributions and $\phi_k(B_{st}|\mathbf{h}_{tk}; \boldsymbol{\beta}_k)$ are the two-component distributions with regression parameters $\mathbf{h}_{t1}$ and $\mathbf{h}_{t2}$, and coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. We estimate this Gaussian mixture model using the `flexmix` package in R (Grün and Leisch 2007). We summarize the results of the two-component regressions in the Electronic Companion (EC.4). Here again, these results show that this is an effective procedure to estimate B_{st} for general surgeries at the UCLA RRMC.

3.2. Estimation of $D_{st}|B_{st}$

We first used the approach outlined in Dexter and Epstein (2018) to verify that staff scheduling did not affect anesthesiologist workload. We provide more details in the Electronic Companion (EC.5). We then choose a lognormal specification for $F(D_{st}|B_{st})$, as it provides a good fit (as shown in the Electronic Companion). In addition, the lognormal specification has been used in the literature for modeling demand for surgical services (He et al. 2012, Strum et al. 1997). While the normal and Weibull distribution worked well in Strum et al. (1997), we found the lognormal distribution to be a better fit with our data. The specification of the regression model for D_{st} was:

$$\log(D_{st}) = \gamma\log(B_{st}) + \xi_s \quad \forall s \in S,\ t \in T. \quad (22)$$

Here, $\xi_s \sim \mathcal{N}(0, \sigma_s'^2)$. We present the results of the estimation of $D_{st}|B_{st}$ across each service in the Electronic Companion (EC.6). These results validate the choice of the lognormal specification to estimate $D_{st}|B_{st}$.

4. Estimation Procedure for Implicit Cost Parameters

To estimate the implicit cost parameters, we adapt the approach followed in the estimation of discrete choice models (McFadden 1974, McFadden and Manski 1981). To enable this, we assume that the staff planner does not know the numerical value of the implicit costs but is aware of the cost trade-offs when making staff planning decisions. Therefore, the planner has subconscious relative weights in mind and uses these costs imperfectly. We observe the staff planner's historical daily decisions on how many anesthesiologists were actually called from the on-call consideration list. We then employ a maximum likelihood optimization to estimate the implicit cost parameters in a manner that best explains the staff planner's decisions observed in the data. The estimation procedure for implicit cost parameters consists of the following steps:

1. We develop a decision model of the staff planner.

2. Based on this decision model, we derive the likelihood of obtaining the observed data as a function of the cost parameters.
3. Finally, we estimate the implicit cost parameters, which maximize the likelihood of observing the data.

We next describe each step in detail.

4.1. Decision Problem of Staff Planner

The literature related to operating room staff planning shows experimental evidence that operating room planners demonstrate errors and biases from the optimal solution (Wachtel and Dexter 2010). Therefore, we model the staff planner as a bounded rational decision-maker who is not a perfect optimizer but makes errors owing to the limited availability of information or because of cognitive limitations. Furthermore, consistent with quantal choice theory (McFadden 1976), we assume that when the planner faces alternative staff planning options, instead of selecting the optimal staffing plan, he or she selects better options with higher probability.

The above evidence that the staff planner is a bounded rational decision-maker precludes the use of data on the monthly decisions for estimating the cost parameters. These decisions include the number of anesthesiologists on regular duty and the number of anesthesiologists on the on-call consideration list for each specialty. This is a two-stage stochastic dynamic problem. Thus, modeling the monthly decisions of the staff planner would require a structural model of *dynamic* discrete choices. Estimating parameters in dynamic discrete choices requires the assumption that the decision-maker is a rational agent. In the literature related to the structural estimation of dynamic discrete choices, this is a standard assumption and referred to as the rational expectations assumption (Aguirregabiria and Mira 2010). Because we assume that the staff planner is not rational but is bounded rational and makes errors in staff planning, we do not assume rational expectations, and we exclude the monthly data in our estimation procedure.

Alternatively, we use data on daily decisions and the logit choice model to evaluate the probability of the staff planner selecting a certain number of anesthesiologists to call from the on-call consideration list. The logit model is suitable in our context for two reasons. First, it allows for discrete choices, such as the number of anesthesiologists. Second, it leads to an analytically tractable maximum likelihood model. Our context is similar to Su (2008), who uses the multinomial logit choice model and provides empirical evidence that a logit choice model provides a good fit for a bounded rational newsvendor.

According to the logit choice model, the probability of selecting a decision x is proportional to $e^{U(x)}$, where $U(x)$ is the utility of selecting the decision x (McFadden 1974). Consequently, if the domain of decisions is X, the probability of selecting choice x is given by:

$$p(x) = \frac{e^{U(x)}}{\sum_{x \in X} e^{U(x)}}. \tag{23}$$

Next, we use the above logit choice probability to derive the likelihood of the staff planner calling a certain number of anesthesiologists from the on-call consideration list.

4.2. Deriving the Likelihood Function for Staff Planning Decisions

For conciseness, we represent $\mathcal{U}(\mathbf{c}, z_{st}, y_{st}B_{st})$ as follows:

$$\begin{aligned}\mathcal{U}(\mathbf{c}, \tau, z_{st}, y_{st}, B_{st}) = \Big\{ &\Big[c_q z_{st} + c'_q (y_{st} - z_{st} - \tau)^+\Big] \\ &+ \mathbf{E}_{D_{st}|B_{st}}\Big[c_u\big(\tilde{D}_{st} - h(x_{st} + z_{st})\big)^+ \\ &+ c_o\big(h(x_{st} + z_{st}) - \tilde{D}_{st}\big)^+\Big]\Big\}.\end{aligned} \tag{24}$$

The utility of calling z_{st} anesthesiologists from the on-call consideration list for a given choice of cost parameter $\mathbf{c}$, threshold parameter τ, booked time B_{st} and y_{st} over all other feasible z'_{st}, is given as the negative of the cost incurred, or, $-\mathcal{U}(\mathbf{c}, \tau, z_{st}, y_{st}B_{st})$. Therefore, from Equation (23), the probability of choice z_{st} is:

$$p_{st}(\mathbf{c}, \tau, z_{st}, y_{st}B_{st}) = \frac{\exp(-\mathcal{U}(\mathbf{c}, \tau, z_{st}, y_{st}B_{st}))}{\sum_{z'_{st} \le y_{st}} \exp(-\mathcal{U}(\mathbf{c}, \tau, z'_{st}, y_{st}B_{st}))}. \tag{25}$$

Therefore, the likelihood of observing z_{st} for all s,t in the data for a given choice of $\mathbf{c}$ will be given by:

$$\mathcal{L}(\mathbf{c}) = \Pi_{s \in S} \Pi_{t \in T} p_{st}(\mathbf{c}, \tau, z_{st}, y_{st}B_{st}). \tag{26}$$

4.3. Determining Costs to Maximize the Likelihood Function

Maximizing the likelihood function, as described in Equation (26), is challenging because computing the likelihood requires the multiplication of $|S| \times |T|$ probabilities. The resultant likelihood becomes extremely small, and we run into floating-point errors when this function is maximized. In order to mitigate this, it is common practice to maximize the log-likelihood (Cameron and Trivedi 2005). Since the logarithm function is monotonically increasing, the optimal

solution will not change. The estimate of $\mathbf{c}$, which maximizes the log-likelihood, is given by:

$$\hat{\boldsymbol{c}} = \arg\max_{\mathbf{c}} \log \mathcal{L}(\mathbf{c}). \tag{27}$$

Using Equation (26), this simplifies to:

$$\hat{\boldsymbol{c}} = \arg\max_{\mathbf{c}} \sum_{s\in S, t\in T} \log\{p_t(\mathbf{c}, \tau, z_{st}, y_{st}B_{st})\}. \tag{28}$$

We first show that the above optimization problem is concave in $\mathbf{c}$ and then propose an estimation procedure.

PROPOSITION 3. $\log\mathcal{L}(\mathbf{c})$ *is concave in* $\boldsymbol{c}$.

In light of Proposition 3, a local solution of a nonlinear solver would be the global optimum. We use the nonlinear solver NLOPT (https://nlopt.readthedocs.io/en/latest/) with a Python programming interface to solve the maximum likelihood problem for a given dataset. For computational stability, during the nonlinear optimization, we normalize c_q to 1. We also employ a nonparametric bootstrap analysis for our estimation procedure. The bootstrap analysis allows us to compute an approximation of the confidence interval of the cost estimates. To perform bootstrap analysis, we follow the procedure described in (Greene 2000). We take J samples with replacement from our dataset. We compute the cost estimates for each sample by solving Equation (28) for the sampled dataset. Thus, we have J cost estimates $\{\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_J\}$. The mean of the cost estimates is given by, $\bar{\mathbf{c}} = \frac{1}{J}\sum_j \hat{\mathbf{c}}_J$, and we use the 2.5th and 97.5th percentile of these cost estimates to obtain the 95% confidence interval of the estimates. In our estimation so far, we assumed that the threshold τ was fixed. To find the best value of τ, we first calculated $\bar{\tau} = \max_{s,t}\{y_{st} - z_{zt}\}$. We then repeat the estimation procedure for implicit costs for different values of τ between 0 and $\bar{\tau}$. Finally, we choose τ as the value that maximizes the log-likelihood function defined in Equation (28). Here, we found $\tau = 1$ maximizes the log-likelihood. This means that one anesthesiologist per specialty incurs no cost per day for being on call but not actually called. We report the associated values of the implicit costs in Table 2. Note that the cost estimates are scaled so that $c_q = 1$. Additionally, we performed sensitivity analysis around the value of $\tau = 1$ by computing the out-of-sample root mean square error for each associated implicit cost parameter. We summarize this analysis in the Electronic Companion (EC.7), and validates $\tau = 1$ to be the best fit for the data.

Table 2 Maximum Likelihood Estimates of Implicit Cost Parameters

Cost parameters	Maximum likelihood estimate*	95% Confidence intervals (Bootstrap)
c'_q	1.63	(1.42, 1.83)
c_u	0.28	(0.12, 0.34)

Note: *Values scaled such that $c_q = 1$.

We observe in Table 2 that the estimated cost of not calling an anesthesiologist on the on-call consideration list is 1.63 times the cost of actually calling the anesthesiologist. This seems plausible, as the anesthesiologist loses not only the additional income from being on-call but potentially forgoes the opportunity to make income from other sources during that day. Dexter and O'Neill (2001) discuss the impact of these implicit costs of on-call staffing, but such costs have not been quantified in the literature thus far. Incorporating such implicit costs is important because not including them would lead to a longer on-call consideration list. While maintaining a longer on-call consideration list may provide the staff planner the flexibility to react to updated information without incurring supplemental financial expenses at the hospital, this would lead to more anesthesiologists being on the on-call list but not getting called. Olmstead et al. (2014) discuss the inconvenience to employees from being on the on-call list. This inconvenience could potentially lead to higher employee dissatisfaction (Gander et al. 2007). This, in turn, can lead to increased employee turnover, which could be detrimental to the hospital.

When we scale c_q to 1, the corresponding value of the explicit costs of overtime $c_o = 0.18$. This implies that the idle cost of an anesthesiologist is 1.55 times the overtime cost. This result is consistent with Olivares et al. (2008), who found that the cost of OR idle time was observed to be 60% higher than the cost of OR overtime. Our study demonstrates that a similar effect is in place for managing on-call anesthesiologists. Furthermore, given that the overtime cost is \$180 per hour, the implicit cost of idle time is \$180 × 1.55 = \$280/hour. Since the cost of idle time should be consistent with the hourly cost of regular time, we used these costs to compute the annual cost of an anesthesiologist based on our estimate. The anesthesiologists at the operating service department work seventeen 8-hour shifts per month. This implies the annual cost should be \$280/hour × 8 hours/shift × 17 shifts/month × 12 months/year = \$456,960/year. At the UCLA RRMC, this includes an overhead rate of 30% of salary to account for health and retirement benefits. This implies an annual salary of \$456,960/1.3 = \$351,507, which is close to the median anesthesiologist salary of \$433,000 at the UCLA Medical Center and \$392,000 nationwide (https://www.medscape.com/slideshow/2019-compensation-anesthesiologist-

6011324). This shows that the staff planner has a good sense of these costs and also validates the implicit cost of idle time estimated by our methodology.

To better understand how the estimates of implicit costs changed with factors such as the data time frame, day of the week, and service, we conducted additional analyses, summarized in the Electronic Companion (EC.8–EC.10). From this analysis, we can conclude that the implicit costs were quite stable and did not vary significantly with these factors. This shows that the operating services department made consistent staffing decisions, and no service was preferred over the other. This is very desirable from the perspective of staff morale.

Finally, the staff planner's problem can be broadly considered as a newsvendor problem, with overstock costs corresponding to the implicit costs of not calling an anesthesiologist from the on-call list and the costs of idle capacity. Similarly, the understock costs will be the costs of calling an anesthesiologist from the on-call list and overtime costs. Studies have shown that decision-makers exhibit systematic biases (Bostian et al. 2008, Ho et al. 2010, Schweitzer and Cachon 2000) whenever there are such newsvendor trade-offs between overstock and understock costs. One such common and well-studied bias is anchoring decisions on mean demand. This means that instead of ordering the optimal expected profit-maximizing quantity, decision-makers order a quantity between the optimal quantity and the quantity required to meet the mean demand (Bostian et al. 2008). Wachtel and Dexter (2010) also discuss a situation in which staff planners for anesthesiologists demonstrate anchoring on mean demand. As described in the Electronic Companion (EC.11), using the approach in Bostian et al. (2008), we also found evidence to indicate that the staff planner's decisions could be driven by a mean anchoring bias. Quantal choice theory has been used to explain the mean anchoring bias in several applications in operations management (Chen and Song 2019). Therefore, this provides more validation to represent the staff planner's decisions using quantal choice theory.

5. Computational Analysis

In this section, we first perform computational analysis to validate the performance of the estimation procedure described in section 4. Then we show the benefits of using the solution method described in section 2.2 over current practice. We also use our model to evaluate the impact of changes in costs, booked time variability, and the impact of hiring more anesthesiologists for particular services.

5.1. Validation of Estimated Cost Parameters

In order to validate the cost estimation procedure, we demonstrate that our model can accurately predict the decisions of the staff planner using the estimated costs. We follow a 10-fold cross-validation procedure to quantify the prediction accuracy of our model. Kohavi (1995) provide a detailed discussion of the advantages of using k-fold models for cross-validation. They propose $k = 10$ for discrete models such as the multinomial logit. In a 10-fold cross-validation approach, we divide our dataset Δ into 10 mutually exclusive subsets (folds) $\{\Delta_1, \ldots, \Delta_{10}\}$ of approximately equal size. We then use the estimation procedure (described in section 4) ten times. Each time, the cost parameters are estimated using dataset $\Delta \backslash \Delta_i$. Let these estimated parameters be $\hat{c}_i$. Next, given these estimates, we use Equation (25) to compute the predicted choice probability $\hat{p}_{st}(\hat{c}_i, z_{st}, y_{st}, B_{st})$ for each feasible z_{st} for the dataset Δ_i. Then, because the staff planner's choice is modeled as a multinomial logit, the predicted decision of the staff planner will be the decision that has the highest predicted probability. Thus, the predicted decisions for the test dataset Δ_i will be:

$$\hat{z}^i_{st} = \arg\max_{z_{st}} \{\hat{p}_{st}(\hat{c}_i, z_{st}, y_{st}, B_{st})\} \quad \forall (s,t) \in \Delta_i \forall i \in \{1, 2, \ldots, 10\}. \tag{29}$$

We compute the root mean square error (*RMSE*) of the above predicted decisions $\hat{z}^i_{st}$ with respect to the actual historical decisions of the staff planner $\tilde{z}^i_{st}$ for each of the 10 datasets Δ_i. Then, we compute the average RMSE across the 10 sets of predictions as:

$$\overline{RMSE} = \frac{1}{10}\sum_{i=1}^{10} \sqrt{\frac{\sum_{(s,t)\in\Delta_i} \left(\tilde{z}^i_{st} - \hat{z}^i_{st}\right)^2}{|\Delta_i|}}. \tag{30}$$

We also compute the accuracy of the model as the percentage of times the model predicted the correct decision. If $\hat{z}^i_{st} = \tilde{z}^i_{st}$, we denote $I_{\hat{z}^i_{st} = \tilde{z}^i_{st}} = 1$. Therefore, the accuracy for the dataset Δ_i is $acc_i = \frac{1}{|\Delta_i|}\sum_{s,t\in\Delta_i} I_{\hat{z}^i_{st} = \tilde{z}^i_{st}}$. The average of the accuracy across the 10-folds would be $\bar{acc} = \frac{1}{10}\sum_{i=1}^{10} acc_i$. We found that the estimation procedure is able to exactly predict z_{st} about 49% of the time. In addition, the error in prediction accuracy was also small, with the average RMSE around 0.48. We also calculated the mean average percentage error between the prediction and staff planner's decisions for the overtime and idle time hours. The results are summarized in the Electronic Companion (EC.12) and show the predicted and actual decisions are close.

We also modeled the staff planner's decision to determine the number of anesthesiologists called from the on-call list (z_{st}) as a linear regression of the observable characteristics, such as the number of anesthesiologists on regular duty (x_{st}), the number of anesthesiologists on the on-call list (y_{st}), and the total booked hours for surgery (B_{st}). Estimating operational parameters assuming a linear managerial decision rule has been applied previously in Foreman et al. (2010). The results, summarized in the Electronic Companion (EC.13), show that the average RMSE for the linear fit is 0.89. The logit choice model is a better fit to model the staff planner's decisions because it better captures the nonlinear dependence of z_{st} on y_{st}, x_{st}, and B_{st}. This, in turn, provides validity for the implicit cost estimation procedure described in section 4.

5.2. Comparison of Decisions and Costs with Current Practice

The current planning process to make these decisions uses an experience-based practitioner's heuristic. Such heuristics have been reported in the literature (Cardoen et al. 2010, Dexter and O'Neill 2001, Rath et al. 2017). At the hospital, we studied the practitioner's heuristic comprises of two stages. In the first stage, the practitioner makes monthly decisions by first calculating the mean and standard deviation of daily demand for a service on a given day. They do this done by using historical data for each day of a week in a given month. As per the practitioner's heuristic, the anesthesiologists on regular duty ($\tilde{x}_{st}$) are used to meet the mean daily demand. The anesthesiologists on the on-call list ($\tilde{y}_{st}$) are chosen to cover three standard deviations of the daily demand. Together, ($\tilde{x}_{st}, \tilde{y}_{st}$) constitute the decisions in the first stage of staff planning. In the second stage, once booking information for the day of surgery is available, the staff planner decides to call a certain number ($\tilde{z}_{st}$) of anesthesiologists from the on-call list previously decided. We model this second stage decision-making process in detail in section 4.1.

The practitioner's decision-making is sub-optimal for the following reasons. The first-stage decision-making does not consider the costs of these decisions and does not effectively incorporate uncertainty or the second-stage problem. In the second stage, as discussed in section 4.1, the practitioner is modeled as a bounded rational newsvendor who makes sub-optimal decisions.

We use the estimated implicit costs to fully specify the *MSPP* and $DSPP_{st}$. We can now compute the total costs of using a model-based solution and compare this to the cost incurred by current practice. When calculating the cost benefits of using the model-based solution described in section 2.2 with respect to the staff planner's actual decisions, we first define the *ex-post* cost of a decision (x_{st}, y_{st}, z_{st}) as:

$$\begin{aligned}\mathfrak{U}(x_{st}, y_{st}, z_{st}) = \Big\{ &\Big[c_q z_{st} + c'_q (y_{st} - z_{st} - \tau)^+\Big] \\ &+ \Big[c_u \big(\tilde{D}_{st} - h(x_{st} + z_{st})\big)^+ \\ &+ c_o \big(h(x_{st} + z_{st}) - \tilde{D}_{st}\big)^+\Big]\Big\} \end{aligned} \quad (31)$$

Here, $\mathfrak{U}(x_{st}, y_{st}, z_{st})$ is the cost when decisions (x_{st}, y_{st}, z_{st}) are taken for day t, and the actual realization of the total durations of surgeries of service s is $\tilde{D}_{st}$.

Let ($x^m_{st}, y^m_{st}, z^m_{st}$) be the decisions computed by the model-based solution procedure described in section 2.2 and ($\tilde{x}_{st}, \tilde{y}_{st}, \tilde{z}_{st}$) are the actual decisions of the staff planner. We employ $\mathfrak{U}(x_{st}, y_{st}, z_{st})$ to compare the benefits of the model-based solutions to the actual decisions of the staff planner by calculating the percentage relative cost improvement as:

$$\delta_{st} = 100\% \times \frac{|\mathfrak{U}(x^m_{st}, y^m_{st}, z^m_{st}) - \mathfrak{U}(\tilde{x}_{st}, \tilde{y}_{st}, \tilde{z}_{st})|}{\mathfrak{U}(\tilde{x}_{st}, \tilde{y}_{st}, \tilde{z}_{st})}. \quad (32)$$

We report the average cost improvement by service and overall average cost improvement in Table 3. This table shows the average cost savings using the model-based solution on historical data is 16.49%. In addition, we observe that the model-based solution improves costs across all the services. But, we note that there is a significant difference in cost savings across services. We found that this was due to the differences in the scale of the forecast errors between the booked anesthesiology hours (B_{st}) and the used anesthesiology hours (D_{st}) between services. When these errors were small, the cost savings between the practitioner's heuristic and our methods were small. However, when these errors were large, the cost savings were much higher, as our methods were more suited to deal with such errors. We provide more details in the Electronic Companion (EC.14). We also developed two benchmark models to better assess the performance of the model-based solution. These are

Table 3 Daily Average Percent Cost Saving of Model Based Solution Over Current Practice

Service	Daily average cost saving (%)	95% Confidence interval
Cardiothoracic	8.59	(6.76, 10.03)
General	15.02	(12.43, 19.02)
Neuro	28.88	(21.88, 34.27)
Pediatric	18.98	(15.76, 21.4)
Average	16.49	(14.87, 19.03)

Table 4 Breakdown of Cost Improvement for Current Practice, Benchmark Models, and Model-based Heuristic

	Current practice	Benchmark Model 1	Benchmark Model 2	Model based solution
Average cost of overtime	12,197	14,984	13,833	9257
Average cost of calling anesthesiologists	2770	586	973	3220
Average explicit costs	14,967	15,570	14,806	12,477
Average cost of not calling anesthesiologists	6716	486	671	4385
Average idle costs	8204	12,854	12,381	6919
Average implicit cost	14,920	13,340	13,052	11,304
Total average cost	29,886	28,910	27,858	23,781
Average annual total cost	10,639,558	10,291,960	9,917,448	8,466,000
% daily average cost savings from current practice	—	2.61%	5.48%	16.49%

Table 5 Comparison of Staffing Plan of Current Practice, Benchmark Models, and the Model-Based Heuristic

	Current practice	Benchmark Model 1	Benchmark Model 2	Model based solution
Average daily overtime (hours)	67.76	83.24	76.85	51.43
Average daily idle time (hours)	29.3	30.48	29.35	24.71
Average number of anesthesiologists on regular duty	17.48	20.85	18.79	18.46
Average number of anesthesiologists on on-call consideration list	6.89	1.04	1.77	5.91
Average number of anesthesiologists called	2.77	0.586	0.973	3.22
Average number of anesthesiologists not called	4.12	0.299	0.671	2.69
Percentage of days with no on-call consideration list	31.35	7.85	4.86	56.22

summarized in the Electronic Companion (EC.15). The first benchmark model considers only explicit costs and constraints to ensure that that the number of on-call positions is lower than historical averages and the call-in rate is higher than historical averages. The second benchmark model also considers only explicit costs and a cost for schedule variability. As shown in Table 4, the costs of the model-based solution significantly improved upon both these models reaffirming the value of our approach.

To better understand the reasons for this improvement, we compared the model-based solution with the benchmark models and the staff planner's plan in more detail. We summarize the results in Table 5 and show that the model-based solution has the lowest average daily overtime and idle time. This is because the algorithm employed to solve the MSPP optimally chooses x_{st} and y_{st} to minimize total expected costs. More specifically, since regular staffing has lower costs than on-call staffing, the model-based solution and the benchmark solutions have higher regular staff (x_{st}) and lower on-call staff (y_{st}) than current practice. This is also shown on Table 5. Also, observe from this table that, on average, the model-based solution uses more anesthesiologists from the on-call consideration list. While this allows for greater flexibility to react to the uncertainty in the booked time (B_{st}), there are costs to having more flexibility. However, the model-based solution still manages to reduce overall costs because it creates an on-call consideration list for fewer days.

Additionally, we assessed the impact of the solutions provided by current practice, the better performing second benchmark model, and the model based heuristic on the anesthesiologist population at an individual level. To do so, we computed p, the average fraction of time an anesthesiologist is called in after being on the on-call list using the algorithm described in the Electronic Companion (EC.16). We summarize the results in Table 6. We then calculated $C(p)$, the coefficient of variation of p_i across all the i anesthesiologists. This will represent a measure variability in outcomes (for example, the variability in call/no-call) across the anesthesiologist population. We shows these results in Table 7. These tables show that the performance of the model-based solution is very comparable to the benchmark model in terms of variability in the fraction of time called, and both of them outperform current practice, providing a more stable and less variable schedule at the individual level. This is important to verify before the implementation of any model-based aggregate staffing plan.

We also analyzed the solution of the model-based heuristic by service the percentage of days when there were no on-call consideration lists and when physicians were not called. These results are summarized in the Electronic Companion (EC.17). The results here show services that have the least reduction in the coefficient of variation when we update the demand distribution of used anesthesiology hours will get the least benefit from using an on-call staffing plan. Thus, it will be beneficial to staff these services using regular shifts.

Table 6 p **for Current Practice, Benchmark Model 2, and Model-Based Solution**

Service	p in current practice	p in Benchmark Model 2	p in Model-based solution
Cardiothoracic	0.378	0.285	0.340
General	0.410	0.914	0.894
Neuro	0.435	0.418	0.333
Pediatric	0.450	0.324	0.384
Average	0.410	0.730	0.684

Table 7 $C(p)$ **for Current Practice, Benchmark Model 2, and Model-Based Solution**

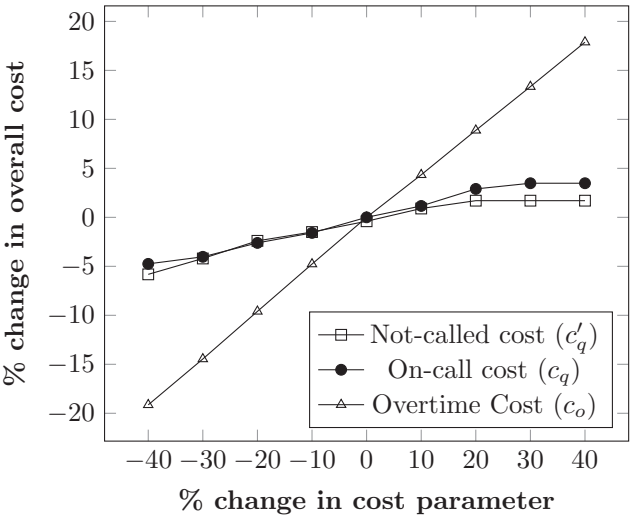
Service	$C(p)$ in current practice	$C(p)$ in Benchmark Model 2	$C(p)$ in model-based solution
Cardiothoracic	0.312	0.256	0.273
General	0.420	0.149	0.157
Neuro	0.214	0.105	0.180
Pediatric	0.118	0.074	0.104
Average	0.349	0.134	0.166

Finally, for this model to be accepted by the anesthesiologists, it is important it captures the implicit costs considered by the staff planner, and these costs have to be consistent with past practice. Our maximum likelihood procedure estimates these costs from the past decisions of the staff planner. This provides reassurance to the anesthesiologists that we have not only captured implicit costs but have estimated their value based on past decisions that were acceptable to them. In addition, the optimization approach more precisely balances the implicit and explicit costs, which leads to lower total costs. As noted above, compared to past practice, our model reduces total costs on average by 16.49%. Here, the reduction in explicit cost was 11.19%. In addition, the reduction in implicit cost was 35.12%, which was greater than the percentage reduction in total costs. Thus, the model-based solution should be at least as acceptable as the solution provided by the staff planner.

5.3. Impact of Changes in Cost

Anesthesiologists are among the most expensive labor categories in the United States, and the mean annual wage has undergone an increase of 14% between 2016 and 2017 (Bureau of Labor Statistics 2018). Increases in salaries imply a proportional increase in on-call (c_q) and overtime payments (c_o). Our model-based solution allows us to evaluate the impact of these cost increases. In Figure 1, we plot the impact of the change in on-call and overtime costs. From this figure, as expected, we can see that the total cost increases with the on-call and overtime costs. However, we can also observe that on a percentage basis, the overall cost is more sensitive to changes in the overtime cost than the on-call cost. This is because overtime costs

Figure 1 Impact of Change in Cost Parameters



are incurred on more days than on-call costs. Thus, a percentage change in overtime cost leads to a greater relative change in the overall cost. We also observed how the solution changed when we ran the model for lower values of c_q shown in Figure 1. Here we found that with decreasing c_q, the optimal solution decreases the number of anesthesiologists on regular duty (x_{st}) while increasing the number of anesthesiologists on the on-call consideration list (y_{st}) and the number of anesthesiologists actually called (z_{st}). We also found that the rate of increase in z_{st} was higher than the rate of increase in y_{st}. Conversely, when we ran the model for higher values of c_q, x_{st} increased, while y_{st} and z_{st} decreased. Detailed results are provided in the Electronic Companion (EC.18).

We also considered the impact of changes in c'_q, the cost of not calling an anesthesiologist from the on-call list on overall costs. These results are also presented in Figure 1 and show that overall costs are least sensitive to these costs. This is because for these costs to incur, an on-call list needs to be generated. As indicated in Table 4, this does not happen on 56% of the days. Even when this list is generated, one needs to exceed a threshold τ of anesthesiologists not-called from the on-call list before these costs are accrued.

We also performed additional analysis to understand the impact of cross-training anesthesiologists, changing the variability of booked time (B_{st}), and changing the number of available anesthesiologist by service (n_{st}). We describe these in detail in sections EC.19, EC.20, and EC.21 in the Electronic Companion.

6. Conclusions

In this study, we consider the anesthesiologist staffing problem typically found in large multi-specialty

hospitals with no limit on the supply of anesthesiologists. Furthermore, these anesthesiologists are willing to be available on-call and paid only if needed, learning the previous day. In this problem, the planner makes monthly and daily staffing decisions about the number of anesthesiologists across each service to minimize overall costs. We model the staff planning problem as a two-stage integer stochastic dynamic program, provide its structural properties, and use this to develop a sample average approximation-based algorithm to solve this problem.

While some of the cost components of this model are explicitly known, other cost components are implicit. We assume that the staff planner is aware of the trade-offs between explicit and implicit costs but is not a perfect optimizer and makes errors in decisions. To capture this, we develop a decision model of a bounded rational staff planner. Using this decision model and available historical data of decisions taken by the staff planner, we estimate the implicit costs. This leads to a fully specified model of staff planning. We then compare the costs of the model-based solution with the costs resulting from the historical decisions of the staff planner. Based on this analysis, we find that our approach can potentially save around 16% in costs, which translates to a total of about $2.17 million on an annual basis in explicit and implicit costs.

In addition, the estimated costs and the optimization model have generated several managerial insights. First, the cost of not calling an anesthesiologist on the call list is significantly more expensive than actually calling the anesthesiologist. This implies that staff planners need to effectively incorporate these costs when constructing on-call lists. Second, the costs of idle time are substantially higher than the costs of overtime. Thus, it is important for staff planners to consider this aspect when determining how many anesthesiologists they need to call from regular duty. Together, the first two insights suggest that it is important to have a data-based understanding of implicit costs in order to make effective staff planning decisions. Third, average daily idle time and overtime costs can be reduced by ensuring that the optimal number of total anesthesiologists are available on the day of the surgery. Furthermore, it may be efficient to have more anesthesiologists on the on-call consideration list, as long as the days requiring an on-call list are chosen carefully. The model-based approach outperforms the current practice as it makes these decisions more effectively. Fourth, our analysis summarized in the Electronic Companion (EC.20) showed that a small reduction in demand variability could considerably reduce costs. Such variance reduction could be achieved by earlier and more timely sharing of demand information between other hospital departments and operating services. Fifth, we show in the Electronic Companion (EC.21) that the marginal benefits of hiring across specialties are notably different. A good understanding of these differences using a data-driven analytical model can reduce overall staffing costs.

Our study has the following limitations. First, it is possible that there is some unobserved heterogeneity across individual anesthesiologists, depending on seniority or other factors. Some anesthesiologists may bear a higher cost of not getting called or have costlier idle time. While it is possible to incorporate this heterogeneity and estimate the different costs across the individual anesthesiologists, we were restricted by our lack of data availability at the individual anesthesiologist level. Second, in the current staffing plan, schedulers adjust the monthly plan only once, and they do this the day before the surgery. However, it may be possible to update the staff planning when each elective procedure is booked. This has been suggested by Tiwari et al. (2014) and Xie and Zenios (2015). In such a dynamic schedule updating framework, there will also be implicit costs. Our procedure can potentially be extended to evaluate these implicit costs. However, we could not to perform this analysis because the UCLA RRMC only recorded the booking data when it was finalized, the day before the procedures. Third, this work does not consider the next stage that determines the work schedules for each individual anesthesiologist either for a week or month and deciding which particular anesthesiologist will be scheduled to work on regular duty or placed on the on-call list. This could necessitate changes in the aggregate schedule provided by the model. In such situations, the model solution could overestimate the true cost savings. Finally, our analysis on the impact of hiring anesthesiologists by service is restricted to the costs considered in the model. However, there could be additional costs of hiring anesthesiologists, such as recruitment costs, bonuses, and on-boarding costs. Furthermore, the decision to hire anesthesiologists specialized in certain services would depend on the hospital's longer-term strategy of attracting demand for certain kinds of procedures or hiring faculty physicians of certain services to meet teaching requirements at the medical school. Since we did not have information on these aspects and the additional costs, we were unable to conduct a more comprehensive and longer-term analysis to determine the right sizing of the anesthesiology staff by service.

This study opens up several opportunities for future research. First, we could extend this framework to other industries outside of healthcare. While this study adds to the evidence that idle time is considered more expensive in the healthcare context, it is not obvious whether that is true for other industries like retail, call centers, and airlines that have

overtime, on-call, and idle-time costs. Second, as described above, we can extend our framework to the context of dynamic staff planning, where staff planning has more than two stages. However, this will require significant modifications to the model and solution procedure.

In conclusion, we believe that the methods presented in this study provide an effective way to estimate implicit costs and to conduct optimized staff planning.

Acknowledgments

We are indebted to Dr. Aman Mahajan, former director of Perioperative Services at the UCLA Medical Center for his support during this project. We also thank the department editor Professor Chelliah Sriskandarajah, the senior editor, and three anonymous referees for their excellent comments during the review process.

References

Aguirregabiria, V. 1999. The dynamics of markups and inventories in retailing firms. *Rev. Econ. Stud.* **66**(2): 275–308.

Aguirregabiria, V., P. Mira. 2010. Dynamic discrete choice structural models: A survey. *J. Econom.* **156**(1): 38–67.

Aksin, Z., B. Ata, S. M. Emadi, C. Su. 2017. Impact of delay announcements in call centers: An empirical approach. *Oper. Res.* **65**(1): 242–265.

Allenby, G. M., N. Arora, J. L. Ginter. 1998. On the heterogeneity of demand. *J. Mark. Res.* **35**(3): 384–389.

Allon, G., A. Federgruen, M. Pierson. 2011. How much is a reduction of your customers' wait worth? An empirical study of the fast-food drive-thru industry based on structural estimation methods. *Manuf. Serv. Oper. Manag.* **13**(4): 489–507.

Bansal, S., M. Nagarajan. 2017. Product portfolio management with production flexibility in agribusiness. *Oper. Res.* **65**(4): 914–930.

Bard, J. F., H. W. Purnomo. 2005. Hospital-wide reactive scheduling of nurses with preference considerations. *IIE Trans.* **37**(7): 589–608.

Bernstein, E., S. Kesavan, B. Staats. 2014. How to manage scheduling software fairly. *Harv. Bus. Rev.* **92**(12). https://hbr.org/2014/09/how-to-manage-schedulingsoftware-fairly (accessed date September 12, 2019).

Birge, J. R. 1985. Decomposition and partitioning methods for multistage stochastic linear programs. *Oper. Res.* **33**(5): 989–1007.

Bostian, A. A., C. A. Holt, A. M. Smith. 2008. Newsvendor "pull-to-center" effect: Adaptive learning in a laboratory experiment. *Manuf. Serv. Oper. Manag.* **10**(4): 590–608.

Boyd, S., L. Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press, Cambridge.

Brunner, J. O., J. F. Bard, R. Kolisch. 2009. Flexible shift scheduling of physicians. *Health Care Manag. Sci.* **12**(3): 285–305.

Bureau of Labor Statistics. 2018. Highest Paid Occupations. Available at https://www.bls.gov/ooh/highest-paying.htm (accessed date August 13, 2018).

Cameron, A. C., P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press, Cambridge.

Cardoen, B., E. Demeulemeester, J. Beliën. 2010. Operating room planning and scheduling: A classification scheme. *Int. J. Health Manag. Inf.* **1**(1): 71–83.

Chen, Y., Y. Song. 2019. Quantal theory in operations management. S. Villa, G. Urrea, J. A. Castañeda, E. R. Larsen, eds. *Decision-Making in Humanitarian Operations*. Springer, Berlin, 169–191.

Deshpande, V., M. Arikan. 2012. The impact of airline flight schedules on flight delays. *Manuf. Serv. Oper. Manag.* **14**(3): 423–440.

Dexter, F., R. H. Epstein. 2018. Influence of annual meetings of the American Society of Anesthesiologists and of large national surgical societies on caseloads of major therapeutic procedures. *J. Med. Syst.* **42**(12): 259.

Dexter, F., L. O'Neill. 2001. Weekend operating room on call staffing requirements. *AORN J.* **74**(5): 664–671.

Dexter, F., A. Macario, L. O'Neill. 1999. A strategy for deciding operating room assignments for second-shift anesthetists. *Anesth. Analg.* **89**(4): 920.

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

Dexter, F., J. Ledolter, R. E. Wachtel. 2005. Tactical decision making for selective expansion of operating room resources incorporating financial criteria and uncertainty in subspecialties' future workloads. *Anesth. Analg.* **100**(5): 1425–1432.

Duan, N., W. G. Manning, C. N. Morris, J. P. Newhouse. 1983. A comparison of alternative models for the demand for medical care. *J. Bus. Econ. Stat.* **1**(2): 115–126.

Easton, F. F. 2014. Service completion estimates for cross-trained workforce schedules under uncertain attendance and demand. *Prod. Oper. Manag.* **23**(4): 660–675.

Fisher, M., K. Rajaram, A. Raman. 2001. Optimizing inventory replenishment of retail fashion products. *Manuf. Serv. Oper. Manag.* **3**(3): 230–241.

Foreman, J., J. Gallien, J. Alspaugh, F. Lopez, R. Bhatnagar, C. C. Teo, C. Dubois. 2010. Implementing supply-routing optimization in a make-to-order manufacturing network. *Manuf. Serv. Oper. Manag.* **12**(4): 547–568.

Fügener, A., G. M. Edenharter, P. Kiefer, U. Mayr, J. Schiele, F. Steiner, R. Kolisch, M. Blobner. 2016. Improving intensive care unit and ward utilization by adapting master surgery schedules. *A & A Case Rep.* **6**(6): 172–180.

Gander, P., H. Purnell, A. Garden, A. Woodward. 2007. Work patterns and fatigue-related risk among junior doctors. *Occup. Environ. Med.* **64**(11): 733–738.

Greene, W. H. 2000. *Econometric Analysis*, 4th edn. International edition. Prentice Hall, New Jersey, 201–215.

Grün, B., F. Leisch. 2007. Fitting finite mixtures of generalized linear regressions in R. *Comput. Stat. Data Anal.* **51**(11): 5247–5252.

Gurnani, H., C. S. Tang. 1999. Note: Optimal ordering decisions with uncertain cost and demand forecast updating. *Management Sci.* **45**(10): 1456–1462.

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

Healthcare Insights. 2014. The evidence is clear: Analytics key to controlling labor costs inefficiency is no longer an option. *White Paper*.

Ho, T. H., N. Lim, T. H. Cui. 2010. Reference dependence in multilocation newsvendor models: A structural analysis. *Management Sci.* **56**(11): 1891–1910.

Kantor, J. 2015. Starbucks to revise policies to end irregular schedules for its 130,000 baristas. *The New York Times*.

Kesavan, S., B. R. Staats, W. Gilland. 2014. Volume flexibility in services: The costs and benefits of flexible labor resources volume flexibility in services. *Management Sci.* **60**(8): 1884–1906.

Kim, K., S. Mehrotra. 2015. A two-stage stochastic integer programming approach to integrated staffing and scheduling with application to nurse management. *Oper. Res.* **63**(6): 1431–1451.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *ICJAI*. **14**(2): 1137–1145.

Kong, N., A. J. Schaefer, S. Ahmed. 2013. Totally unimodular stochastic programs. *Math. Program.* **138**(1–2): 1–13.

May, J. H., D. P. Strum, L. G. Vargas. 2000. Fitting the lognormal distribution to surgical procedure times. *Decis. Sci.* **31**(1): 129–148.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. P. Zaremka, ed. *Frontiers in Econometrics*. Academic Press, New York, NY, 105–142. ISBN 0-12-776150-0.

McFadden, D. L. 1976. Quantal choice analysis: A survey. *Ann. Econ. Soc. Meas.* **5**(4): 363–390.

McFadden, D. L., C. F. Manski. 1981. *Econometric Models of Probabilistic Choice*. MIT Press, Cambridge, MA.

McIntosh, C., F. Dexter, R. H. Epstein. 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. *Anesth. Analg.* **103**: 1499–1516.

Min, Y., A. Agresti. 2002. Modeling nonnegative data with clumping at zero: A survey models for semicontinuous data. *JIRSS* **1**(May): 7–33.

Nahmias, S., Y. Cheng. 2009. *Production and Operations Analysis*, Vol. 4. McGraw-Hill/Irwin, New York.

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

Olmstead, J., D. Falcone, J. Lopez, L. Mislan, M. Murphy, T. Acello. 2014. Developing strategies for on-call staffing: A working guideline for safe practices. *AORN J.* **100**(4): 369–375.

Pinker, E. J., R. C. Larson. 2003. Optimizing the use of contingent labor when demand is uncertain. *Eur. J. Oper. Res.* **144**(1): 39–55.

Rath, S., K. Rajaram, A. Mahajan. 2017. Integrated anesthesiologist and room scheduling for surgeries: Methodology and application. *Oper. Res.* **65**(6): 1460–1478.

Rogers, A. E., W. T. Hwang, L. D. Scott, L. H. Aiken, D. F. Dinges. 2004. The working hours of hospital staff nurses and patient safety. *Health Aff.* **23**(4): 202–212.

Rust, J. 1987. Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica* **55**(5): 999–1033.

Schultz, R., L. Stougie, M. H. Van Der Vlerk. 1998. Solving stochastic programs with integer recourse by enumeration: A framework using Gröbner basis. *Math. Program.* **83**(1–3): 229–252.

Schweitzer, M. E., G. P. Cachon. 2000. Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Sci.* **46**(3): 404–420.

Sisko, A. M., S. P. Keehan, J. A. Poisal, G. A. Cuckler, S. D. Smith, A. J. Madison, K. E. Rennie, J. C. Hardesty. 2019. National health expenditure projections, 2018–27: Economic and demographic trends drive spending and enrollment growth. *Health Aff.* **38**(3): 491–501.

Slaugh, V. W., A. A. Scheller-Wolf, S. R. Tayur. 2018. Consistent staffing for long-term care through on-call pools. *Prod. Oper. Manag.* **27**(12): 2144–2161.

Smith, M., R. Saunders, L. Stuckhardt, J. M. Mcginnis. 2012. *Best Care at Lower Cost the Path to Continuously Learning Health Care in America Committee on the Learning Health Care System in America*. National Academic Press, Washington, DC.

Stimpfel, A. W., D. M. Sloane, L. H. Aiken. 2012. The longer the shifts for hospital nurses, the higher the levels of burnout and patient dissatisfaction. *Health Aff.* **31**(11): 2501–2509.

Strum, D. P., L. G. Vargas, J. H. May, G. Bashein. 1997. Surgical suite utilization and capacity planning: A minimal cost analysis model. *J. Med. Syst.* **21**(5): 309–322.

Su, X. 2008. Bounded rationality in newsvendor models. *Manuf. Serv. Oper. Manag.* **10**(4): 566–589.

Sun, R. R., O. V. Shylo, A. J. Schaefer. 2015. Totally unimodular multistage stochastic programs. *Oper. Res. Lett.* **43**(1): 29–33.

Tiwari, V., W. R. Furman, W. S. Sandberg, 2014. Predicting case volume from the accumulating elective operating room schedule facilitates staffing improvements. *Anesthesiology* **121**(1): 171–183.

Trinkoff, A. M., R. Le, J. Geiger-Brown, J. Lipscomb, G. Lang. 2006. Longitudinal relationship of work hours, mandatory overtime, and on-call to musculoskeletal problems in nurses. *Am. J. Ind. Med.* **49**(11): 964–971.

Wachtel, R. E., F. Dexter. 2010. Review of behavioral operations experimental studies of newsvendor problems for operating room management. *Anesth. Analg.* **110**(6): 1698–1710.

Wild, B., C. Schneewei. 1993. Manpower capacity planning—A hierarchical approach. *Int. J. Prod. Econ.* **30**: 95–106.

Xie, S., S. A. Zenios. 2015. Forecasting and dynamic adjustment of staffing levels in hospital operating rooms. Working paper, Stanford Graduate School of Business.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

EC.1: Proofs of Propositions.
EC.2: Figures.
EC.3: Computational Analysis for Performance of Proposed Heuristic Method.
EC.4: Estimation Results for Distribution of Booked Hours by Service (B_{st}).
EC.5: Analysis of Schedule Impact on Anesthesiologist Workload.
EC.6: Estimation Results for Distribution Hours by Service Conditioned on Booked Hours ($D_{st}|B_{st}$).
EC.7: Sensitivity with Respect to τ.
EC.8: Estimation and Validation Results for Model with Different Implicit Costs for Each Service.
EC.9: Estimation Results for Model with Different Implicit Costs for Two Halves of the Data Set.
EC.10: Estimation Results for Model with Different Implicit Costs for Each Day of Week.
EC.11: Evidence of Mean Anchoring Bias.
EC.12: Predictive Performance of Decision Model on Outcome Variables: Overtime and Idle Time Hours.
EC.13: Modeling Decision Model of Staff Planner as a Linear Decision Rule.
EC.14: Analysis of Differences in Percentage Savings Across Services.
EC.15: Benchmark Models.
EC.16: Algorithm to Compute p.
EC.17: Analysis of Differences in On-Call Use Across Services.
EC.18: Solution Changes with Changes in On-Call Costs.
EC.19: Impact of Cross-Training Anesthesiologists.
EC.20: Impact of Changes in Booked Time Variability.
EC.21: Impact of Hiring Anesthesiologists by Service.