Effectiveness of Surgical Scheduling Optimization: a Reinvestigation under the “To-Follow” Practice and Perioperative Uncertainties

Miao Bai Ph.D.
Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, bai.miao@mayo.edu

Robert H. Storer Ph.D., Gregory L. Tonkay Ph.D.
Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA 18015, rhs20@lehigh.edu, glk0@lehigh.edu

Terril E. Theman M.D.
Healthcare Systems Engineering, Lehigh University, Bethlehem, PA 18015, tet212@lehigh.edu

We reinvestigate the effectiveness of surgical scheduling optimization in scheduling practice as we observed in hospitals and learned from interviews with practitioners. We model the practice of surgical schedule construction and real-time schedule adjustment and account for commonly-encountered perioperative uncertainties. Moreover, we include the widely-implemented “to-follow” practice that contradicts common assumptions found in previous literature. To solve the resulting scheduling optimization problems of practical complexity, we propose a gradient descent solution algorithm based on recursive sensitivity analysis, which outperforms benchmark methods and identifies near-optimal solutions.

In numerical experiments, we illustrate that the “to-follow” practice can mitigate the uncertainties intrinsic to surgical suite operations. Previous surgical scheduling researchers did not include the ubiquitous “to-follow” practice that modern hospitals use, an oversight that may significantly reduce the efficacy of these prior recommendations. We also show that it is preferable to schedule a single surgeon for the whole day in the operating room (OR) when the “to-follow” practice is implemented. The value of real-time schedule adjustment optimization, which can reduce chaos and confusion in current practice, is more significant in surgical suites with high uncertainty and frequent surgeon changes in ORs. We also provide insights into the time when patients are requested to arrive for surgery and the frequency of real-time schedule adjustment in current practice.

Subject classifications:
Area of review: Healthcare Management
1. Introduction

Central to surgical patient flow in a hospital, surgical suites are the most capital-intensive medical resources and their operations have substantial impacts on other medical departments. Ineffective management of surgical suites may result in suboptimal patient care, wasteful utilization of expensive medical resources and declining financial performance (Erdogan and Denton 2010). However, due to the unpredictable nature of perioperative operations, surgical suite management is challenging in practice. Therefore sophisticated strategies, particularly mathematical approaches, have been studied to improve the utilization of these costly medical resources (Gupta 2007).

Among a variety of mathematics-based management strategies, surgical scheduling optimization has been a research focus for decades and has been shown to be capable of improving surgical suites utilization and enhancing the patient experience (Cardoen et al. 2010). Surgical scheduling optimization studies typically follow a modeling-optimization paradigm. It starts with mathematically modeling the surgical practice and subsequently optimizes scheduling decisions in the model, which would work well practically as long as the model accurately reflects the practice.

In this study, by following the same modeling-optimization paradigm, we reinvestigate the effectiveness of surgical scheduling optimization in the real world of surgery. To reflect the scheduling practice in modern surgical suites, we conducted observations and interviewed practitioners to learn current practice at multiple U.S. hospitals, including national academic medical centers, regional hospitals and specialized ambulatory surgical centers. Based on what we have observed, we were able to model the surgical scheduling mechanism and study the effectiveness of surgical scheduling optimization.

Modeling

We model behaviors that have not been discussed in previous studies. First and most importantly, we model the ubiquitous “to-follow” practice that modern hospitals use. In the “to-follow” practice, patients are requested to arrive well before the scheduled start time $T_{ss}$ of their surgeries. After the patient is prepared, the surgery begins as soon as the surgical team is ready and the OR is available. One surgical case follows another according to the schedule. However if a case finishes ahead of its predicted time allocation, or if there is a last-minute case cancellation, then other scheduled cases are moved up on the schedule “to follow” the finished procedure as soon as possible and often ahead of the $T_{ss}$. This contrasts with common assumptions in previous studies that no surgery can start before $T_{ss}$ (e.g. Robinson and Chen 2003, Denton and Gupta 2003, Mancilla and Storer 2012, Lee and Yih 2014, Bai et al. 2017).

Second, we model 3 sets of surgical scheduling decisions that operating room (OR) managers make on a daily basis. Prior to the day of surgery, OR managers construct surgical schedules, which include determination of $T_{ss}$ and the time when patients are requested to arrive at hospitals ($T_{eta}$). As the surgery day unfolds, real-time adjustments are typically done based on experience by
the OR manager to alleviate deviation from the surgical schedule. Most existing studies focus on $T_{ss}$ but few discuss $T_{sta}$ or real-time adjustments (see literature review in §2). To our knowledge, we are the first to study all 3 sets of decisions in surgical scheduling optimization. Moreover, in real-time schedule adjustment, we model the rescheduling paradigm and “advanced notice time” requirements used in practice.

Third, we account for the practice of surgeons sharing ORs and various commonly-encountered perioperative uncertainties. The OR sharing practice has a substantial impact on surgical scheduling since surgeons must be given a start time for their first surgery and this surgery cannot begin until surgeons are ready to proceed. Perioperative uncertainties radically complicate surgical scheduling and may result in inefficient resource utilization and ultimately in a loss of revenue (Gupta 2007). We model common perioperative uncertainties, including patient availability uncertainty, emergency surgeries, surgery cancellations, recovery resource availability and duration uncertainty in preoperative operations, surgeries and postoperative recovery. Previous studies only consider some of these (see §2).

**Optimization** With practical factors in our modeling of the surgical practice, the scheduling optimization problem becomes extremely complex. Perioperative operations, including preoperative preparation, surgery and postoperative recovery, are highly variable in duration (May et al. 2011). Surgical emergencies and surgery cancellation bring uncertainty to the number of cases that can be performed. Recovery resources are shared among streams of patients coming from different ORs, which complicates the problem with additional coordinating decisions (Erdogan and Denton 2010, Bai et al. 2017). Furthermore, compared to previous studies, our problem becomes more complicated with additional decisions on requested patient arrival time and consideration of the “to-follow” practice.

Surgical scheduling optimization of practical complexity has remained an open challenge (Erdogan and Denton 2010, Cardoen et al. 2010, May et al. 2011). We notice that a stream of studies (e.g. Kim et al. 2015, van Ryzin and Vulcano 2008) have successfully solved stochastic optimization problems by recursive sensitivity analysis and gradient-descent algorithms. Notably, similar solution methodologies have been recently applied to surgical scheduling problems considering duration uncertainty and recovery resource constraints (Zhang and Xie 2015, Bai et al. 2017). Our problem is more complex because of additional practical considerations, but previous studies inspired us at the early stage of this research.

Following the same algorithm paradigm, we develop a gradient descent algorithm, where a sample-path gradient is derived by an analytically complicated, but computationally efficient recursive sensitivity analysis. Our algorithm identifies near-optimal solutions and outperforms benchmark methods found in previous studies.
**Findings** Based on numerical results, we provide insights into surgical scheduling optimization for both research studies and practical usage.

1. Surgical suite operations are subject to much uncertainty. Implementing the “to-follow” practice is an effective way to mitigate this uncertainty. Perhaps this is why the “to-follow” practice is implemented in virtually all surgical suites.

2. Most previous studies do not consider the “to-follow” practice but rather assume no surgery may start until the scheduled start time. Based on this assumption, surgical scheduling optimization is shown to be effective in cost saving and performance improvement. However, when the “to-follow” practice is implemented, these hypothetical gains could quickly disappear.

3. It is preferable to schedule a single surgeon for the whole day in the OR. Scheduling multiple surgeons into a single OR can blunt the benefits of the “to-follow” practice since the first surgery of incoming surgeons cannot begin until surgeons are ready to operate.

4. Real-time schedule adjustment in the current surgical practice takes most of the OR managers time and can cause significant chaos and confusion. More formal optimization algorithms can perform this task quite well and save significant time and money. The value of schedule adjustment increases in surgery suites with frequent surgeon changes in ORs and high uncertainty (especially with regard to emergency and cancellation rates), which may make the algorithm used to construct the schedule in the first place seems to be of little consequence.

**Paper Organization** The rest of the paper is organized as follows. Relevant literature is reviewed in §2 and surgical scheduling practice is detailed as we observed in §3. Methodologies for mathematical modeling and scheduling optimization are presented in §4. Numerical results and discussions are provided in §5 with conclusions and future research discussed in §6.

2. Related Literature
Surgery-to-OR allocation and surgery sequencing are determined jointly by OR managers and surgeons based on a variety of factors which are considered out of the scope of this study. Reviews on general OR management can be found in (Cayirli and Veral 2003, Gupta 2007, Erdogan and Denton 2010, May et al. 2011).

**Surgical Schedule Construction** The majority of surgical scheduling work focuses on surgical schedule construction which considers perioperative uncertainties. Two statistical studies (Dexter and Traub 2000, Wachtel and Dexter 2007) investigate when a patient should be ready for surgery so that medical resources are not left idle. Many studies have considered surgical duration uncertainty (e.g. Robinson and Chen 2003, Denton and Gupta 2003, Mancilla and Storer 2012, Zhang and Xie 2015). Freeman et al. (2015) address surgical duration uncertainty and surgical emergencies with a scenario-based approach. Some use Monte Carlo simulation to investigate the impact of

As opposed to existing research, we examine the combined effect of duration uncertainties, recovery resource constraints, patient availability, surgical emergencies and surgery cancellations. Moreover, we consider the “to-follow” practice and study decisions on requested patient arrival time.

**Real-time Schedule Adjustment** As Dexter et al. (2004) illustrates, real-time schedule adjustment is of practical benefit and is thus frequently implemented in surgical suites. However, as noted in (Gupta 2007, Erdogan and Denton 2010), it remains an open challenge with a paucity of relevant literature. We have found only two relevant studies. Van Essen et al. (2012) examine real-time surgical emergency management when surgical durations are assumed deterministic. They make rescheduling decisions by solving an integer programming model to accommodate emergency surgeries. Erdem et al. (2012) investigate a similar problem with the assumption of deterministic surgical durations, in which emergency surgeries could be rejected with a penalty. A GA is used to solve an integer programming model for rescheduling decisions. Different from these two papers, we study real-time schedule adjustment against various perioperative uncertainties, not just emergency surgeries. Moreover, we model a rescheduling framework with notice time requirements as we observed in practice.

We have identified only one paper that adopts the construction-adjustment scheduling framework. Ewen and Mönch (2014) examine a surgical scheduling problem considering duration uncertainty and resource constraints. They propose a GA-based heuristic to assign surgeries to surgeons and then determine scheduled start time using a quartile of duration distributions. The heuristic is run periodically to reassign patients between surgeons as a schedule adjustment. Note that surgery-to-surgeon assignment is out of the scope of our study, and in our study patients are not allowed to be reassigned between surgeons on the day of surgery (which is what we observed in practice).

In our work, we consider a number of practical issues, including the “to-follow” practice, decisions on requested patient arrival times, rescheduling notice time and various perioperative uncertainties. Despite the complexity introduced by these practical considerations, we develop a gradient descent algorithm that identifies near-optimal solutions and outperforms benchmark methods.
3. Surgical Scheduling Process
In this section, we describe the surgical scheduling mechanism that we observed in practice, including important factors such as the “to-follow” practice, perioperative uncertainties and rescheduling notice time.

3.1. “To-follow” Practice
In current surgical practice, medical resources (including human resources) are treated as more costly than patient time. Previous analysis shows that the cost of an OR and a surgical team per hour is 53 times more than the median compensation of all occupations in the U.S. (Bai et al. 2016). If medical resources are left idle when they could have been used for surgeries, it is a significant financial loss to hospitals and may eventually increase patients’ healthcare costs.

With the intent to reduce idle time of medical resources, the “to-follow” policy is widely implemented according to our survey of hospitals. It has two implications in practice: (1) patients are requested to arrive at the hospital well before the scheduled start time of their surgeries ($T_{ss}$). (2) On the day of surgery, a surgeon is ready for surgery at the $T_{ss}$ of his or her first case. A surgery is started when the surgeon and patient are ready and the OR is available. If a case finishes ahead of its predicted time allocation or if a case is cancelled at the last minute (discussed in 3.3), the following scheduled case could start even ahead of its $T_{ss}$.

Note that the “to-follow” practice contrasts with the standard assumption in previous literature where patients arrive punctually at $T_{ss}$ and no surgery can start before its $T_{ss}$ (e.g. Robinson and Chen 2003, Denton and Gupta 2003, Mancilla and Storer 2012, Lee and Yih 2014, Bai et al. 2017).

3.2. OR Sharing Policy and Surgical Schedule Construction
Before the daily operating room schedule is created, allocation of physical operating rooms and specific time slots are determined through the joint efforts of surgeons, nursing and hospital administrators as detailed in E-companion EC.1. As a result, one may see that multiple surgeons are scheduled to use a specific OR over the course of the day, or it could be that one surgeon is kept in a particular OR where all of his/her cases are performed. Since surgeons must be given a start time for their first surgery and the surgery cannot begin until surgeons are ready for surgery, the “to-follow” practice is disrupted by OR sharing (waiting for incoming surgeons). Note neither OR time allocation nor surgical sequencing decision is considered within the scope of this study, but these certainly affect the effectiveness of surgical scheduling optimization. Therefore we model them as parameters, rather than as decision variables, in numerical experiments in §5.1.

Given the surgeons scheduled in ORs and the sequence of surgeries, a surgical schedule is constructed before the surgery date. It includes two sets of decisions, of which patients are informed: scheduled start time ($T_{ss}$) of surgeries and the time when patients are requested to arrive at the
hospital \((T_{eta})\). Note that we study both \(T_{eta}\) and \(T_{ss}\) as decision variables, which differs from the sole decision \(T_{ss}\) or \(T_{eta}\) in previous literature (see §2).

In practice, a surgical schedule is constructed in a heuristic manner. \(T_{ss}\) is typically determined based on estimated surgical duration (Cardoen et al. 2008, Gul et al. 2011, Ozen et al. 2015), and \(T_{eta}\) is selected to be 1.5-2 hours before \(T_{ss}\), a practice customary to our surveyed hospitals and many other U.S. hospitals (ABSMC 2016, NMH 2016). Note that \(T_{eta}\) is well before \(T_{ss}\) as part of the “to-follow” practice to reduce the idle time of medical resources. Although being addressed by heuristics in practice, surgical schedule construction is undeniably challenging because of perioperative uncertainties.

### 3.3. Perioperative Uncertainties

A day in the surgical suite rarely unfolds as planned in the surgical schedule. It is frequently disrupted by various perioperative uncertainties.

**Pre-op and Patient Availability** Patients are requested to arrive by their \(T_{eta}\) on the day of surgery. After arrival, patients go through a series of preoperative activities (pre-op) to get ready for surgery, which may include a pre-operative examination, receiving medications, etc.

Patients’ arrival times and the pre-op durations are random around \(T_{eta}\) and the expected pre-op duration \(L_{E}^{pre}\). Therefore the time when a patient is actually ready for surgery, before which a surgery cannot start, is random w.r.t the expected patient-ready time \(T_{Er} = T_{eta} + L_{E}^{pre}\). In most surveyed hospitals, \(T_{Er}\) is 60-75 minutes before \(T_{ss}\).

**Last-minute Surgery Cancellation** A surgery may have to be cancelled at the last minute due to patient “no-show” or issues identified in pre-op (e.g. unexpected changes in medical conditions). We observed in practice that a surgery is declared cancelled at its \(T_{Er}\), which is roughly when the reason for cancellation is discovered in pre-op, or when a patient is declared as a “no-show”.

If surgery is cancelled at the last minute, OR managers will attempt to move following surgeries up in the surgical schedule in the “to-follow” practice. Occasionally surgeries on the waiting list (“add-on” cases) are placed into the schedule in place of cancelled cases if OR managers ascertain that this will not delay following surgical cases. This practice can be handled with minor modifications as described in E-companion §EC.2.

**Recovery Resources** At the completion of surgery, patients from different ORs are transferred to a shared recovery area, the post-anesthesia care unit (PACU). If the PACU is at its full capacity, patients wait in the OR for a PACU space under the care of an anesthesiologist and the OR is blocked from the following surgeries. PACU admission can be governed by different priority rules such as First-Come-First-Served (FCFS) and Sickest-Patient-First (Bai et al. 2017). We use the FCFS rule that is implemented in our surveyed hospitals.
Duration Uncertainty Both surgical durations and patients’ length of stay (LOS) in the PACU are highly uncertain and have been shown to fit truncated lognormal distributions (Strum et al. 2000, Dexter et al. 2001, Marcon and Dexter 2006, e.g.). For simplicity, turnover time (time for cleaning and preparing an OR between surgeries) is included as part of the surgical duration in this study.

Surgical Emergencies Emergencies in need of immediate surgical treatments, are typically attended to by “on-call” surgeons who are designated in advance (Lord et al. 2014). After arrival, emergency patients enter ORs without undergoing the regular pre-op procedures. In some hospitals, ORs are reserved specifically for surgical emergencies. However, we only consider surgical emergencies that are performed in non-emergency ORs and that may disrupt the surgical schedule of elective surgeries. Note that we model the arrival of surgical emergencies as Poisson processes as in (Van Essen et al. 2012, Helm et al. 2011).

3.4. Real-time Surgical Schedule Adjustment
To alleviate deviation from the surgical schedule, OR managers frequently check the progress of perioperative operations and evaluate the need for schedule adjustments.

Adjustment Frequency and Actions OR managers conduct evaluations at fixed time intervals (e.g. every 30 minutes) and also at the arrival of surgical emergencies or surgery cancellations. When OR managers determine the need of a schedule adjustment, they typically reschedule scheduled start time ($T_{ss}$) of surgeries that have not started. Other adjustment actions listed in Appendix A, including requesting patients to arrive at an earlier or later $T_{eta}$, are infrequently used in our surveyed hospitals. This observation is consistent with the finding in (Van Essen et al. 2012) that rescheduling $T_{ss}$ is the most frequently used rescheduling action.

Note that surgeries are performed one after another regardless of the $T_{ss}$ in the “to-follow” practice, except when multiple surgeons share an OR. In this situation no subsequent operation can start before the assigned surgeon is ready. Therefore only by rescheduling the time when surgeons are ready for surgery $T_{dr}$, which is also $T_{ss}$ of their first surgeries, will make a difference on the day of surgery when using the “to-follow” practice.

Rescheduling Notice Time When asking surgeons to arrive and get ready for surgery earlier (i.e. rescheduling $T_{dr}$ to an earlier time), sufficient time $t_n$ should be allowed for surgeons to travel to the hospital and to get ready to operate. This requirement has been discussed in (Van Essen et al. 2012, Erdem et al. 2012).

We identify an additional notice time requirement when $T_{dr}$ is rescheduled to a later time. For example, if a surgeon’s first surgery with $T_{ss} = 2$pm cannot start until 2:30pm, rescheduling its $T_{ss}$ to 2:30pm theoretically reduces surgeon idle time by 30 minutes. However, if the surgeon is
notified about this rescheduling at 1:50pm, it is likely that he/she is on the way to the hospital or is in preparation for surgery and he/she may have difficulty otherwise utilizing this 30 minutes on such short notice. Therefore, in this case, the original $T_{ss}$ is kept unchanged and hence surgeon idle time from 2pm to 2:30pm is penalized. However, with sufficient notice time (e.g. the surgeon is notified at 11am), this rescheduling is allowed.

Therefore, at any time $t$, the earliest time when the first surgery of a surgeon could be rescheduled to is $t + t_n$ to prevent rescheduling on impractically short notice. Notice time varies between surgeons and services, but $t_n = 45$ minutes is believed to be sufficient by surgeons and OR managers who we have consulted.

4. Mathematical Modeling and Solution Algorithm
In this section, we introduce methodologies to study surgical scheduling optimization. Following the modeling-optimization paradigm, we first discuss our modeling of the surgical scheduling practice and the surgical schedule optimization problem and propose recursive sensitivity analysis and a gradient descent algorithm to optimize surgical scheduling decisions.

4.1. Modeling of Surgical Scheduling Practice
We develop a discrete event dynamic system (DEDS) to model the surgical scheduling mechanism described in §3. Performance metrics are proposed to evaluate the operation of a surgical suite.

**Discrete Event Dynamic System** Inputs to the DEDS include two sets of information: a surgical schedule that includes scheduled start time ($T_{ss}$) and requested arrival time ($T_{eta}$) for all patients; and a scenario (realization) of perioperative uncertainties discussed in §3.3.

The DEDS computes time of different events in the surgical suite according to patient flow and surgeon workflow in the surgical suite: patients arrive at the hospital according to $T_{eta}$, finish pre-op, receive surgeries in ORs and enter the PACU for recovery. Surgeons get ready to operate at $T_{ss}$ of their first surgery and perform surgeries one after another following the “to-follow” practice. OR managers monitor the progress of perioperative activities and make real-time adjustments if needed.

Mathematical formulation of the DEDS is presented in E-companion §EC.4 with notation defined in Appendix B.

**Performance Evaluation** Based on event time in the DEDS, the performance of a given surgical schedule in a scenario of perioperative uncertainties is evaluated based on the resource utilization and patient experience. It is measured by a weighted sum of performance metrics, including patient waiting time, surgeon idle time, OR idle time, OR blocking time, OR overtime and PACU overtime. Weights for performance metrics are determined based on their financial values in §5.1.
A patient’s waiting time starts at the completion of pre-op and ends when the surgery starts. An OR is blocked when a patient waits in the OR for PACU admission. Surgeon idle time is when a surgeon is ready for surgery but is not performing surgery. An OR is counted as idle when it is not utilized for surgeries before all the cases scheduled in it are completed or are cancelled. OR and PACU overtime calculates the work after regular work hours. An example in E-companion EC.3 further illustrates different performance metrics. Note that the waiting time of emergency patients is not penalized since they are handled as soon as possible and also because we focus on the performance of surgical schedules for non-emergent surgeries. However, it can be modeled by making a minor modification in the objective function.

4.2. Surgical Scheduling Optimization

In this section, we first discuss stochastic optimization models for surgical scheduling. We then propose to use sample average approximation to derive solutions that asymptotically approach optimal schedules of the stochastic problem.

Surgical Scheduling Optimization Models

Based on the DEDS and performance evaluation, we evaluate the performance, denoted as \( f(\psi, \omega) \), of a surgical schedule \( \psi \) in a scenario \( \omega \) of perioperative uncertainties. Also we can evaluate the expected performance of \( \psi \) in a scenario set \( \Omega \) of infinite number of possible scenarios, denoted as \( E_{\omega \in \Omega} \left[ f(\psi, \omega) \right] \). As shown in Equation (1), surgical scheduling optimization is to find a \( \psi \) in a feasible set of schedules \( \Psi \), which gives the optimal expected performance in scenario set \( \Omega \).

\[
\min_{\psi \in \Psi} F(\psi) = \min_{\psi \in \Psi} E_{\omega \in \Omega} \left[ f(\psi, \omega) \right]
\]  

For surgical schedule construction, the set of feasible schedules \( \Psi_{\text{schedule}} \) is defined so that surgeries follow a sequence agreed by surgeons and OR managers, and all surgeries are scheduled within regular hours (detailed in E-companion §EC.5). Scenario set \( \Omega_{\text{schedule}} \) is constructed based on parameters derived from historical information, including arrival rate of surgical emergencies, surgery cancellation rate, patient punctuality and duration distributions of pre-op, surgeries and LOS in the PACU.

In real-time schedule adjustment, a given schedule \( \psi_{\text{current}} \) is adjusted by rescheduling a selected set of cases with notice time requirement as discussed in §3. Therefore in the corresponding optimization model, the set of feasible schedules \( \Psi_{\text{adjust}} \) contains schedules in \( \Psi_{\text{schedule}} \) that can be adjusted from \( \psi_{\text{current}} \), i.e., \( \Psi_{\text{adjust}} \subseteq \Psi_{\text{schedule}} \).

Different from \( \Omega_{\text{schedule}} \), perioperative uncertainties are partially observed at the time of adjustment and hence the scenario set for real-time adjustment optimization \( \Omega_{\text{adjust}} \) is conditional on
the revealed information. For instance, if a surgery is still in progress after 1 hour at the time of adjustment, its duration should be at least 1 hour in any scenario \( \omega \in \Omega_{\text{adjust}} \).

In both optimization models, perioperative uncertainties have an infinite number of realizations, which makes the problem stochastic and difficult to solve. By using sample average approximation, we can derive schedules that approach the optimal schedule of the stochastic problem.

**Sample Average Approximation** In sample average approximation (SAA), a finite number of scenarios \( \omega_1, \omega_2, \ldots, \omega_{n_\Omega} \) are sampled from \( \Omega \) and the expected performance of a surgical schedule \( E_{\omega \in \Omega} [f(\psi, \omega)] \) is estimated by \( \frac{1}{n_\Omega} \sum_{i=1}^{n_\Omega} f(\psi, \omega_i) \) (Kim et al. 2015, Kleywegt et al. 2002). The objective of the corresponding SAA optimization problem can be written in Equation (2)

\[
\min_{\psi \in \Psi} f_{n_\Omega}(\psi) = \frac{1}{n_\Omega} \sum_{i=1}^{n_\Omega} f(\psi, \omega_i) \tag{2}
\]

We prove that the SAA estimators in our problem are consistent estimators. This implies that when the number of scenarios \( n_\Omega \) is sufficiently large, solutions to the SAA problem converge to solutions to the stochastic problem with probability 1 (w.p.1). Detail of the proof for Theorem 1 is presented in E-companion §EC.6. We select \( n_\Omega \) based on computational results in §5.2.

**Theorem 1** \( f_{n_\Omega}^* \to F^* \) and \( d(\psi_{n_\Omega}^*, \psi^*) \to 0 \) a.s. as \( n \to \infty \). where \( d(x, Y) = \inf_{y \in Y} |x - y| \).

Next, we solve these SAA scheduling optimization problems by recursive sensitivity analysis and a gradient descent solution algorithm.

### 4.3. Recursive Sensitivity Analysis

Based on the observation of \( f(\psi, \omega) \) in the DEDS with a surgical schedule \( \psi \) and a scenario \( \omega \), recursive sensitivity analysis predicts the performance \( f(\psi + \delta, \omega) \) after an infinitesimal perturbation change \( \delta \) on \( \psi \), without needing additional runs of the DEDS. Recursive sensitivity analysis has been applied to stochastic optimization problems in (e.g Ho and Cao 1991, Fu and Hu 1997, Kim et al. 2015, van Ryzin and Vulcano 2008, Zhang and Xie 2015, Bai et al. 2017), which inspired us at an early stage of this study.

We use an example in Fig. 1 to illustrate the mechanism of recursive sensitivity analysis. With a given schedule and a scenario of uncertainties, it is observed that Surgery 2 waits for 1 hour and starts right after Surgery 1 in the DEDS. If Surgery 1 is rescheduled so that it starts and thus finishes earlier by \( \delta \) hour, we can predict without an additional run of the DEDS that Surgery 2 will still start right after Surgery 1 and thus start earlier by \( \delta \) hour, given that \( \delta \) is infinitesimal. Note that Surgery 2 cannot start before the patient is ready for surgery at 9am. Therefore if \( \delta \) is too large, the sensitivity analysis is no longer valid. If additional surgeries are scheduled in the OR, this sensitivity analysis is conducted recursively following the propagation of \( \delta \) among patients.
Differentiability
Note that the impact of small perturbation $\delta$ on $f(\psi, \omega)$ is a gradient if $f(\psi, \omega)$ is differentiable. Therefore we show the differentiability of $f(\psi, \omega)$ in order to derive gradient information.

As detailed in E-companion §EC.4, $f(\psi, \omega)$ is a linear function of event times that are determined by a series of plus operations and min and max functions on DEDS inputs. Thus $f(\psi, \omega)$ is a piecewise linear function of DEDS inputs, and therefore an infinitesimal perturbation on $\psi$ changes $f(\psi, \omega)$ linearly in a sufficiently small neighborhood unless $\psi$ is a point of nondifferentiability for a given $\omega$.

Nondifferentiability occurs at $\psi$ for a given $\omega$ when a perturbation on $\psi$ changes the result of min or max functions in the DEDS (see Ho and Cao 1991, Fu and Hu 1997 for reference). In our DEDS, these cases occur when event sequence changes or when overtime cost is incurred after perturbation. However, we are able to prove that $f(\psi, \omega)$ is differentiable at any $\psi \in \Psi$ w.p.1. This implies that for a given feasible surgical schedule $\psi$, cases corresponding to points of nondifferentiability occur w.p.0. Please refer to E-companion §EC.7 for the detailed proof for Theorem 2.

Theorem 2 $f(\psi, \omega)$ is differentiable at any $\psi \in \Psi$ w.p.1.

Gradient Derivation
Because of the almost sure differentiability, we are able to use recursive sensitivity analysis to derive the gradient of $f(\psi, \omega)$ at a given $\psi$ by Equation (3). $\frac{\partial f(\psi, \omega)}{\partial T_{ss}(p)}$ and $\frac{\partial f(\psi, \omega)}{\partial T_{eta}(p)}$ calculate the partial derivative w.r.t the scheduled start time and the requested arrival time of patient $p$ in scenario $\omega$. Note that only scheduled start time is changed in the real-time schedule adjustment.

\[
\nabla_{\text{schedule}} f_{n_{\Omega}}(\psi, \omega) = \frac{1}{n_{\Omega}} \sum_{i=1}^{n_{\Omega}} \nabla_{\text{schedule}} f_{\psi, \omega} = \frac{1}{n_{\Omega}} \sum_{i=1}^{n_{\Omega}} \left( \frac{\partial f(\psi, \omega)}{\partial T_{ss}(p)} \right), \forall p \in P^s
\]

\[
\nabla_{\text{adjust}} f_{n_{\Omega}}(\psi, \omega) = \frac{1}{n_{\Omega}} \sum_{i=1}^{n_{\Omega}} \nabla_{\text{adjust}} f_{\psi, \omega} = \frac{1}{n_{\Omega}} \sum_{i=1}^{n_{\Omega}} \left( \frac{\partial f(\psi, \omega)}{\partial T_{eta}(p)} \right), \forall p \in P^s
\] (3)
To calculate $\frac{\partial f(\psi, \omega)}{\partial t_{ss}(p)}$ and $\frac{\partial f(\psi, \omega)}{\partial t_{eta}(p)}$, we investigate the propagation of perturbation in the system in a recursive fashion. First, we predict the direct impact on $f(\psi, \omega)$ of perturbation on the surgical schedule $\psi$. Second, we inspect its impact on surgery start time in $\frac{\partial f(\psi, \omega)}{\partial t_{ss}(p, \omega)}$. Third, we study whether this perturbation has influence on the PACU admission process and subsequently on other patients in $\frac{\partial f(\psi, \omega)}{\partial t_{ap}(p, \omega)}$. This analysis is conducted until the propagation stops.

Mathematical detail of our recursive sensitivity analysis is presented in Appendix §C, based on which we obtain $\nabla_{\text{schedule}} f_{n_{\Omega}}(\psi, \omega)$ and $\nabla_{\text{adjust}} f_{n_{\Omega}}(\psi, \omega)$. Next, we propose a gradient descent algorithm to solve the SAA scheduling optimization problem.

4.4. Gradient Descent Solution Algorithm
Given a gradient derived by recursive sensitivity analysis, we propose a gradient descent algorithm (Algorithm 1) to solve the SAA scheduling optimization problems in Equation (1).

In each iteration, we derive the gradient by recursive sensitivity analysis and determine a direction of change that potentially leads to improvement. To determine how far we should move along the direction, we implement a backtracking line search scheme, in which step size is updated based on initial step size and a constant factor (Nocedal and Wright 2006). A step is taken only when the improvement in the objective is at least a threshold percentage, which generates solutions of similar quality to the Armijo conditions but in much shorter time (Nocedal and Wright 2006). Any infeasible schedule as a result of a step is revised to a feasible schedule by a simple heuristic $\Phi(\psi)$ in Appendix §D.

Gradient descent is restarted at a random point in the feasible region (as defined in §4.2) if the objective stops improving for a certain number of iterations. This prevents our algorithm being trapped at nondifferentiability points. Computational results are presented in §5.2 to select the number of scenarios $n_{\Omega}$ and the number of random restarts $n_{R}$. We also demonstrate that our algorithm identifies near-optimal solutions in the same section.

5. Numerical Results and Discussions
We explain the setup of numerical experiments in §5.1. We demonstrate the performance of the proposed gradient descent algorithm by studying its convergence properties in §5.2 and comparing it to existing algorithms in §5.3. In §5.4, we investigate the impact of the “to-follow” practice. In §5.5 and §5.6, we study the impact of various practical factors on surgical scheduling optimization.

5.1. Test Problem Generation
We construct random test problems for numerical experiments based on the data from previous studies and our survey. Each test problem corresponds to a specific surgery day in a surgical suite of $n_{OR} \in [4, 10]$ open ORs.
Data: Scenarios $\omega_1, \ldots, \omega_n$, percentage threshold $c = 10^{-4}$, initial step size $\rho = 1$, updating factor $\alpha = 2$, step size threshold $\gamma_s = 10^{-6}$ (default values selected in pilot tests)

begin
| Objective value $f^*_{n \Omega} \leftarrow \infty$, best solution $\psi^* \leftarrow 0$; |
| for $n=1$ to $n_R$ do |
| Randomly choose $\psi_0 \in \Psi$; Set $\gamma_0 = \rho$; |
| for $m=1$ to $n_{\text{iter}}$ (a sufficient large number) do |
| Calculate $\nabla_{\text{schedule}} f_{n \Omega}(\psi, \omega)$ or $\nabla_{\text{adjust}} f_{n \Omega}(\psi, \omega)$. |
| Attempt to move along the steepest descent direction. |
| $\psi_{\text{temp}} = \Phi(\psi_{m-1} - \gamma_{m-1} \frac{\nabla f_{n \Omega}(\psi_{m-1})}{\| \nabla f_{n \Omega}(\psi_{m-1}) \|})$ |
| Take the step if the improvement is sufficiently large: |
| if $f_{n \Omega}(\psi_{\text{temp}}) \leq f_{n \Omega}(\psi_{m-1}) \ast (1 - c)$ then |
| $\psi_m = \psi_{\text{temp}}$, $f_{n \Omega}(\psi_m) = f_{n \Omega}(\psi_{\text{temp}})$, $\gamma_m = \alpha \gamma_{m-1}$; |
| else $\psi_m = \psi_{m-1}$, $f_{n \Omega}(\psi_m) = f_{n \Omega}(\psi_{m-1})$, $\gamma_m = \frac{\gamma_{m-1}}{\alpha}$; |
| if $\gamma_m < \gamma_s$ then break; |
| end |
| end |
| if $f_{n \Omega}(\psi_m) < f^*_{n \Omega}$ then $f^*_{n \Omega} = f_{n \Omega}(\psi_m)$, $\psi^* = \psi_m$; |
end

Algorithm 1: Gradient descent algorithm with random restarts

To model different OR sharing practice, we include a parameter $p_{\text{one}}$ to indicate the probability by which an OR is reserved exclusively for a single surgeon. In the test problem construction, each OR is randomly labeled as exclusive or shared w.r.t $p_{\text{one}}$. If a surgical practice tends to reserve an exclusive OR for each surgeon, $p_{\text{one}}$ is closer to 1, and vice versa. As discussed in §3, a shared OR corresponds to the case where an OR is allocated to a group of surgeons, or where an exclusive OR is partially released or where an OR is used for open booking.

Each OR is booked with 2-6 heterogeneous surgeries generated based on the data set in (Lee and Yih 2014). If an OR is labeled as shared, these surgeries are divided into a random number ($> 1$) of partitions, each of which corresponds to a set of surgeries booked by a surgeon. Surgical durations and LOS in the PACU follow truncated lognormal distributions given in the same study.

Each surgery could be cancelled at the last minute w.p. $p_{\text{cancel}}$ and the arrival rate of surgical emergencies is $\lambda_{\text{ER}}$ case per hour per OR. We assume that $p_{\text{cancel}}$ and $\lambda_{\text{ER}}$ are the same for all surgeries and for all ORs, though our approach can handle distinct values. Default values of $p_{\text{one}}$, $p_{\text{cancel}}$ and $\lambda_{\text{ER}}$ are selected based on our survey in Table 1. We study their impact on surgical scheduling optimization in §5.5.
The capacity of the recovery resource $n_{PACU}$ is selected so that the ratio of $\frac{n_{PACU}}{n_{OR}}$ is 0.7 as recommended in (Lee and Yih 2014) (with rounding to make $n_{PACU}$ an integer). The length of regular work hours for ORs and the PACU is 8 hours. The expected duration of pre-op is assumed to be 0.5 hour for all scheduled patients based on our survey and we discuss in §5.6 why this assumption does not affect the results of this study. The uncertainty in patient-ready time for surgery is assumed to follow a triangular distribution $\text{Tri}(-0.5, 0, 0.5)$, which is used in (Ewen and Mönch 2014) to model patient punctuality.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Meaning</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{one}$</td>
<td>Probability of an OR being exclusively reserved for one surgeon</td>
<td>0.25</td>
</tr>
<tr>
<td>$p_{cancel}$</td>
<td>Last-minute surgery cancellation rate</td>
<td>0.04</td>
</tr>
<tr>
<td>$\lambda_{ER}$</td>
<td>Emergency arrival rate per hour per OR</td>
<td>0.01</td>
</tr>
</tbody>
</table>

We select weight parameters for performance metrics based on the analysis in (Bai et al. 2016), which computes their monetary value ($ per hour) based on national statistics and previous studies. Note that OR time or surgeon time is regarded more costly than patient time.

- OR idle $c_{oi} = 714.2$
- OR overtime $c_{oo} = 1249.9$
- Patient waiting $c_{pw} = 17.1$
- Surgeon idle $c_{si} = 197.7$
- OR blocking $c_{b} \sim U[0, 714.2]$
- PACU overtime $c_{po} = 148.9 \times n_{PACU}$

### 5.2. Convergence Test

In this section, we study the quality of solutions obtained by the proposed algorithm. We use the method outlined by Linderoth et al. (2006) to statistically derive the gap (referred to as “optimality gap”) between the objective values of solutions obtained and the optimal objective value. The bias of estimation can be reduced by increasing the number of scenarios $n_\Omega$ and the number of random restarts $n_R$ in Algorithm 1. Therefore different values of $n_\Omega$ and $n_R$ are tested in the experiment. This experiment design is detailed in E-companion §EC.8.

We observe similar convergence results in all test problems and illustrate with a random test problem in Fig 2. The optimality gap decreases as $n_\Omega$ increases and the small gap at $n_\Omega = n_R = 5000$ signifies that the proposed algorithm identifies near-optimal solutions. Also as noted by Linderoth et al. (2006), the choice of $n_\Omega$ and $n_R$ has little impact on the objective values of solutions obtained by the proposed algorithm, which reflects that the quality of solutions is not significantly impacted by $n_\Omega$ and $n_R$.

**Takeaway** Based on this test, we show that the proposed algorithm identifies near-optimal solutions. Also we determine to use $n_\Omega = 1000$ and $n_R = 50$ in the subsequent experiments.
5.3. Comparison of the Proposed Algorithm and Benchmark Methodologies

In this section, we demonstrate that our proposed algorithm outperforms benchmark algorithms and is effective in cost reduction in surgical suites.

As shown in Table 2, for surgical schedule construction, we compare the performance of a simple but widely-used heuristic $Heu$, two best-performing algorithms $ONP$ and $OP$ in previous studies (Bai et al. 2017, Denton et al. 2007) as well as the proposed gradient descent algorithm $GDc$. To evaluate the benefits of optimization on requested patient arrival time $Teta$, we include a variant $GDc^*$ which optimizes scheduled start time $Tss$ by the proposed gradient descent algorithm, but selects $Teta$ by heuristics.

For real-time schedule adjustment, due to the lack of relevant literature as discussed in §2, we could not find any benchmark algorithm to compare with. However we show that our gradient descent algorithm $GDa$ can significantly improve the end-of-day performance even when a day starts with suboptimal surgical schedules.

**Surgical Schedule Construction Optimization** Surgical schedule construction algorithms are tested in simulations as follows. First, 100 random test problems are generated with default parameters in Table 1, each of which corresponds to a day in a surgical suite of randomly-generated settings. Second, each method is implemented to select $Tss$ and $Teta$ and thus we have 5 surgical schedules for each test problem. Third, each surgical schedule is evaluated in simulations to estimate the expected cost. A run of simulation starts each surgery day with a constructed surgical schedule and continues as modeled in §4.1 in the “to-follow” practice without real-time adjustments.

Note that in $Heu$, $ONP$, $OP$ and $GDc^*$, $Teta = Tss - 1.5$ is used for all patients in the surgical schedule, which is customary to our surveyed hospitals and many other hospitals in the U.S. (ABSMC 2016, NMH 2016). We study the impact of $Teta$ in §5.6.
## Table 2  Surgical Scheduling Optimization Methods under Comparison

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Methods</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Construct</td>
<td>Benchmark</td>
<td>Heu</td>
<td>Heuristic widely-used in practice that constructs scheduled start time ( T_{ss} ) based on mean surgical durations (Gul et al. 2011); requested patient arrival time ( T_{eta} ) by heuristics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ONP</td>
<td>Optimization considering duration uncertainty but not recovery resource for ( T_{ss} ) (Denton et al. 2007); ( T_{eta} ) by heuristics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OP</td>
<td>Optimization considering duration uncertainty and recovery resource for ( T_{ss} ) (Bai et al. 2017); ( T_{eta} ) by heuristics</td>
</tr>
<tr>
<td>Proposed</td>
<td></td>
<td>GDc*</td>
<td>Gradient descent algorithm that optimizes ( T_{ss} ), but selects ( T_{eta} ) by heuristics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GDc</td>
<td>Gradient descent algorithm that optimizes both ( T_{ss} ) and ( T_{eta} ) in schedule construction</td>
</tr>
<tr>
<td>Adjust</td>
<td>Proposed</td>
<td>GDa</td>
<td>Gradient descent algorithm that enables real-time adjustment on ( T_{ss} ); rescheduling every ( t_{int} = 0.5 ) hour (default) in addition to event-driven adjustments as described in §3.4</td>
</tr>
</tbody>
</table>

*In Heu, the first surgery in an OR is scheduled at time 0 and following surgeries are scheduled with interval equal to their expected surgical duration

**None of Heu, ONP and OP accounts for the “to-follow” practice. GDc* optimizes \( T_{ss} \) considering the “to-follow” practice but selects \( T_{eta} \) by heuristics. GDc optimizes both \( T_{ss} \) and \( T_{eta} \) considering the “to-follow” practice.

## Table 3  Comparison between different algorithms in surgical schedule construction for a surgical suite with 4 to 10 ORs

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Avg daily cost reduction (% improvement)</th>
<th>Statistically Significant Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>ONP v.s. Heu</td>
<td>$0.45 \times 10^3 (3.2%)$</td>
<td>66/100</td>
</tr>
<tr>
<td>OP v.s. Heu</td>
<td>$0.27 \times 10^3 (1.9%)$</td>
<td>38/100</td>
</tr>
<tr>
<td>GDc* v.s. Heu</td>
<td>$0.74 \times 10^3 (5.3%)$</td>
<td>86/100</td>
</tr>
<tr>
<td>GDc* v.s. ONP</td>
<td>$0.29 \times 10^3 (2.2%)$</td>
<td>54/100</td>
</tr>
<tr>
<td>GDc* v.s. OP</td>
<td>$0.47 \times 10^3 (3.4%)$</td>
<td>85/100</td>
</tr>
<tr>
<td>GDc v.s. GDc*</td>
<td>$0.13 \times 10^3 (1.0%)$</td>
<td>73/100</td>
</tr>
<tr>
<td>GDc v.s. Heu</td>
<td>$0.87 \times 10^3 (6.2%)$</td>
<td>97/100</td>
</tr>
<tr>
<td>GDc v.s. ONP</td>
<td>$0.43 \times 10^3 (3.1%)$</td>
<td>82/100</td>
</tr>
<tr>
<td>GDc v.s. OP</td>
<td>$0.60 \times 10^3 (4.4%)$</td>
<td>92/100</td>
</tr>
</tbody>
</table>

*Note: Heu, ONP, OP, GDc* and GDc indicate that a surgery day start with schedules constructed by these algorithms without any real-time adjustment.

As shown in Table 3, our proposed algorithms GDc and GDc* outperform all benchmark methods. By optimizing \( T_{ss} \), GDc* outperforms benchmark methods by 2.2%-5.3% reduction in average daily cost. Optimization of \( T_{eta} \) in GDc contributes additional cost saving over GDc* and GDc outperforms benchmark methodologies by $0.43 - 0.87 \times 10^3$ reduction in average daily cost.

**Real-time Schedule Adjustment Optimization** We apply the proposed schedule adjustment GDa to surgery days started with schedules constructed by GDc and Heu, which are the best-performing and the worst-performing methodology in surgical schedule construction. We show that even starting a day with suboptimal schedules generated by Heu, GDa can “correct” the overall performance to a level similar to that achieved by a schedule generated by GDc.
This numerical test is conducted in the same 100 test problems as those in the test of surgical schedule construction algorithms. A simulation starts with a surgical schedule constructed by \( GDc \) or \( Heu \) and is paused at each rescheduling point. \( GDa \) is executed to update \( T_{ss} \) based on scenarios newly-generated as described in §4.2. The simulation is then resumed with the updated \( T_{ss} \). We implement \( GDa \) at 0.5-hour intervals (\( t_{int} = 0.5 \)) in addition to adjustment at surgical emergency arrivals and surgery cancellations, as described in §3.4. Impact of rescheduling frequency \( t_{int} \) is studied in §5.6.

### Table 4 Effectiveness of real-time schedule adjustment for a surgical suite with 4 to 10 ORs

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Avg daily cost reduction (% improvement)</th>
<th>Statistically Significant Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Heu + GDa ) v.s. ( Heu )</td>
<td>$1.02 \times 10^3$ (7.3%)</td>
<td>100/100</td>
</tr>
<tr>
<td>( GDc + GDa ) v.s. ( GDc )</td>
<td>$0.45 \times 10^3$ (3.5%)</td>
<td>96/100</td>
</tr>
<tr>
<td>( GDc ) v.s. ( Heu )</td>
<td>$0.87 \times 10^3$ (6.2%)</td>
<td>97/100</td>
</tr>
<tr>
<td>( GDc + GDa ) v.s. ( Heu + GDa )</td>
<td>$0.29 \times 10^3$ (2.2%)</td>
<td>56/100</td>
</tr>
</tbody>
</table>

*Note: \( Heu + GDa \) and \( GDc + GDa \) indicates that surgery days start with schedules constructed by \( Heu \) and \( GDc \) and real-time adjustment \( GDa \) is implemented in real-time.

Our proposed real-time adjustment algorithm \( GDa \) improves the overall performance of a surgical suite. According to Table 4, \( GDa \) contributes $1.02 \times 10^3$ (7.3%) average cost savings per day if a surgery day starts with a surgical schedule generated by \( Heu \). This number is $0.45 \times 10^3$ (3.5%) when a day starts with schedules generated by \( GDc \). Note that \( GDc \) gives surgical schedules with the best-expected performance (over an infinite number of scenarios), but these schedules are not necessarily optimal for each individual scenario. Therefore real-time adjustment enables further improvement based on the realization of what happens in the course of a day.

After real-time adjustment by \( GDa \), the overall performance of a day starting with a schedule obtained by \( Heu \) is very similar to a day starting with a schedule constructed by \( GDc \). Cost difference between \( GDc + GDa \) and \( Heu + GDa \) is only significant in 56% of test problems with an average difference of $0.29 \times 10^3$. This difference is significantly smaller than the difference of $0.87 \times 10^3$ between \( GDc \) and \( Heu \).

**Takeaway** The proposed gradient descent algorithms for surgical schedule construction and real-time schedule adjustment outperform benchmark algorithms and effectively reduce the overall cost of a surgical schedule.

### 5.4. Impact of the “To-follow” Practice

The “to-follow” practice has not been considered in previous literature (e.g. Robinson and Chen 2003, Denton and Gupta 2003, Mancilla and Storer 2012, Lee and Yih 2014, Bai et al. 2017). In this section, we evaluate the benefits of implementing the “to-follow” practice and investigate the
impact of not considering it in surgical schedule construction. Note that we focus on its impact on surgical schedule construction since we could not locate benchmark real-time adjustment algorithms due to the lack of literature as discussed in §2.

We study two practical elements of the “to-follow” practice: (1) to request patients to arrive well before $T_{ss}$; (2) “to-follow” previous surgery, i.e., to start a surgery before $T_{ss}$ when the patient and the surgeon are ready and the OR is available. Accordingly, we compare the performance of surgical schedules (1) in which patients are requested to arrive 1.5 hours earlier than $T_{ss}$ ($t_{\text{ahead}} = T_{ss} - T_{eta} = 1.5$) and 0.5 hour earlier than $T_{ss}$ ($t_{\text{ahead}} = 0.5$) and (2) in simulations with and without the practice of allowing surgeries to start before $T_{ss}$. Note that $t_{\text{ahead}} = 1.5$ is customary to our surveyed hospitals and many other hospitals in the U.S. (ABSMC 2016, NMH 2016). Given that the expected duration of pre-op is 0.5 hour in the test problems, when ahead time is $t_{\text{ahead}} = 0.5$ hour patients are expected to be ready for surgery at $T_{ss}$, which is the earliest time when a surgery can start in this case.

![Figure 3](image_url)  
**Fig. 3** Impact of the “to-follow” practice on surgical schedule construction for a surgical suite with 4 to 10 ORs (confidence intervals of average daily cost)

**Benefits of “To-follow” Practice** Performance of surgical schedules is evaluated in simulations in the same 100 test problems in §5.3. As shown in Fig. 3, it is beneficial to have patients arrive earlier at $t_{\text{ahead}} = T_{ss} - T_{eta} = 1.5$, compared to having patients arrive at $t_{\text{ahead}} = 0.5$). This observation is the same no matter if $\text{Heu}, \text{OP}, \text{ONP}$ or $\text{GDc}^*$ is used to determine $T_{ss}$, or whether surgeries
can start before $T_{ss}$. This is closely related to the current practice where surgeon time and OR time are considered more costly than patient time. Requesting patients to arrive earlier can reduce the probability of medical resources being left idle when previous surgeries finish earlier than expected. This observation is consistent with findings in (Dexter and Traub 2000, Wachtel and Dexter 2007). Note that in §5.6, we discuss the choice of $t_{ahead}$ to reduce overall cost without excessive waiting time for patients.

We also observe that allowing surgeries to start before $T_{ss}$ substantially reduces the average daily cost when $t_{ahead} = 1.5$, but the difference is not significant when $t_{ahead} = 0.5$. When patients arrive and get ready for surgeries well before $T_{ss}$, starting the surgery immediately when the OR and the surgeon are available reduces the costly idle time of medical resources. However, if $t_{ahead} = 0.5$, patients are ready for surgery roughly around $T_{ss}$, which is also the earliest time when a surgery can start. In this circumstance, allowing or not allowing surgeries to start before $T_{ss}$ has little impact on the time when surgeries actually start and thus has limited influence on the overall performance.

Note that $GDc$ optimizes both $T_{ss}$ and $T_{eta}$ and thus it is not evaluated with $T_{eta}$ selected by heuristics. Consistent with previous observations, allowing surgeries to start before $T_{ss}$ substantially reduces the overall cost for surgery days that start with schedules constructed $GDc$. However, $GDc$ performs poorly compared with other optimization approaches when surgeries are not allowed to start before $T_{ss}$. This behavior is reasonable since schedules constructed by $GDc$ will be naturally suboptimal whenever the “to-follow” practice is not implemented.

**Overestimated Effectiveness of Schedule Construction Optimization** We notice that the effectiveness of optimization methodologies is substantially weakened with the “to-follow” practice. After the implementation of the “to-follow” practice, i.e., when $t_{ahead} = 1.5$ and surgeries are allowed to start before $T_{ss}$, the advantage margin of optimization methods $OP$, $ONP$, $GDc^*$ over $Heu$ shrinks significantly, compared to the case where $t_{ahead} = 0.5$ and surgeries are not allowed to start before $T_{ss}$.

As shown in Table 3, best-performing algorithms $OP$ and $ONP$ in (Denton et al. 2007, Bai et al. 2017) show only $0.27 \times 10^3$ and $0.45 \times 10^3$ reduction in average daily cost over $Heu$ with the “to-follow” practice. Moreover, these improvements are statistically significant only in 38% and 66% of test problems. Considering the potential expense of development and implementation, $OP$ and $ONP$ may not have a clear advantage over heuristic $Heu$ that is being used in practice.

**Takeaway** The “to-follow” practice is beneficial in the current surgical practice to compensate for the uncertainty intrinsic to surgical suite operations. However, without considering the “to-follow” practice, effectiveness of optimization algorithms including best-performing algorithms in previous studies could be significantly overestimated. In contrast, the simple heuristic that is being used in practice ($Heu$) performs reasonably well when the “to-follow” practice is implemented.
5.5. Impact of the OR Sharing Practice, Cancellations and Surgical Emergencies

Ideally, both of our algorithms for surgical schedule construction and real-time schedule adjustment should be implemented to maximize cost savings. However when budget, time and other resources are constrained in surgical practice, it is of particular interest to OR managers to prioritize these two interventions. In this section, we obtain general insights on the prioritization of two interventions by evaluating their effectiveness in different circumstances, including different OR sharing practice and perioperative uncertainties.

Impact of OR Sharing Practice Evaluation is conducted in test problems with different $p_{\text{one}}$, which affects the extent of OR sharing among surgeons as discussed in §1 and in §3.2. If $p_{\text{one}} = 1$, every OR is used by a single surgeon while every OR is shared by two or more if $p_{\text{one}} = 0$. For each value of $p_{\text{one}} = \{0, 0.25, 0.5, 0.75, 1\}$, 100 random test problems are generated with default $p_{\text{cancel}}$ and $\lambda_{ER}$ in Table 1. Heu, GDc and GDa are tested in all 500 test problems following the same experiment design in §5.3.

In Fig. 4, the first thing to notice is that the overall cost is lowest when an OR is scheduled for a single surgeon for the whole day ($p_{\text{one}} = 1$). Scheduling multiple surgeons into a single OR can blunt the benefits of the “to-follow” practice since the first surgery of incoming surgeons cannot begin until surgeons are ready to operate.

When $p_{\text{one}} = 0$ or 0.25, real-time adjustment GDa is effective in improving overall performance: GDa can save more than $1.00 \times 10^3$ and $0.45 \times 10^3$ in average daily cost over Heu and GDa. The end-of-day performance of GDc+GDa and Heu+GDa is very similar. However, its effectiveness
vanishes as $p_{\text{one}}$ increases. This observation can be explained by the “to-follow” practice where surgeries start one after another regardless of the scheduled start time $T_{ss}$ if an OR is used by a single surgeon. In this case, improvement achieved by real-time adjustment of $T_{ss}$ is limited, given that patients are ready for surgeries well before their $T_{ss}$.

As $p_{\text{one}}$ increases, the improvement margin of $GDc$ over $Heu$ also decreases, which signifies the decreasing effectiveness of surgical schedule construction optimization. However, $GDa$ becomes almost ineffective when $p_{\text{one}}$ is close to 1, which makes $GDc$ of dominant importance in cost reduction.

Therefore it is preferable to schedule a single surgeon for the whole day in the OR with the “to-follow” practice. If most ORs are shared by multiple surgeons, the real-time adjustment is of greater usefulness in performance improvement, while optimization of surgical schedule construction is of dominant significance when most ORs are used by a single surgeon.

**Impact of Surgery Cancellation Rates** Test problems are constructed with different last-minute surgery cancellation rates $p_{\text{cancel}} = 0, 0.04, 0.1$ and default values of $p_{\text{one}}$ and $\lambda_{ER}$ in Table 1. Note $p_{\text{cancel}} = 0.04$ is the result of our survey, $p_{\text{cancel}} = 0.1$ is reported in a previous study (Dexter et al. 2005) and $p_{\text{cancel}} = 0$ is an extreme case for testing purpose.

![Fig. 5 Surgical scheduling optimization under different last-minute cancellation rates for a surgical suite with 4 to 10 ORs](image)

(a) Confidence intervals of average daily cost reduction  (b) Number of test problems with statistically significant cost reduction

In Fig. 5, the value of both surgical schedule construction optimization $GDc$ and schedule adjustment $GDa$ increases as $p_{\text{cancel}}$: $GDc$ shows a growing improvement margin over $Heu$. The improvement of $GDc+GDa$ and $Heu+GDa$ over $GDc$ and $Heu$ is also increasing.

When $p_{\text{cancel}} = 0$ or 0.04, $GDc+GDa$ and $Heu+GDa$ give similar end-of-day performance, which signifies that real-time schedule adjustment handles last-minute cancellations reasonably well. On
the other hand, when \( p_{\text{cancel}} = 0.1 \), both \( GDc \) and \( GDa \) are needed to compensate for the randomness caused by unusually high last-minute cancellation rates.

**Impact of Surgical Emergency Arrival Rates** We evaluate the impact of different emergency arrival rates \( \lambda_{ER} = 0, 0.01, 0.05 \) in test problems with default \( p_{\text{one}} \) and \( p_{\text{cancel}} \) in Table 1. \( \lambda_{ER} = 0.01 \) is set up based on our survey while \( \lambda_{ER} = 0 \) and 0.05 are extreme cases for testing purpose. Note that as described in §3.3, we only consider surgical emergencies that are performed in ORs that are not specifically reserved for emergencies.

As seen in Fig 6, as \( \lambda_{ER} \) increases, \( GDc \) sustains a stable improvement over \( Heu \) while \( GDa \) is increasingly effective on both \( GDc \) and \( Heu \). End-of-day performance of \( GDc+GDa \) and \( Heu+GDa \) show similar end-of-day performance. This demonstrates that real-time schedule adjustment is effective in the management of surgical emergencies, which agrees with findings from previous studies on surgical emergency management (Van Essen et al. 2012, Erdem et al. 2012).

![Confidence intervals of average daily cost reduction](image1)
![Number of test problems with statistically significant cost reduction](image2)

**Fig. 6** Surgical scheduling optimization under different emergency arrival rates for a surgical suite with 4 to 10 ORs

**Takeaway** It is preferable to schedule a single surgeon for the whole day in the OR with the “to-follow” practice. Optimization of surgical schedule construction is of dominant significance when most ORs are used by a single surgeon. On the other hand, the value of schedule adjustment increases in surgical suites with frequent surgeon changes and high emergency and cancellation rates, making the algorithm used to construct the schedule in the first place of little consequence. Administrators are advised to choose the most cost-effective approach for service improvement, based on the OR sharing policy, emergency and cancellation rates in their surgical suites.
5.6. Selection of Requested Patient Arrival Time and Real-time Adjustment Frequency

We have shown in §5.3 that Heu, a heuristic being used in practice that determines scheduled start time $T_{ss}$ based on mean surgical duration and selects requested patient arrival time as $T_{eta} = T_{ss} - 1.5$, gives satisfactory performance compared to existing optimization algorithms in the “to-follow” practice. In this section, we first explore whether the performance of this widely-used heuristic can be further improved by selecting a different $T_{eta}$.

Selection of Requested Patient Arrival Time

On the day of surgery, patients are asked to arrive at the hospital at $T_{eta}$ to be ready for surgery after the pre-op. With an ideal $T_{eta}$, a patient is ready for surgery at the time when the surgeon and OR become available so that no waiting and idleness occur. We examine Heu in which all patients are requested to arrive at time $t_{ahead} = T_{ss} - T_{eta}$ before their $T_{ss}$, with different values of $t_{ahead} = 0.5$, 1, 1.5, 2 and 2.5. Note that $t_{ahead}$ is commonly 1.5-2 hours in practice (ABSMC 2016, NMH 2016) and when $t_{ahead} = 0.5$, the earliest time when a surgery can start is roughly $T_{ss}$ as discussed in §5.4.

![Fig. 7](image)

**Fig. 7** Confidence intervals of average daily cost under different $t_{ahead} = T_{ss} - T_{eta}$ for a surgical suite with 4 to 10 ORs

In Fig. 7, we show test results in 100 random test problems with default settings in Table 1. Proposed algorithms $GDc$ and $GDa$ are not affected by the choice of $t_{ahead}$. It is first noted that $GDc$, in which $t_{ahead}$ is made a variable for patients, always outperforms Heu with a fixed $t_{ahead}$ for patients by more than $0.85 \times 10^3$ in daily cost saving. This observation agrees with many previous studies (e.g. Robinson and Chen 2003, Denton and Gupta 2003, Dexter and Traub 2000, Wachtel and Dexter 2007) in that variable $t_{ahead}$ in patient scheduling helps improve the overall performance.
For scheduling heuristic $\text{Heu}$, increasing $t_{\text{ahead}}$ from 0.5 to 1.5 hours reduces the overall cost of the surgical suite with or without real-time adjustment $GDa$. This is because a larger $t_{\text{ahead}}$ can reduce the chance of surgeon waiting and OR being idle in the “to-follow” practice, in the case where the previous surgery finishes earlier. This result justifies why patients are requested to arrive earlier in the “to-follow” practice.

However, the overall performance of either $\text{Heu}$ or $\text{Heu}+GDa$ is not sensitive beyond $t_{\text{ahead}} = 1.5$, which agrees with the practice where $t_{\text{ahead}} = 1.5-2$. With $t_{\text{ahead}} > 1.5$, the benefit of reducing idle time of medical resources is cancelled out by the excessively long patient waiting time. Therefore healthcare administrators are advised to study patient arrivals to avoid unnecessary patient waiting time.

So far we have assumed that the expected duration of pre-op is 0.5 hour for all surgeries, but we understand that surgeries may need pre-op of different duration in some hospitals. In such cases, results of this paper can be readily reinterpreted w.r.t. the time when patients are expected to be ready for surgery $T_{Er}$ instead of their requested arrival time $T_{eta}$, since $T_{Er} = T_{eta} + 0.5$ as discussed in §3.3. For example, discussions on $t_{\text{ahead}} = T_{ss} - T_{eta} = 0, 0.5, 1, 1.5, 2$ and 2.5 in this section also apply to the corresponding $T_{ss} - T_{Er} = 0, 0.5, 1, 1.5$ and 2.

**Selection of Real-time Adjustment Frequency** In practice, frequent real-time adjustments may incur inconvenience and dissatisfaction of patients and medical staff. However, opportunities for improvement may be missed if $t_{\text{int}}$ is too large. Therefore in the second part of this section, we study whether a different adjustment interval $t_{\text{int}}$ of real-time adjustment is beneficial.

As discussed in §3.4, real-time schedule adjustment $GDa$ is carried out at fixed time intervals $t_{\text{int}}$ and also at emergency arrivals and surgery cancellations. With surgical schedules constructed by $GDc$ and $\text{Heu}$ for 100 test problems with default settings in Table 1, $GDa$ with different frequency $t_{\text{int}} = 0.25, 0.5, 1$ and 2 hours are tested.

As shown in Fig. 8, $GDa$ consistently improves end-of-day performance of surgical schedules constructed by $GDc$ and $\text{Heu}$ in more than 90% of the test problems. The magnitude of cost reduction achieved by $GDa$ shrinks as $t_{\text{int}}$ increases, i.e., with less frequent adjustment. However, the improvement margins achieved by $GDa$ with $t_{\text{int}} = 0.25, 0.5$ and $t_{\text{int}} = 1$ are not significantly different. The practice of $t_{\text{int}} = 0.5$ hour as we observed works reasonably well. We suggest that OR managers carefully select a $t_{\text{int}}$ based on the needs of each surgical practice in order to reduce unnecessarily frequent schedule adjustment.

**Takeaway** Requesting that all patients arrive the same amount of time ($t_{\text{ahead}}$) before their surgeries (as in the current practice) is not necessarily optimal. Moreover, the overall performance of the surgical suite is not sensitive when $t_{\text{ahead}}$ is beyond 1.5, which agrees with current experience.
We recommend the requested patient arrival time be carefully considered to avoid excessive patient waiting and not waste resources.

We notice that frequent real-time adjustment does not guarantee significant performance improvement and the practice of $t_{int} = 0.5$ hour as we observed works reasonably well. Given that real-time adjustment already occurs with surgical emergencies and procedural cancellations, the frequency of additional real-time adjustments should be determined so that a balance is maintained between inconvenience caused by excessive adjustments and missed improvement opportunities due to infrequent adjustments.

6. Conclusions and Future Research

In this study, we reinvestigate the effectiveness of surgical scheduling optimization. We were able to model the surgical scheduling mechanism based on our observations in hospitals and on interviews with practitioners. We account for 1) the widely-implemented “to-follow” practice, which contrasts with common assumptions found in previous studies; 2) scheduling decisions that OR managers make on a daily basis in surgical schedule construction and real-time schedule adjustment; and 3) OR sharing policy and various commonly-encountered perioperative uncertainties, only some of which have been discussed in previous work. To tackle this complex real-world problem we develop a gradient descent solution algorithm based on recursive sensitivity analysis, which identifies near-optimal solutions and which outperforms benchmark methods found in previously published studies in numerical experiments.

Through numerical experiments, we illustrate the “to-follow” practice is an effective way to mitigate the uncertainties in surgical suite operations, which may be why it is widely implemented.
in practice. However, previous studies that have shown significant effectiveness of surgical schedule optimization, do not consider the “to-follow” practice. We show that these hypothetical gains from schedule optimization algorithms could quickly disappear when the “to-follow” practice is implemented.

Based on computational tests with various practical factors, we also provide insights into surgical schedule construction and real-time surgical schedule adjustment. We show that it is preferable to schedule a single surgeon for the whole day in the OR, while scheduling multiple surgeons into a single OR can reduce the benefits of the “to-follow” practice because of waiting for incoming surgeons. Real-time schedule adjustment, which alleviates significant chaos and confusion in current practice, can be performed quite well with our proposed optimization algorithm. The value of such real-time adjustment is more significant in surgical suites with frequent surgeon changes in ORs and high emergency and cancellation rates, which may make algorithms used to construct the schedule in the first place seems to be of little consequence.

We show that having patients arrive a fixed amount of time before their surgeries as in the current practice, is not necessarily optimal. However, current practice, where patients are requested to arrive 1.5-2 hours ahead, works reasonably well, considering the overall performance of the surgical suite is not time-sensitive when this early time goes beyond 1.5 hours. Therefore we recommend that the requested patient arrival time be carefully selected to avoid excessive patient waiting and not waste valuable resources. We also find that increasing the frequency of real-time adjustment does not guarantee performance improvement. But the current practice of rescheduling every 30 minutes, in addition to rescheduling at the arrival of surgical emergencies and at surgery cancellations, work reasonably well.

The methodological framework in this paper provides a platform of great flexibility for studies on surgical scheduling optimization. This framework allows healthcare researchers to study the overall effects of surgical scheduling optimization, as we do in this paper, and also to solve specific scheduling problems customized to their medical institutions.

**Future Research** In this study, we focus on the optimization of scheduled surgical cases. However, management of add-on surgeries, which can be thought of as cases on a waiting list, is out of the scope of this study. Add-on cases could wait for available OR time on a surgery day without any guarantee as to whether the case can be done that day or when the case will start. Consequently, surgeons will need to wait without an estimated surgery start time while patients are kept NPO all day. Even so, the add-on case could still be pushed back to the following day. It will be of great practical importance to coordinate add-on cases into available OR time and most importantly to provide an estimate of the probability whether a case can be performed with an estimated surgery start time.
Another future research direction is on surgical capacity allocation. We used a parameter to model the OR sharing policy in test problems and demonstrate its impact on surgical scheduling optimization. A natural follow-up research problem is how to adjust surgical capacity allocation so that the integrated decisions of surgical capacity allocation and surgical scheduling can improve resource utilization and enhance the patient experience. It will be of practical interest to investigate this problem by considering many of the factors that have been discussed in this paper.

References


Mathias JM (2011) Surgical scheduling: taking an important role to the next level. *OR manager* 27(3).


**Appendix A: Infrequent Actions of Real-time Schedule Adjustment**

We list some real-time schedule adjustments that are infrequently used in our observed practice.

1. Either surgery-to-OR assignments or surgery sequence is rarely altered since they are approved by surgeons and OR managers as described by §EC.1. This is also noted in (Van Essen et al. 2012).
2. Scheduled surgeries are not cancelled as an adjustment action. This practice is supported by the finding that it is financially beneficial to complete all surgeries even when overtime is required (Dexter et al. 2004).

3. Emergency surgeries are not rescheduled or cancelled for patient safety (May et al. 2011). They are performed as soon as possible.

4. Changing patients’ requests arrival time \( T_{eta} \); i.e., asking patients to arrive earlier, is less frequently used as an adjustment action. Given that \( T_{eta} \) is well before the scheduled start time, patients would typically be on the way to the surgical suite when such change is informed.

**Appendix B: Notation**

We define the notation used in this study in Table 5 and Table 6.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_{OR}, n_{PACU} )</td>
<td>Total number of ORs and PACU beds. ORs are indexed by ( j \in J = {1, 2, \ldots, n_{OR}} )</td>
</tr>
<tr>
<td>( \omega )</td>
<td>A realization/scenario of perioperative uncertainties</td>
</tr>
<tr>
<td>( P_s^p, P_e^p )</td>
<td>Set of scheduled surgeries in OR ( j ); ( P_s^p = \bigcup_{j \in J} P_{sf}^p )</td>
</tr>
<tr>
<td>( P_{sf}^p, P_{sc}^p, \omega )</td>
<td>Set of emergency surgeries in OR ( j ) in scenario ( \omega ); ( P_{sc}^p, \omega = \bigcup_{j \in J} P_{sc}^p, \omega )</td>
</tr>
<tr>
<td>( I^{OR}(p) )</td>
<td>The OR where surgery ( p ) is performed; ( p \in P_s^p \cup P_e^p ).</td>
</tr>
<tr>
<td>( I^d(p) )</td>
<td>The surgeon who performs scheduled surgery ( p \in P_s^p ).</td>
</tr>
<tr>
<td>( I^{alt}(p) )</td>
<td>The position of ( p \in P_s^p ) in the sequence of scheduled surgeries in the OR.</td>
</tr>
<tr>
<td>( I^s(p) )</td>
<td>The position of ( p ) in the sequence of surgeries of its surgeon; ( p \in P_s^p \cup P_e^p ).</td>
</tr>
<tr>
<td>( L^{intra}(p, \omega) )</td>
<td>Surgical duration of patient ( p ) in scenario ( \omega ); ( p \in P_s^p \cup P_{sc}^p, \omega ).</td>
</tr>
<tr>
<td>( L^{post}(p, \omega) )</td>
<td>LOS in the PACU of patient ( p ) in scenario ( \omega ); ( p \in P_s^p \cup P_{sc}^p, \omega )</td>
</tr>
<tr>
<td>( L^{pre}_p )</td>
<td>Expected duration of pre-op of patient ( p \in P_s^p ).</td>
</tr>
<tr>
<td>( L^{pre}(p, \omega) )</td>
<td>Randomness of actual ready time of patient ( p \in P_s^p ) around the expected patient-ready time in scenario ( \omega ); ( L^{pre}(p, \omega) = +\infty ) if ( p ) is cancelled in scenario ( \omega ).</td>
</tr>
<tr>
<td>( P_{sf}^p, P_{sc}^p )</td>
<td>Set of scheduled surgeries in OR ( j ) that are performed and cancelled in scenario ( \omega ), respectively; i.e. ( P_{sf}^p, \omega = { p</td>
</tr>
<tr>
<td>( t_r(p, \omega) )</td>
<td>Arrival time of an emergency patient ( p \in P_{sc}^p, \omega ) in scenario ( \omega ).</td>
</tr>
<tr>
<td>( t_{int}, t_{nt}, t_w )</td>
<td>Rescheduling interval, rescheduling notice time and length of regular work hours</td>
</tr>
<tr>
<td>( c_{pw}, c_{si}, c_{oi} )</td>
<td>Weight of patient waiting time, surgeon idle time and OR idle time in the objective.</td>
</tr>
<tr>
<td>( c_b, c_{oo}, c_{po} )</td>
<td>Weight of OR blocking time, OR overtime and PACU overtime in the objective.</td>
</tr>
</tbody>
</table>

**Table 5** Sets and parameters

**Appendix C: Gradient Derivation**

To calculate \( \frac{\partial f(\psi, \omega)}{\partial t_{int}(p)} \) and \( \frac{\partial f(\psi, \omega)}{\partial t_{eta}(p)} \), we investigate the propagation of perturbation in the system. First, we predict the impact on the the objective function in Equations (4) and (5) based on conditions in Table 7.
Table 6 Variables

<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(T_{ss}(p), T_{eta}(p))</td>
<td>Scheduled start time and requested patient arrival time of a surgery (p \in P^s).</td>
</tr>
<tr>
<td>(t_{ahead}(p))</td>
<td>(t_{ahead}(p) = T_{ss}(p) - T_{eta}(p), p \in P^s).</td>
</tr>
<tr>
<td>(T_{ss}, T_{eta})</td>
<td>(T_{ss} = {T_{ss}(p) \mid \forall p \in P^s}, T_{eta} = {T_{eta}(p) \mid \forall p \in P^s}).</td>
</tr>
<tr>
<td>(T_{Er}(p))</td>
<td>Expected patient-ready time of a surgery (p \in P^s, T_{Er}(p) = T_{eta}(p) + L_{pE}^{pre}(p)).</td>
</tr>
<tr>
<td>(T_r(p, \omega))</td>
<td>Actual ready time of (p \in P^s) in scenario (\omega; T_r(p, \omega) = T_{Er}(p) + L_{pE}^{pre}(p, \omega)).</td>
</tr>
<tr>
<td>(T_{as}(p, \omega), T_{ap}(p, \omega))</td>
<td>Actual start time and PACU admission time of patient (p \in P^s \cup P^{ef, \omega}) in (\omega).</td>
</tr>
<tr>
<td>(T_{df}(d))</td>
<td>The time when surgeon (d) is ready for surgery.</td>
</tr>
<tr>
<td>(T_{da}(d, \omega), T_{df}(d, \omega))</td>
<td>Total surgical duration and the time of completion for surgeon (d) in scenario (\omega).</td>
</tr>
<tr>
<td>(T_{as}(j, \omega), T_{of}(j, \omega))</td>
<td>Total surgical duration and the time of completion for OR (j) in scenario (\omega).</td>
</tr>
<tr>
<td>(T_{pj}(\omega))</td>
<td>Time of completion for the PACU in scenario (\omega).</td>
</tr>
<tr>
<td>(P_{ej, \omega}, P^{ef, \omega})</td>
<td>Set of emergency surgeries in OR (j) that are counted against the scheduled surgeries in (\omega). Note that we include surgical emergencies that arrive before the last scheduled surgery in the OR is started or cancelled. (P_{ej, \omega} = \bigcup_{1 \leq j} P_{ej, \omega}).</td>
</tr>
</tbody>
</table>

Second, we inspect whether this perturbation impacts the actual surgery start time \((T_{as}(p, \omega))\) in Equation (6) based on conditions in Table 8. Third, the impact on the PACU admission process and the subsequent influence on other patients is measured in Equation (7) based on conditions in Table 9. This analysis is conducted recursively until the propagation stops. Note we only investigate right derivatives because of the almost sure differentiability of \(f(\psi, \omega)\) at a given \(\psi\).

\[
\begin{align*}
\frac{\partial f(\psi, \omega)}{\partial T_{ss}(p)} &= \begin{cases} 
\frac{\partial f(\psi, \omega)}{\partial T_{ss}(p, \omega)} - c_{si} & \text{if } a1 \\
-c_{si} & \text{if } a2 \cup a4 \cup \{b2 \cap (c1 \cup c2 \cup c3)\} \\
0 & \text{if } a3 \cup a5 \cup \{(b3 \cup b4 \cup b5) \cap (c1 \cup c2 \cup c3)\} \\
\frac{\partial f(\psi, \omega)}{\partial T_{ss}(p, \omega)} - c_{si} & \text{if } b1 \cap \{c1 \cup c2 \cup c3\}
\end{cases} \\
\frac{\partial f(\psi, \omega)}{\partial T_{eta}(p)} &= \begin{cases} 
\frac{\partial f(\psi, \omega)}{\partial T_{eta}(p, \omega)} + c_{oi} + c_{ao} & \text{if } b3 \cap c1 \\
\frac{\partial f(\psi, \omega)}{\partial T_{eta}(p, \omega)} + c_{oi} & \text{if } b3 \cap c2 \\
c_{si} + c_{oi} + c_{ao} & \text{if } b4 \cap c3 \\
c_{si} + c_{oi} & \text{if } b4 \cap c3
\end{cases}
\end{align*}
\]

\[
\begin{align*}
\frac{\partial f(\psi, \omega)}{\partial T_{as}(p, \omega)} &= \begin{cases} 
c_{pw} - c_{b} + c_{si} & \text{if } d1 \cap c1 \\
c_{pw} - c_{b} & \text{if } d2 \cap c1 \\
-c_{b} & \text{if } d3 \cap c1 \\
c_{pw} + \frac{\partial f(\psi, \omega)}{\partial T_{ap}(p, \omega)} - c_{b} + c_{si} & \text{if } d1 \cap c2 \\
c_{pw} + \frac{\partial f(\psi, \omega)}{\partial T_{ap}(p, \omega)} - c_{b} & \text{if } d2 \cap c2 \\
\frac{\partial f(\psi, \omega)}{\partial T_{as}(p, \omega)} - c_{b} & \text{if } d3 \cap c2
\end{cases}
\end{align*}
\]
Conditions related to scheduled surgeries ($p \in P^s_{f,\omega}$)

1. $p \in P^s_{f,\omega}$ and $T_{as}(p,\omega) = T_{ss}(p)$
2. $p \in P^s_{f,\omega}$ and $T_{as}(p,\omega) = T_{r}(p,\omega)$, $I^s(p) = 0$ or 1
3. $p \in P^s_{f,\omega}$ and $T_{as}(p,\omega) = T_{r}(p,\omega)$, $I^s(p) > 1$
4. $p \in P^s_{f,\omega}$ and $I^s(p) = 0$ or 1, $T_{as}(p,\omega) \neq T_{r}(p,\omega)$, $T_{as}(p,\omega) \neq T_{ss}(p)$.
5. $p \in P^s_{f,\omega}$ and $I^s(p) > 1$, $T_{as}(p,\omega) \neq T_{r}(p,\omega)$, $T_{as}(p,\omega) \neq T_{ss}(p)$.

1. If a scheduled surgery starts at $T_{as}(p,\omega) = T_{ss}(p)$, increasing its $T_{ss}(p)$ delays $T_{ss}(p,\omega)$ and reduces surgeon idle time; increasing its $T_{eta}(p)$ does not change $T_{ss}(p,\omega)$ and hence reduces patient waiting time.
2. If a surgery starts at $T_{as}(p,\omega) = T_{r}(p,\omega)$, perturbing its $T_{ss}(p)$ does not change $T_{as}(p,\omega)$ but increasing $T_{eta}(p)$ delays $T_{ss}(p,\omega)$. Patient waiting time is not changed in either case.
3. If it is the first case for a surgeon ($I^s(p) = 0$ or 1), delaying its $T_{ss}(p)$ may also reduce surgeon idle time.

Conditions related to cancelled surgeries ($p \in P^{sc,\omega}$)

1. $p \in P^{sc,\omega}$ and $I^s(p) = 1$, $\exists p' \in P^{f,\omega}_{OR}(p) \text{ s.t. } I^d(p') = I^d(p)$ and $T_{as}(p',\omega) = T_{ss}(p)$
2. $p \in P^{sc,\omega}$ and $I^s(p) = 1$, $\exists p' \in P^{f,\omega}_{OR}(p) \text{ s.t. } I^d(p') = I^d(p)$ and $T_{as}(p',\omega) = T_{ss}(p)$
3. $p \in P^{sc,\omega}$ and $I^s(p) = 2$, $\exists p' \in P^{f,\omega}_{OR}(p) \text{ s.t. } I^d(p') = I^d(p)$ and $T_{as}(p',\omega) = T_{r}(p)$
4. $p \in P^{sc,\omega}$ and $I^s(p) = 3$, $T_{eta}(p) = T_{df}(I^d(p,\omega))$
5. $p \in P^{sc,\omega}$ and $p \notin (b1) \cup (b2) \cup (b3) \cup (b4)$

1. If the first case for a surgeon with at least 2 cases is cancelled ($I^s(p) = 1$), delaying its $T_{ss}(p)$ reduces surgeon idle time and may also delay the following surgery of the same surgeon.
2. If a cancelled case is not the first or last case for a surgeon with at least 2 cases ($I^s(p) = 2$), delaying its $T_{eta}(p)$ may delay the following surgery for the same surgeon.
3. If the last case is cancelled for a surgeon with at least 2 cases ($I^s(p) = 3$), delaying its $T_{eta}(p)$ may increase surgeon idle time.
4. If a cancelled case does not fulfill condition (b1) to (b4), changing its $T_{ss}(p)$ or $T_{eta}(p)$ has no impact on $f(\psi,\omega)$.

Conditions related to the last cancelled surgery in an OR

1. $p \in P^{sc,\omega}$ and $T_{eta}(p) = T_{df}(I^{OR}(p,\omega)) \geq t_w$
2. $p \in P^{sc,\omega}$ and $T_{eta}(p) = T_{df}(I^{OR}(p,\omega)) < t_w$
3. $p \in P^{sc,\omega}$ and $T_{eta}(p) < T_{df}(I^{OR}(p,\omega))$

If the last event in an OR is a surgery cancellation, delaying its $T_{eta}(p)$ may increase OR idle time and OR overtime.

**Table 7** Conditions for $\frac{\partial f(\psi,\omega)}{\partial T_{ss}(p)}$ and $\frac{\partial f(\psi,\omega)}{\partial T_{eta}(p)}$ calculations

\[
\frac{\partial f(\psi,\omega)}{\partial T_{ss}(p)} = \begin{cases} 
    c_{oi} + c_{oo} + c_{po} & \text{if } f1 \cap g1 \\
    c_{oi} + c_{oo} + \frac{\partial f(\psi,\omega)}{\partial T_{ss}(p')} & \text{if } f1 \cap g3 \\
    c_{oi} + c_{oo} & \text{if } f1 \cap g4 \\
    c_{oi} + c_{po} & \text{if } f2 \cap g1 \\
    c_{oi} & \text{if } (f2 \cap g2) \cup (f2 \cap g4) \\
    c_{po} + \frac{\partial f(\psi,\omega)}{\partial T_{as}(p')} & \text{if } f2 \cap g3 \\
    \frac{\partial f(\psi,\omega)}{\partial T_{as}(p')} + \frac{\partial f(\psi,\omega)}{\partial T_{as}(p')} & \text{if } f3 \cap g1 \\
    c_{po} & \text{if } (f3 \cap g2) \cup (f3 \cap g4) \\
    0 & \text{if } f4 \cap g1 \\
    \frac{\partial f(\psi,\omega)}{\partial T_{eta}(p')} & \text{if } f4 \cap g3 \\
\end{cases}
\]
Appendix D: Projection Algorithm $\Phi(\psi)$

In Algorithm 1, if a step results in an infeasible solution, $\Phi(\psi)$ in Algorithm 2 revises it to a feasible schedule.
Algorithm 2: \( \Phi(\psi) \)

for \( j = 1 \) to \( n_{OR} \) do
  for \( p \in P^s_j \) do
    if \( p' \in P^s_j \), \( I^{all}(p) = I^{all}(p') + 1 \), \( T_{ss}(p) < T_{ss}(p') \) then
      \( T_{ss}(p) = T_{ss}(p') \)
    else if \( T_{ss}(p) > t_w \) then
      \( T_{ss}(p) = t_w \)
    else
      \( T_{ss}(p) = T_{ss}(p) \)
    end
  end
  if \( p' \in P^s_j \), \( I^{all}(p) = I^{all}(p') + 1 \), \( T_{eta}(p) + L_E^{pre}(p) \leq T_{eta}(p') + L_E^{pre}(p') \) then
    \( T_{eta}(p) = T_{eta}(p') + L_E^{pre}(p') - L_E^{pre}(p) + \delta \) where \( \delta \) is a sufficiently small number
  else
    \( T_{Er}(p) = T_{Er}(p) \)
  end
end
E-companion

Appendix EC.1: Surgical Capacity Allocation and Surgical Sequencing

In this section, we introduce surgical capacity allocation and surgical sequencing policy in practice.

First, surgeons and OR managers reach an agreement on how to divide surgical capacity (OR time) among surgical groups or surgeons. This agreed allocation is implemented for several weeks to a few months until the next allocation is approved (May et al. 2011). The current practice at many hospitals (Erdogan and Denton 2010, Cardoen et al. 2008), including those we have surveyed, is to implement a hybrid policy: some OR time is reserved for specific surgeons or surgical groups within an uninterrupted time block (i.e. block booking, e.g. in Fig. EC.1), and the remaining OR time is used to accommodate non-block surgeons’ case requests and unscheduled cases on a first-come-first-serve basis (i.e. open booking, e.g. in Fig. EC.2).

It is common that time in an OR is divided among multiple surgeons, which is the end result of two practice. First, although more than 50% of the total OR time in our respondent hospitals is reserved for block booking, the size of allotted time blocks could be shorter than 8 hours. Moreover, if a time block is assigned to a group of surgeons, it is further divided among the group. This observation is consistent with those in (Mathias 2011, Pandit et al. 2007, Li et al. 2016). Second, if surgeons cannot book enough surgeries to fully utilize their assigned block time within a certain time period, cases of other surgeons will be placed into the available time in an open booking pattern. As a result, an OR could be shared by multiple surgeons.

Surgical sequencing is determined by surgeons and OR managers. In a reserved OR time block, the block-owner surgeon sequences their surgeries based on various priority rules, such as outpatient first, children first and longest case first (Cardoen et al. 2008). Surgeries scheduled via open booking are sequenced on the FCFS basis (Erdogan and Denton 2010). One exception is that surgeries of the same surgeon are scheduled consecutively in the same OR whenever possible (e.g. Dr. I in Fig. EC.2). Surgical sequencing in practice is not necessarily correlated with surgical duration, which “behave overall as if there were random sequencing” (Marcon and Dexter 2007).
Appendix EC.2: Cancellations and Add-ons

Add-on cases are usually performed at the end of a day after scheduled surgeries in an OR are finished (Marcon and Dexter 2006). On rare occasions, they may be inserted in place of cancelled cases when OR managers are ascertained that the following cases will not be delayed. To model this practice, a composite distribution for the surgical duration is constructed. For example, we can construct a distribution so that a surgery is performed w.p. $p_1$, cancelled and replaced by a type-1 add-on case w.p. $p_2$; cancelled and replaced by a type-2 add-on w.p. $p_3$; and cancelled without replacement w.p. $p_4$. These probabilities can be derived based on historical data.

Appendix EC.3: Illustration of Performance Measures

Fig. EC.3 demonstrates the patient flow in a surgical suite with two ORs and a single PACU bed. Surgeries P1 and P2 are scheduled for Dr. A in OR1. Surgeries P3 and P4 are performed in OR2 by Dr.B and Dr.C respectively. All surgeons are ready to work at the $T_{ss}$ of his or her first surgery; i.e., P1 for Dr.A, P3 for Dr. B and P4 for Dr. C.

Surgery P1 starts at its $T_{ss}$. Patient waiting time is penalized since the patient is ready before $T_{ss}$. After surgery, patient P1 enters the PACU for recovery. OR1 and Dr. A remain idle since Surgery P1 finishes before its expected finish time. According to the “to-follow” practice, surgery P2 starts when the patient is ready even before $T_{ss}$. After surgery, patient P2 enters the PACU.

In OR2, patient P3 waits before the surgery starts. Surgery P3 runs longer than expected, which forces surgery P4 to start late. Thus the idle time of Dr. C and the waiting time of patient P4 are penalized. After surgery, patient P3 enters the PACU and surgery P4 is started. After surgery, patient P4 is blocked in the OR until patient P2 is discharged from the PACU. Since OR2 remains idle during OR blocking, both the blocking time and the OR idle time are penalized. Surgeon idle time is not penalized here since Dr. C is done for the day when surgery P4 is complete. OR 2 operates until the PACU admission of patient P4 and hence the OR overtime is penalized. PACU overtime is penalized since the last patient P4 is discharged after regular hours.

Appendix EC.4: Development of the DEDS

EC.4.1. Notation in the DEDS

An event in this system could be the arrival of a surgical emergency, a scheduled surgery being cancelled, start of a surgery, a patient’s admission into the PACU and a patient’s release from the PACU. A state $s_n$ is defined by sets of pending events; i.e. $s_n = \{A_n, C_{j,n}, W_{j,n}, O_{j,n}, R_{n,j} \mid j \in J\}$. State-related notation in the DEDS is defined as follows.

- $I^e(p, \omega)$: The position of emergency surgery $p$ among all the emergency surgeries in the OR in scenario $\omega$; $p \in P^{e,\omega}$. 
• $e = (p, \delta, \tau, \theta)$: represents a pending event that has not yet occurred in a given state. It contains information about the associated patient $p(e)$, the priority tag $\delta(e)$, the time tag $\tau(e)$ and the type of this event $\theta(e)$. $\tau(e)$ indicates the earliest time when an event could occur. Both $\delta(e)$ and $\tau(e)$ are used to determine the sequence of events.

Type-0 pending events ($\theta(e) = 0$) are associated with arrival of emergency surgeries, where $\tau(e)$ reflects the time of arrival ($\tau(e) = t_{e}(p(e), \omega)$) and priority tags are set as $\delta(e) = -\infty$. Type-1 pending events correspond to cancellation of scheduled surgeries and pending events with $\theta(e) = 2$ and $\theta(e) = 3$ are respectively related to the start of scheduled and emergency surgeries. Scheduled surgeries are performed in their scheduled sequence. Therefore $\delta(e) = I^{all}(p(e)) \geq 0$ for type-1 or type-2 pending events. Time tag $\tau(e)$ of a type-1 pending event is the corresponding expected patient-ready time, while $\tau(e)$ of a type-2 event is the time when the patient and the surgeon are ready. Emergency surgeries have higher medical priority (smaller $\delta(e)$) than scheduled surgeries and they are performed in their arrival sequence. Therefore the priority tag of a type-3 pending event is $\delta(e) = I^{e}(p(e), \omega) - |P'^{e, \omega}| - 1 < 0$. The corresponding $\tau_{n}(e)$ is the emergency arrival time.

Type-4 pending events are associated with patients’ PACU admission. Their $\tau(e)$ are patient’s surgery finish time. Type-5 events correspond to patients’ discharge from the PACU. Their $\tau(e)$ is the time of the PACU discharge. $\tau(e)$ of a type-4 or a type-5 pending event is initialized to $+\infty$ and is updated after the surgery or the PACU recovery is started. The priority tag of a type-4 or a type-5 pending event is $\delta(e) = -\infty$. Different pending events are initialized as follows.

$$e = (p, \delta, \tau, \theta)$$
We assume ORs are open from time 0 for simplicity, though our methodology is capable of handling distinct start times. Accordingly, the initial state \( s_0 \) contains emergency arrivals in all ORs in scenario \( \omega \). Note that emergency surgeries have higher priority than scheduled surgeries and scheduled surgeries are performed in the planned sequence. Therefore pending events in \( W_j,0 \) are sequenced by \( \delta \).

A special case occurs when the next surgery start (with minimal \( \delta \)) in \( W_j,0 \) is earlier than the next surgery cancellation in \( C_{j,n} \). This needs special treatment since both medical priority and chronological orders are involved. Note that emergency surgeries have higher priority \( (\delta < 0) \) than scheduled surgeries and scheduled surgeries are performed in the planned sequence. Therefore pending events in \( W_j,0 \) are sequenced by \( \delta \).

First we determine the next pending event \( e_{j,n}^{CW} \) in an OR \( j \) in state \( s_n \) with both pending surgery starts and pending surgery cancellations \( (W_j,n \cup C_{j,n}) \). This needs special treatment since both medical priority and chronological orders are involved. Note that emergency surgeries have higher priority \( (\delta < 0) \) than scheduled surgeries and scheduled surgeries are performed in the planned sequence. Therefore pending events in \( W_j,0 \) are sequenced by \( \delta \).

A special case occurs when the next surgery start (with minimal \( \delta \)) in \( W_j,0 \) is earlier than the next surgery cancellation in \( C_{j,n} \). Next event in \( W_j,n \cup C_{j,n} \) is the surgery cancellation if it has a higher priority. This ensures that a scheduled surgery is started only after its previous scheduled

\[
\begin{align*}
(p, -\infty, t_c(p, \omega), 0) & \quad \text{(Arrival of surgical emergencies)} \\
(p, I^{all}(p), T_{Er}(p), 1) & \quad \text{(Cancellation of scheduled surgeries)} \\
(p, I^{all}(p), \max \{T_r(p, \omega), T_{dr}(I^d(p))\}, 2) & \quad \text{(Start of scheduled surgeries)} \\
(p, I^{e}(p, \omega) - |P^{e,\omega}| - 1, t_c(p, \omega), 3) & \quad \text{(Start of emergency surgeries)} \\
(p, -\infty, +\infty, 4) & \quad \text{(Patients’ admission into the PACU)} \\
(p, -\infty, +\infty, 5) & \quad \text{(Patients’ discharge from the PACU)}
\end{align*}
\]
surgery has been finished or cancelled. \( e_{CW}^{j,n} \) is derived as shown in Equation (EC.1). Note that if there is a tie in \( \tau \), we break the tie by arbitrarily picking the surgery with a smaller \( \delta \).

\[
e_{CW}^{j,n} = \begin{cases} 
\arg \min_{e \in C_{J,n}} \tau(e) & \text{if } W_{j,n} = \emptyset \\
\arg \min_{e \in W_{j,n}} \delta(e) & \text{if } C_{j,n} = \emptyset \\
\emptyset & \text{if } W_{j,n}, C_{j,n} = \emptyset \\
\arg \min_{e \in C_{J,n}} \tau(e) & \text{if } \tau(\arg \min_{e \in W_{j,n}} \delta(e)) \geq \min_{e \in C_{J,n}} \tau(e) \\
\arg \min_{e \in W_{j,n}} (\tau(e), \omega) & \text{if } \tau(\arg \min_{e \in C_{J,n}} \delta(e)) < \min_{e \in W_{j,n}} \tau(e) \text{ and } \min_{e \in C_{J,n}} \delta(e) \geq \delta(\arg \min_{e \in W_{j,n}} \tau(e)) \\
\arg \min_{e \in W_{j,n}} \delta(e) & \text{if } \tau(\arg \min_{e \in C_{J,n}} \delta(e)) < \min_{e \in W_{j,n}} \tau(e) \text{ and } \min_{e \in C_{J,n}} \delta(e) < \delta(\arg \min_{e \in W_{j,n}} \tau(e)) 
\end{cases} 
\quad (EC.1)
\]

Now we determine the next event \( \hat{e}_{n+1} \) in the DEDS in set \( U_n \). When the PACU is at its full capacity (\( |R_n| = n_{PACU} \)), \( \hat{e}_{n+1} \) cannot be a PACU admission, \( \hat{e}_{n+1} \notin \bigcup_{j \in J} O_{J,n} \). If an OR is occupied (\( O_{j,n} \neq \emptyset \)), \( \hat{e}_{n+1} \) cannot be a surgery start in the corresponding OR, \( \hat{e}_{n+1} \notin W_{j,n} \). When \( O_{j,n} = \emptyset \), a pending event \( e_{CW}^{j,n} \) could occur as shown in Equation (EC.1). Therefore \( U_n \) is defined as follows.

\[
U_n = \begin{cases} 
A_n \cup R_n \cup \bigcup_{O_{j,n} \neq \emptyset, j \in J} C_{J,n} \cup \bigcup_{O_{j,n} = \emptyset, j \in J} e_{CW}^{j,n} & \text{if } |R_n| = n_{PACU} \\
A_n \cup R_n \cup \bigcup_{O_{j,n} \neq \emptyset, j \in J} C_{J,n} \cup \bigcup_{O_{j,n} = \emptyset, j \in J} e_{CW}^{j,n} \cup \bigcup_{j \in J} O_{J,n} & \text{if } |R_n| < n_{PACU} 
\end{cases}
\]

\( \hat{e}_{n+1} \) and \( \gamma_{n+1} \) are determined based on the time tag \( \tau \). Note that if patients are waiting for an OR or a PACU space, the time tag of these pending events could be smaller than \( \gamma_n \), the time when the OR or a PACU bed becomes available. If there is a tie in \( \tau \), we break the tie by arbitrarily picking the surgery with the smallest \( \delta \) or the smallest surgery index (\( I^{al}(p) \) or \( I^e(p, \omega) \)).

\[
\gamma_{n+1} = \max(\min_{e \in U_n} \tau(e), \gamma_n) \\
\hat{e}_{n+1} = \arg \min_{e \in U_n} \tau(e)
\]

The type of \( \hat{e}_{n+1} \) determines the derivation of \( s_{n+1} = \{ A_{n+1}, C_{J,n+1}, W_{J,n+1}, O_{J,n+1}, R_{n+1}, j \in J \} \).

If \( \hat{e}_{n+1} \) is a type-0 emergency arrival event, it is removed from \( A_n \) and a type-3 pending event (start of an emergency surgery) is added to the corresponding \( W_{J,n+1} \).

\[
\text{if } \theta(\hat{e}_{n+1}) = 0 \\
C_{J,n+1} = C_{J,n}, \forall j \in J \\
W_{J,n+1} = \begin{cases} 
W_{J,n} \cup \{(p(\hat{e}_{n+1}), I^e(p(\hat{e}_{n+1}), \omega) - |P^e,\omega| - 1, t_c(p(\hat{e}_{n+1}), \omega), 3)\} & \text{if } j = I^{OR}(p(\hat{e}_{n+1})) \\
W_{J,n} & \text{otherwise} 
\end{cases} \\
O_{J,n+1} = O_{J,n}, \forall j \in J \\
R_{n+1} = R_n \\
A_{n+1} = A_n \setminus \{\hat{e}_{n+1}\}
\]


If \( \hat{e}_{n+1} \) is a type-1 event (cancellation of a scheduled surgery), it is removed from the corresponding \( C_{j,n} \). If there is no pending cancellation or surgery start events in \( C_{j,n} \cup W_{j,n} \), pending emergency arrival events in the corresponding OR are removed from \( A_n \).

\[
\text{if } \theta(\hat{e}_{n+1}) = 1 \\
C_{j,n+1} = \begin{cases} 
C_{j,n} \setminus \{\hat{e}_{n+1}\} & \text{if } j = I^{OR}(p(\hat{e}_{n+1})) \\
C_{j,n} & \text{otherwise}
\end{cases} \\
W_{j,n+1} = W_{j,n}, \forall j \in J \\
O_{j,n+1} = O_{j,n}, \forall j \in J \\
R_{n+1} = R_n \\
A_{n+1} = \begin{cases} 
A_n \setminus \bigcup_{e \in A_n} e & \text{if } C_{j,n} \cup W_{j,n} = \emptyset, j = I^{OR}(p(\hat{e}_{n+1})) \\
A_n & \text{otherwise}
\end{cases}
\]

If \( \hat{e}_{n+1} \) is a type-2 event (start of a scheduled surgery), it is removed from the corresponding \( W_{j,n} \) and a type-4 pending event (PACU admission) is added to the corresponding \( O_{j,n} \). The actual surgery start time of the associated patient is updated accordingly. Pending emergency arrival events in the corresponding OR will be removed from \( A_n \) if \( C_{j,n} \cup W_{j,n} \) is empty.

\[
\text{if } \theta(\hat{e}_{n+1}) = 2 \\
C_{j,n+1} = C_{j,n}, \forall j \in J \\
W_{j,n+1} = \begin{cases} 
W_{j,n} \setminus \{\hat{e}_{n+1}\} & \text{if } j = I^{OR}(p(\hat{e}_{n+1})) \\
W_{j,n} & \text{otherwise}
\end{cases} \\
O_{j,n+1} = \begin{cases} 
O_{j,n} \cup \{(p(\hat{e}_{n+1}), -\infty, \gamma_{n+1} + L^{intra}(p(\hat{e}_{n+1}), \omega), 4)\} & \text{if } j = I^{OR}(p(\hat{e}_{n+1})) \\
O_{j,n} & \text{otherwise}
\end{cases} \\
R_{n+1} = R_n \\
A_{n+1} = \begin{cases} 
A_n \setminus \bigcup_{e \in A_n} e & \text{if } C_{j,n} \cup W_{j,n} = \emptyset, j = I^{OR}(p(\hat{e}_{n+1})) \\
A_n & \text{otherwise}
\end{cases} \\
T_{as}(p(\hat{e}_{n+1}), \omega) = \gamma_{n+1}
\]

If \( \hat{e}_{n+1} \) is a type-3 event (start of an emergency surgery), it is removed from the corresponding \( W_{j,n} \) and a type-4 pending event (PACU admission) is added to the corresponding \( O_{j,n+1} \). The actual surgery start time of the associated patient is updated accordingly.

\[
\text{if } \theta(\hat{e}_{n+1}) = 3
\]
\[
\begin{aligned}
C_{j,n+1} &= C_{j,n}, \forall j \in J \\
W_{j,n+1} &= \begin{cases} W_{j,n} \setminus \{\hat{e}_{n+1}\} & \text{if } j = I^R(p(\hat{e}_{n+1})) \\
W_{j,n} & \text{otherwise} \end{cases} \\
O_{j,n+1} &= \begin{cases} O_{j,n} \cup \{(p(\hat{e}_{n+1}), -\infty, \gamma_{n+1} + L^{\text{intra}}(p(\hat{e}_{n+1}), \omega), 4)\} & \text{if } j = I^R(p(\hat{e}_{n+1})) \\
O_{j,n} & \text{otherwise} \end{cases} \\
R_{n+1} &= R_n \\
A_{n+1} &= A_n \\
T_{as}(p(\hat{e}_{n+1}), \omega) &= \gamma_{n+1}
\end{aligned}
\]

If \( \hat{e}_{n+1} \) is a type-4 event (PACU admission), it is removed from the corresponding \( O_{j,n} \) and a type-5 pending event (PACU discharge) is added to \( R_n \). The PACU admission time of the associated patient is updated accordingly.

\[
\begin{aligned}
\text{if } \theta(\hat{e}_{n+1}) = 4 \\
\begin{aligned}
C_{j,n+1} &= C_{j,n}, \forall j \in J \\
W_{j,n+1} &= W_{j,n}, \forall j \in J \\
O_{j,n+1} &= \begin{cases} O_{j,n+1} = O_{j,n} \setminus \{\hat{e}_{n+1}\} & \text{if } j = I^R(p(\hat{e}_{n+1})) \\
O_{j,n} & \text{otherwise} \end{cases} \\
R_{n+1} &= R_n \cup \{(p(\hat{e}_{n+1}), -\infty, \gamma_{n+1} + L^{\text{post}}(p(\hat{e}_{n+1}), \omega), 5)\} \\
A_{n+1} &= A_n \\
T_{ap}(p(\hat{e}_{n+1}), \omega) &= \gamma_{n+1}
\end{aligned}
\end{aligned}
\]

If \( \hat{e}_{n+1} \) is a type-5 event (PACU discharge), it is removed from \( R_n \).

\[
\begin{aligned}
\text{if } \theta(\hat{e}_{n+1}) = 5 \\
\begin{aligned}
C_{j,n+1} &= C_{j,n}, \forall j \in J \\
W_{j,n+1} &= W_{j,n}, \forall j \in J \\
O_{j,n+1} &= O_{j,n}, \forall j \in J \\
R_{n+1} &= R_n \setminus \{\hat{e}_{n+1}\} \\
A_{n+1} &= A_n
\end{aligned}
\end{aligned}
\]

The DEDS is terminated if all sets of pending events are empty in the \( n_f \)th state; that is,

\[
A_{n_f} \cup \left\{ \bigcup_{j \in J} C_{j,n_f} \bigg\} \cup \left\{ \bigcup_{j \in J} W_{j,n_f} \bigg\} \cup \left\{ \bigcup_{j \in J} O_{j,n_f} \bigg\} \cup R_{n_f} = \emptyset
\]

Other variables can be determined based on outputs of the DEDS. Surgeons are ready for surgery at the \( T_{ss} \) of their first surgery. Total surgical duration of a surgeon is the duration of all their cancelled cases. Surgeons complete all work when all their surgeries are finished or cancelled.

\[
T_{ds}(d, \omega) = \sum_{p \in P_s^{L,d}(p) = d} L^{\text{intra}}(p, \omega)
\]

\[
T_{dr}(d) = \{T_{ss}(p) | p \in P_s, I^d(p) = d, I^s(p) \in \{0, 1\}\}
\]

\[
T_{df}(d, \omega) = \max \left\{ \{T_{as}(p, \omega) + L^{\text{intra}}(p, \omega) | p \in P_s^{L,d}(p) = d\} \cup \{T_{er}(p) | p \in P_s^{C, L}, I^d(p) = d\} \right\}
\]
Work is completed in an OR after all surgeries are either admitted into the PACU or cancelled. Total surgical duration in an OR includes both scheduled and emergency surgeries. PACU work is completed when the last patient is discharged.

\[ T_{os}(j, \omega) = \sum_{p \in P^s, \omega} L_{intra}(p, \omega) \]

\[ T_{sf}(j, \omega) = \max\left\{ \left\{ T_{ap}(p, \omega) \mid p \in P^s, \omega \right\} \cup \left\{ T_{Er}(p) \mid p \in P^e, \omega \right\} \right\} \]

\[ T_{pf}(\omega) = \max\left\{ \left\{ T_{ap}(p, \omega) + L_{post}(p, \omega) \mid p \in P^s, \omega \right\} \right\} \]

According to the definition in §4.1, a surgical schedule denoted as \( \psi = (T_{ss}, T_{eta}) \), can be evaluated in scenario \( \omega \) by Formula (EC.2).

\[
f(\psi, \omega) = c_{pw} \sum_{p \in P^s, \omega} (T_{ss}(p, \omega) - T_r(p, \omega)) + c_{si} \sum_{d \in D} (T_{df}(d, \omega) - T_{dr}(d) - T_{ds}(d, \omega)) + c_{ao} \sum_{j \in J} (T_{of}(j, \omega) - T_{os}(j, \omega)) + c_{b} \sum_{p \in P^s, \omega} (T_{ap}(p, \omega) - T_{as}(p, \omega) - L_{intra}(p, \omega)) + c_{oo} \sum_{j \in J} \max(T_{of}(j, \omega) - t_w, 0) + c_{po} \max(T_{pf}(\omega) - t_w, 0) \tag{EC.2}
\]

**Appendix EC.5: Feasible Region \( \Psi \)**

Feasible region \( \Psi \) consists of the feasible region \( \Psi_{ss} \) for \( T_{ss} \) and the feasible region \( \Psi_{eta} \) for \( T_{eta} \). \( \Psi_{ss} \) is defined as follows: the first surgery in an OR is scheduled at time 0 since we assume that all OR open at time 0. The order of \( T_{ss} \) is the same as the sequence of scheduled surgeries. In addition, all surgeries should be scheduled within the regular work hours. If we define \( n_{all} = \sum_{j \in J} |P^s_j| \), \( \Psi_{ss} \) can be written as follows.

\[
\Psi_{ss} = \begin{cases} 
T_{ss} \in \mathbb{R}^{n_{all}} \\
T_{ss}(p) = 0 \text{ if } I^{all}(p) = 1; \\
T_{ss}(p) \geq T_{ss}(\bar{p}) \text{ if } I^{all}(p) - I^{all}(\bar{p}) > 0, I^{OR}(p) = I^{OR}(\bar{p}); \\
T_{ss}(p) \leq t_w 
\end{cases}
\]

The feasible region \( \Psi_{eta} \) is defined so that the corresponding \( T_{Er} \) is no later than the \( T_{ss} \) and the order of \( T_{Er} \) is the same as the surgery sequence.

\[
\Psi_{eta} = \begin{cases} 
T_{eta} \in \mathbb{R}^{n_{all}} \\
T_{Er}(p) > T_{Er}(\bar{p}), \text{ if } I^{all}(p) - I^{all}(\bar{p}) > 0, I^{OR}(p) = I^{OR}(\bar{p}); \\
T_{Er}(p) \leq T_{ss}(p); \\
T_{Er}(p) = T_{eta}(p) + L_{pre}(p)
\end{cases}
\]
Appendix EC.6: Consistency of Estimators

In this section, we prove the consistency of the SAA estimators. First, we show the objective function $F(\psi)$ is continuous in the feasible region of $\Psi$.

**Lemma EC.1.** $F(\psi)$ is continuous in the feasible region $\Psi$

*Proof* The proof is similar to proofs for Proposition 1 in (Kim et al. 2015) and Proposition 1 in (Bai et al. 2017). $f(\psi, \omega)$ is a.s. continuous at a given $\psi$, that is, $f(\psi + \Delta \psi, \omega) \to f(\psi, \omega)$ a.s. as $\Delta \psi \to 0$. Note that all activities can be finished within a finite time horizon since the number of surgeries is finite and surgery durations and LOS in the PACU are bounded. Therefore $f(\psi, \omega)$ is bounded by a sufficiently large number, given the finite weights in the objective.

Consequently the Dominated Convergence Theorem (e.g. Kim et al. 2015) can be applied to random variables $f(\psi + \Delta \psi, \omega) - f(\psi, \omega)$. For any $\psi, \psi + \Delta \psi \in \Psi$, we have

$$\lim_{\Delta \psi \to 0} [F(\psi + \Delta \psi) - F(\psi)] = \lim_{\Delta \psi \to 0} \{E_\omega [f(\psi + \Delta \psi, \omega) - f(\psi, \omega)]\}$$

(Dominated Convergence Theorem)

$$= E_\omega \left[ \lim_{\Delta \psi \to 0} \{f(\psi + \Delta \psi, \omega) - f(\psi, \omega)\} \right] = 0$$

Since the result above can be applied to any $\psi \in \Psi$, $F(\psi)$ is continuous in $\Psi$. □

Given the continuity of $F(\psi)$, we can prove Lemma EC.2 by following the proof for Theorem 7.48 in (Shapiro et al. 2009).

**Lemma EC.2.** $f_{n\Omega}(\psi) \to F(\psi)$ uniformly on $\Psi$, a.s. as $n\Omega \to \infty$

We define $F^*$ and $\pi^*$ as the optimal objective value and the set of optimal solutions to the stochastic problem. The SAA objective function $f_{n\Omega}(\psi)$ may be discontinuous and therefore we define $f^*_{n\Omega} = \inf_{\psi \in \Psi} [f_{n\Omega}(\psi)]$. Due to the boundedness and compactness of $\Psi$, we can always find a sequence $\{\psi_k\}$ in $\Psi$ such that $\lim_{k \to \infty} f_{n\Omega}(\psi_k) = f^*_{n\Omega}$ and $\psi^*_k = \lim_{k \to \infty} \psi_k$.

Last, we can follow the proof for Theorem 9 in (Kim et al. 2015) and Theorem 5.3 in (Shapiro et al. 2009) to prove Theorem 1.

**Theorem 1.** $f^*_{n\Omega} \to F^*$ and $d(\psi^*_k, \psi^*) \to 0$ a.s. as $n \to \infty$. where $d(x, Y) = \inf_{y \in Y} |x - y|$.

Appendix EC.7: Proof of Differentiability

Nondifferentiability may occur when a perturbation causes a change to the result of some min or max functions in the DEDS, which results in event sequence change or incurrence of overtime penalty (see (Ho and Cao 1991, Fu and Hu 1997) for reference). In this section, we identify cases corresponding to points of nondifferentiability and show that these cases occur w.p.0 for a given schedule $\psi \in \Psi$. Note that we only summarize the cases where two events are involved, but one can easily generalize them to cases with more events.
• $\Xi_1$: If an emergency arrives at $T_{ej}(j,\omega)$, the number of emergency surgeries may be changed after perturbation, because we do not consider emergencies that arrive after all scheduled surgeries have been started or cancelled. This condition may cause a change in the sequence of the following events. However, this occurs w.p.0, since we assume that the arrival of emergency surgeries is a Poisson process.

• $\Xi_2$: Patient $p(e)$ is scheduled right before patient $p(\bar{e})$; that is, $I_{\text{all}}(p(\bar{e})) - I_{\text{all}}(p(e)) = 1$ and $I^\text{OR}(p(e)) = I^\text{OR}(p(\bar{e}))$. Patient $p(\bar{e})$ and the corresponding surgeon are ready for surgery at $t_{\text{same}}$ when $p(e)$ is cancelled at $T_Er(p(e)) = t_{\text{same}}$. If $T_Er(p(e))$ is reduced after perturbation, there is no change in the sample penalty. If $T_Er(p(e))$ is delayed after perturbation, the waiting time of $p(\bar{e})$ is increased. This results in two different one-side derivatives. To show this case occurs w.p.0, we consider two conditions. If $p(\bar{e})$ has the same surgeon as $p(e)$, it requires $t_{\text{same}} = T_Er(p(e)) = T_r(p(\bar{e}))$, which occurs w.p.0 since $T_r(p(\bar{e}))$ depends on the random duration of the pre-op. If $p(\bar{e})$ has a different surgeon from $p(e)$, it requires $t_{\text{same}} = T_Er(p(e)) = \max\{T_r(p(\bar{e})), T_{ss}(\bar{e})\}$, but $T_Er(p(e)) < T_Er(\bar{e}) \leq T_{ss}(\bar{e})$ as defined in $\Psi$. Therefore $\Xi_2$ occurs w.p.0.

• $\Xi_3$: Patient $p(e)$ and the corresponding surgeon both become ready for surgery at $t_{\text{same}}$. The OR becomes available when the previous surgery $p(\bar{e})$ enters the PACU at $t_{\text{same}}$. If the PACU admission time of $p(\bar{e})$ is delayed, patient $p(e)$ has longer waiting time; while if it is moved earlier, there is no change to $p(e)$’s waiting time. The actual patient-ready time depends on the random duration of the pre-op. The time when a surgeon is available depends on the scheduled start time of his or her surgeries and also depends on the random duration of his or her previous surgeries (if any). The PACU admission time of $p(\bar{e})$ depends on the random duration of the surgery and other patients’ LOS in the PACU. Therefore this condition occurs w.p.0.

• $\Xi_4$: If two surgeries $p(e)$ and $p(\bar{e})$ finish at the same time and compete for a PACU bed, a perturbation may change the sequence of these two patients’ admission into the PACU. If the surgery finish time of $p(e)$ is perturbed, the patient who enters the PACU first could be different, which results in a different sequence of the following events. This condition occurs w.p.0. since surgery finish time depends on the random surgical duration.

• $\Xi_5$: If a surgery $p(e)$ is finished when a patient $p(\bar{e})$ is discharged from the PACU at time $t_{\text{same}}$, a perturbation may result in different one-side derivatives. If surgery $p(e)$ is finished earlier after perturbation, patient $p(e)$ may be blocked until a PACU bed becomes available at $t_{\text{same}}$. If $p(e)$ is finished later, $p(e)$ can directly enter the PACU. This condition occurs w.p.0. since surgery finish time depends on the random surgical duration.

• $\Xi_6$: If two patients are discharged from the PACU when a patient is waiting for a PACU bed, a perturbation may result in different one-side derivatives. If a patient is discharged earlier after perturbation, the OR block time is reduced. However, if a patient is discharged later, OR block
time is unchanged. This condition occurs w.p.0. since the PACU discharge time depends on the random LOS in the PACU.

- $\Xi_7$: A patient is admitted into the PACU at the end of regular work hours. Delaying his or her PACU admission time incurs the penalty for OR overtime, while moving it earlier does not, which causes unequal one-sided derivatives. This condition occurs w.p.0. since the PACU admission time depends on the random surgical duration and LOS in the PACU.

- $\Xi_8$: The last PACU discharge occurs at the end of regular work hours. Similar to $\Xi_7$, a perturbation may cause unequal one-sided derivatives. This condition occurs w.p.0. since PACU discharge time depends on the random LOS in the PACU.

Nondifferentiability in $f(\psi, \omega)$ occurs at $\bigcup_{i=1,2,\ldots,8} \Xi_i$, which are shown to occur w.p.0 for a given feasible schedule. Consequently, $f(\psi, \omega)$ is differentiable at any feasible schedule w.p.1.

**Theorem 2.** $f(\psi, \omega)$ is a.s. differentiable at any $\psi \in \Psi$.

**Appendix EC.8: Lower and Upper Bound Estimations in Convergence Test**

In (Linderoth et al. 2006), upper and lower bound estimates of the optimal objective are statistically derived based on the result of (Mak et al. 1999):

$$ E[f^*_{n\Omega}] \leq F^* \leq F(\psi) $$

where $F^*$ is the optimal objective of the stochastic problem and $f^*_{n\Omega}$ is the infimum of the SAA problem with $n\Omega$ scenarios.

The optimality gap $(F(\psi) - F^*)$ is evaluated by $(F(\psi) - E[f^*_{n\Omega}])$, the bias of which can be reduced by increasing $n\Omega$. Therefore different values of $n\Omega$ are tested in the experiment. $E[f^*_{n\Omega}]$ can be evaluated statistically based on $f^*_{n\Omega}$. In addition, $f^*_{n\Omega}$ can be estimated by $f^{nR}_{n\Omega}$, the best SAA objective value obtained by the gradient descent algorithm with $nR$ restarts and $n\Omega$ scenarios, the bias of which can be reduced by increasing $nR$.

To estimate the lower bound $E[f^*_{n\Omega}]$, we first generate $M_L = 20$ independent batches of $n_{\Omega,1}$ scenarios, $\{\omega^i_1, \omega^i_2, \ldots, \omega^i_{n_{\Omega,1}}\}$, $i = 1,2,\ldots,M_L$ and solve $M_L = 20$ SAA problems by the gradient descent algorithm with $nR$ random restarts. Let $f^{nR,\ast}_{i,n\Omega,1}$ and $\psi^{nR,\ast}_{i,n\Omega,1}$ be the best objective value and the corresponding best schedule by solving the SAA problem with $nR$ random restarts in $i$th batch of scenarios. If $\psi^{nR,\ast}_{i,n\Omega,1}$ is optimal to the corresponding SAA problem, a 95% confidence interval for $E[f^*_{n\Omega,1}]$ can be estimated by

$$ L - \frac{t_{\alpha/2,M_L-1} s_L}{\sqrt{M_L}}, L + \frac{t_{\alpha/2,M_L-1} s_L}{\sqrt{M_L}} $$
where

\[ L = \frac{1}{M_L} \sum_{i=1}^{M_L} f_{i,n_{\Omega_1}}^{n_{R^*}} \quad \text{and} \quad s_L = \sqrt{\frac{1}{M_L - 1} \sum_{i=1}^{M_L} (f_{i,n_{\Omega_1}}^{n_{R^*}} - L)^2} \]

Upper bound \( F(\psi) \) can be estimated by the objective value of a feasible schedule. For each feasible solution \( \psi_{i,n_{\Omega_1}}^{n_{R^*}} \) in the lower bound estimation, we could evaluate its objective value by sampling \( M_U = 50 \) independent batches of \( n_{\Omega_2} = 20000 \) scenarios.

\[ U = \frac{1}{M_U} \sum_{i=1}^{M_U} f_{i,n_{\Omega_2}}(\psi_{i,n_{\Omega_1}}^{n_{R^*}}) = \frac{1}{M_U} \sum_{i=1}^{M_U} \sum_{k=1}^{n_{\Omega_2}} f(\psi_{i,n_{\Omega_1}}^{n_{R^*}}, \omega_k^i) \]

We select the \( \psi_{i,n_{\Omega_1}}^{n_{R^*}} \) with the smallest \( U \) value and obtain a new estimate in a newly-generated set of \( M_U = 50 \) batches of \( n_{\Omega_2} = 20000 \) scenarios. Then a 95% confidence interval for \( F(\psi) \) is

\[ \left[ U - \frac{t_{\alpha/2,M_U-1}s_U}{\sqrt{M_U}}, U + \frac{t_{\alpha/2,M_U-1}s_U}{\sqrt{M_U}} \right] \]

where

\[ s_U = \sqrt{\frac{1}{M_U - 1} \sum_{i=1}^{M_U} (f_{i,n_{\Omega_2}}(\psi) - U)^2} \]