ISE

Industrial and
Systems Engineering

A Sample-gradient-based Algorithm for a Multiple-OR and PACU Surgery Scheduling Problem

Miao Bai, Robert H. Storer, and Gregory L. Tonkay

Department of Industrial and Systems Engineering, Lehigh University, USA

ISE Technical Report 16T-004

LEHIGH
UNIVERSITY

A Sample-gradient-based Algorithm for a Multiple-OR and PACU Surgery Scheduling Problem

M. Bai, R.H. Storer, G.L. Tonkay

Industrial and Systems Engineering Department,
Lehigh University, Bethlehem, PA 18015

ABSTRACT

In this paper, we study a multiple-OR surgery scheduling problem constrained by shared PACU capacity within the block-booking framework. With surgery sequences predetermined in each OR, a Discrete Event Dynamic System (DEDS) is devised for the problem. A DEDS-based stochastic optimization model is formulated in order to minimize the cost incurred from patient waiting time, OR idle time, OR blocking time, OR overtime and PACU overtime. A sample-gradient-based algorithm is thus proposed for the sample average approximation of our formulation. Numerical experiments suggest that the proposed method identifies near-optimal solutions and outperforms previous methods. We also show that considerable cost savings are possible in hospitals where PACU beds are a constraint.

1 Introduction

Facing an aging population and increasing expenditures, health care providers are under growing pressure to improve efficiency. As a sector accounting for more than 40% of a hospital's revenue and expenses, operating rooms (ORs) have recently been the focus of research (Gupta & Denton, 2008; Gupta, 2007; Erdogan & Denton, 2010; May et al., 2011). OR efficiency criteria, such as utilization, overtime and punctuality have been found to be "well below the achievable target" in most hospitals (Denton et al., 2007). As an important area to reduce cost and improve overall satisfaction of both patients and service providers, OR scheduling has received particularly intensive research attention.

Block-booking, a widely implemented OR scheduling system, is the framework within which we conduct our study. In the block-booking system, ORs are reserved for each surgeon or specialty within a specific time period or block. For example, a surgeon might be assigned blocks on Tuesdays and Thursdays from 7am to 3pm. After consulting with patients, surgeons can book surgeries into their allotted block time until the block is "full". The scheduled start time for each surgery is subsequently determined prior to the day of surgery with the help of OR managers (Erdogan & Denton, 2010). Patients are informed of their scheduled start times and expected to arrive accordingly. On the day of surgery, time overruns of surgeries and temporary unavailability of resources may occur due to the random durations of surgical activities. In response, OR staff make sequential decisions throughout the day to reschedule surgeries based on the information about ongoing activities. Various performance criteria are evaluated in this decision making process, such as patient waiting time, surgeon idle time and OR overtime (Cardoen et al., 2010).

Although the random nature of surgical care has been considered in OR scheduling, few have yet integrated downstream constraints in a surgical suite into their studies. The Post-Anesthesia Care Unit (PACU), a shared downstream resource for patients to recover from surgery, can be a bottleneck to overall performance of a surgical suite (Iser et al., 2008). Unavailability of PACU beds can result in OR blocking, in which case a patient waits for a PACU bed in the OR after his/her surgery. OR blocking may result in suboptimal patient care, delay of surgeries, overtime work for staff and surgery cancellation, which can be costly to both service providers and patients (Gordon et al., 1988; Jonnalagadda et al., 2005).

We study a stochastic surgery scheduling problem in multiple ORs with PACU constraints within the block-booking framework in this paper. This problem has been an open challenge for years according to Erdogan and Denton (2010). We provide a viable optimization-based solution method for this difficult problem and more importantly we show that considerable cost savings are possible in hospitals where PACU beds are a constraint. In this paper, this multistage stochastic problem is formulated as a DEDS-based stochastic optimization model. The objective is to minimize the expected cost of patient waiting time, surgeon idle time, OR blocking time, OR overtime and PACU overtime. A sample-gradient-based algorithm is proposed to solve the sample average approximation (SAA) of our formulation. Based on Monte Carlo simulation results, our proposed method is compared with stochastic approximation, the mean-value method and other solution strategies. A series of numerical experiments are also conducted to provide more insights into this OR scheduling problem.

The remainder of this paper is organized as follows: A brief review of relevant literature is presented in the next section. In Section 3, we define the problem under study and present a DEDS model. In Section 4, the sample-gradient-based algorithm is proposed. Numerical studies are shown in Section 5. Section 6 concludes the paper with some general remarks.

2 Literature Review

Once surgeons have booked surgeries into blocks in the block-booking system, the OR manager needs to schedule particular start times for the given surgeries. In this section, we focus on literature that investigates two topics: surgery scheduling in ORs without PACU constraints and surgery scheduling in ORs with PACU constraints, assuming that no decision on patient admission is involved. More general reviews on the OR management issues can be found in(Cayirli & Veral, 2003; Gupta, 2007; Gupta & Denton, 2008; Erdogan & Denton, 2010; Cardoen et al., 2010; May et al., 2011).

2.1 PACU not Considered

Given that surgeries have been assigned to ORs and no resources, for example PACUs, are shared among ORs, surgery scheduling in multiple ORs is reduceable to scheduling every OR independently. Various methods have been studied to address surgery scheduling in a single OR. Robinson and Chen (2003) study the problem of scheduling a fixed sequence of surgeries in an OR with the objective to minimize surgeon idle cost and patient waiting cost. They solve the sample average approximation of their stochastic linear programming model and propose a closed-form heuristic to set up surgery start times. Zhang and Xie (2015) study a multiple-OR surgery scheduling problem with a given surgery sequence within the open-scheduling framework. The problem is formulated as a discrete-event simulation-based model by assigning surgeries using the First-Come-First-Served rule. The objective is to minimize the weighted cost of surgeon waiting, OR idling and OR overtime. They prove that the sample cost function is unimodal, Lipschitz continuous and almost surely differentiable and the expected cost function is continuously differentiable. Stochastic approximation (SA) is proposed to solve the problem and numerical experiments demonstrate that SA converges to a unique global minimizer.

Another important stream of research is based on stochastic programming. Denton and Gupta (2003) propose a two-stage stochastic linear programming model with simple recourse to find the appointment time for a given sequence of surgeries with random durations in an OR. To minimize the expected cost of patient waiting, OR idling and overtime, this problem is solved by a variant of standard L-shaped method with sequentially partitioning the duration space. A similar case is studied by Begen and Queyranne (2011), but with overage and underage cost associated with scheduled start times in the objective function. Given that all the durations follow a known discrete integer-value joint distribution, the existence of an integer optimal schedule is guaranteed. It is further proved that, if cost parameters satisfy α-monotonicity, the objective function is L-convex and can be optimized in polynomial time.

Denton et al. (2007),and Mancilla and Storer (2012) both examine the surgery sequencing and scheduling problem in a single OR to minimize the expected cost of patient waiting, OR idling and overtime. With random surgery durations, Denton et al. (2007) present a two-stage mixed-integer linear programming model as the sample-average approximation (SAA) to the original stochastic problem. Simple heuristics are proposed to sequence the surgeries, among which Sequencing-by-Variance gives the best performance in the numerical study. Mancilla and Storer(2012) formally prove that this SAA problem with two scenarios is NP-complete when idle costs are equal but waiting costs are different for each job. A Benders' decomposition-based heuristic is proposed to address the problem, which outperforms the Sequencing-by-Variance heuristic in numeral experiments.

Other studies of interest can be found in (Lamiri et al., 2008; Batun et al., 2011; Begen et al., 2012). It is worth noting that none of the papers mentioned above consider PACU constraints. Realizing that PACU can be a bottleneck to patient flow within many surgical suits, some researchers directly address the PACU issue in their OR scheduling studies.

2.2 PACU Considered

PACU capacity impacts surgery scheduling decisions while surgery schedules will in turn influence the performance of the PACU. Some simulation studies examine the impact of different surgery sequencing rules on PACU staffing or utilization (Marcon & Dexter, 2006; Iser et al., 2008). Marcon and Dexter(2006) test 7 sequencing rules and their influence on ORs and PACU. Longest-Duration-First, a rule that is commonly used in practice, is shown to be one of the worst performing rules.

With the PACU in consideration, ORs are linked by this shared resource and scheduling multiple ORs is more difficult than scheduling multiple independent ORs. To overcome the computational challenges, many studies assume deterministic service times when the PACU is considered. Hsu et al. (2003) formulate a two-stage job shop model for sequencing and scheduling surgeries in multiple ORs. In their study, all service times are assumed deterministic and no OR blocking is allowed (patients are sent to the PACU immediately after surgery). The objective is to minimize the number of PACU nurses and PACU makespan. A tabu search heuristic framework is developed based on solving two subproblems iteratively by a greedy heuristic. Their algorithm is shown to be effective in the experiments with four to six ORs and 12 to 26 surgeries.

Pham and Klinkert (2008) borrow the idea of Job Shop planning to schedule patients with deterministic surgical durations into multiple ORs. Surgery-to-OR assignment and scheduled start time for each surgery are to be optimized. OR blocking is allowed in their research and recovery is not started until a patient enters the PACU. A mixed-integer linear optimization model is constructed to minimize the makespan and meanwhile allot the limited medical resources to competing jobs. As reported by the authors, only small to medium-sized instances can be solved by CPLEX within a reasonable time. Augusto et al. (2010) also consider the problem of scheduling a group of patients with deterministic surgical service times into multiple ORs. In their study, recovery starts in ORs if patients are blocked due to unavailability of PACU beds. A deterministic flowshop is modeled to represent flows of patient between wards, ORs and the PACU. A mixed-integer linear programming model is established to minimize the costs associated with patients' completion times and a Lagrangian relaxation-based heuristic is employed to solve the problem by relaxing all resource-related constraints. Their heuristic is reported to solve problems with 10-30 surgeries within 10 minutes.

Other relevant papers can be found in (Price et al., 2011; Min & Yih, 2010b). To our knowledge, only two papers are closely related to our study in that they consider both randomness of medical service times and PACU constraints. Lee and Yih (2012) investigate the problem of scheduling a pool of surgeries with lognormally-distributed durations into multiple ORs with limited downstream PACU capacity. OR blocking is allowed in their research and recovery starts when the patient enters the PACU. The surgery-to-OR assignment and scheduled start times of the surgeries in each OR are found to minimize the total cost associated with patient waiting, OR blocking, OR idling and OR overtime. A Genetic Algorithm (GA) is proposed to solve the problem. Solutions after 500 GA generations are shown to outperform, in terms of objective value, six sequencing rules used in hospitals such as Longest-Duration-First rule.

Subsequent research by Lee and Yih (2014) determines scheduled start times for surgeries in multiple ORs constrained by limited PACU capacity when the sequence of surgeries in each OR is given. Surgical service times are assumed to be triangular fuzzy numbers (TFNs) and recovery is not started until a patient enters the PACU. They formulate a multi-objective problem and solve it in two stages: minimizing OR blocking time and total completion time of all medical processes in the first stage and then patient waiting and OR idle time in the second stage. The first-stage problem is modeled as a flexible job shop (FJSP)with a satisfaction degree objective function, an approximation of which is solved by a Genetic Algorithm. Five heuristics are presented in the second stage to construct feasible schedules, among which the Newsvendor-based heuristic gives the best performance. The proposed algorithm can achieve shorter patient waiting time in simulation studies compared with the GA method in their previous paper(Lee & Yih, 2012).

In contrast to most of the works cited above, we study a multiple-OR and PACU surgery scheduling problem. Given the sequence of surgeries in each OR, random service duration distributions and limited PACU capacity are integrated into our models. Most of the papers presumably follow the First-Come-First Served (FCFS) rule in their simulation studies. We apply FCFS to formulate a Discrete-Event-Dynamic-System (DEDS) for the problem. The objective is to minimize the expected cost of patient waiting time, OR blocking, surgeon idle time, and OR and PACU overtime. With ideas borrowed from Perturbation Analysis, a sample-gradient-based algorithm is proposed to solve the sample average approximation to our stochastic problem.

3 Problem Statement and Model Development

3.1 Problem Description

We study the problem of determining scheduled start times (SST) for elective surgeries in multiple ORs with shared PACU resources after surgeons have booked surgeries into ORs. In our study, it is assumed that each surgeon is dedicated to one OR over the whole time horizon under study and each will submit a sequence of surgeries to the OR manager. Such sequences are determined by surgeons based on their experience and personal preference, and cannot be altered. It is assumed that all surgeries have to be scheduled within regular work hours. For example surgeries can only be scheduled between 7am and 3pm.

We assume that the first surgery in each OR is scheduled at the beginning of the day (time 0) as in other studies including (Robinson & Chen, 2003; Marcon & Dexter, 2006; Iser et al., 2008). If the first surgery is allowed to start later than time 0 (late-start) and the late start is not penalized, the overall cost can be improved by 0.34% (statistically significant) on average at the expense of 16.4% increase in solution time in our numerical experiments over 100 test problems. We also observe that allowing late-start brings in as much as 15% cost reduction in some test problems. Therefore one may choose to implement a late-start policy based on the particular problem under study and our algorithm can solve the late-start problem with minor modifications. To allow a late-start, SST of the first patient in each OR becomes a decision variable that is allowed to change to positive values. In this paper, we assume the first surgery in each OR is scheduled at time 0.

To accommodate the random nature of medical service times, surgical durations and length of stay (LOS) in the PACU of patients are assumed to follow some known random distributions that are independent from the choice of SST. Surgical durations and LOS in the PACU of different patients are mutually independent. We further assume that these random distributions are truncated, which ensures that surgical durations and LOS in the PACU will be both upper and lower bounded. Truncated random distributions have been adopted in a number of papers to model durations of surgical procedures, such as (Marcon & Dexter, 2006; Denton et al., 2010; Mancilla & Storer, 2012; Begen & Queyranne, 2011; Begen et al., 2012). These truncated random distributions could be constructed based on historical data and patients' conditions. Turnover times (pre-operative and post-operative tasks) are not explicitly formulated, but instead contained in the surgical durations in our formulation.

Informed of the scheduled start time prior to the day of surgery, patients and surgeons are assumed to be punctual at the allotted time on the day of surgery, and thus surgeries cannot be started before their SST. After surgery, patients are transferred to the PACU if a PACU bed is available, otherwise they are blocked in the OR and wait for a spot in the PACU. Patients are released from the PACU after spending LOS in the PACU. It is assumed that both transport from OR to PACU and release from the PACU take no time.

We require that surgeries in each OR are performed in the same sequence as their SST thus no surgery resequencing is allowed. Surgeons are assumed to start working from the beginning of the time horizon (time 0) since the first surgery in an OR is scheduled at time 0 and the patient is assumed punctual. We also assume that every surgery is started as soon as the OR, surgeon and patient are ready to enforce the non-anticipative rule, which requires that decisions do not depend on information from a later time. Furthermore, considering the finite number of patients and bounded surgical durations and LOS in the PACU, the time horizon under study is bounded.

After a surgery is finished, we assume that a patient will be sent to the PACU instantly if there is a PACU bed available and no other patient is blocked. If more than one patient is waiting for a PACU bed, we employ a rule to determine the sequence of PACU admission. Note that gradient derivation is unchanged and the proposed solution approach is still workable no matter what PACU admission rule is adopted, since the gradient is calculated only based on event times in the Discrete Event Dynamic System (DEDS) as we show in Section 3.2.2. The only adjustment required to accommodate different PACU admission rules is to modify the DEDS updating rules (A.0.3) and (A.0.4) in Appendix A.

Three PACU admission rules have been tested, namely First-Come-First-Served (FCFS), sickest-patient-first (SPF), and a practical rule recommended by an anonymous reviewer. According to FCFS, the patient with the earliest surgery finish time will get the next PACU bed if more than one patient is waiting for a PACU bed. SPF makes sure that the patient with the highest medical priority will be admitted into the PACU first if multiple patients are competing for the PACU bed. The reviewer suggests that the patient from the OR with the closest following SST will be admitted to the PACU first and the last patient in an OR gets lower priority into the PACU. In our numerical experiments, the rule suggested by the anonymous reviewer outperforms FCFS by 1.45% (statistically significant) in average solution quality. The reviewer's rule outperforms FCFS by as much as 8.84% in some test problems, while FCFS can be slightly better (1.94%) than the reviewer's rule in some instances. The sickest-patient-first rule (SPF) is implemented with randomly-generated patient medical priorities. Compared with SPF, FCFS achieves 0.55% (statistically significant) improvement in average solution quality. The maximal improvement of FCFS over SPF is 9.71% in some test problems, while there are cases in which SPF outperforms FCFS by 7.65%.

Instead of searching for the best PACU admission rule, we would like to focus our discussion on the surgery scheduling problem. Our solution approach lays the framework for solving the surgery scheduling problem with PACU capacity constraints, which can accommodate different PACU admission rules. FCFS is used as the PACU admission rule in this paper to demonstrate the development of our solution method.

The surgery schedule is evaluated by the expected cost of patient waiting time, surgeon idle time, OR blocking time, OR overtime and PACU overtime. Patient waiting time is the time a patient has to wait between his/her scheduled start time (SST) and actual start time (AST). Surgeon idle time is defined as the time a surgeon is waiting for the start of the next surgery after finishing one. OR blocking time is the time a patient is held in an OR after surgery before being sent to the PACU. OR overtime is calculated for every OR and reflects the total amount of time surgeons have worked past regular work hours. PACU overtime is the time that PACU work has exceeded regular hours.

The example in Figure 1 demonstrates the patient flow on the day of surgery and the associated performance measures. Three patients in OR 1 and two patients in OR 2 are scheduled with 1 shared PACU bed. In OR 1, the first surgery is finished and sent to the PACU earlier than scheduled, so the surgeon is idle while waiting for the start of the second surgery. Though started on time, the second surgery in OR 1 takes longer than scheduled. Therefore Patient 3 has to wait unit until Patient 2 is sent to the PACU. Surgery 3 in OR 1 continues past regular work time, which increases OR overtime.

In OR 2, the first patient is blocked in the OR after surgery because there is no available PACU bed. This patient has to wait until Patient 1 from OR 1 finishes recovery. When Patient 1 in OR 2 is blocked, the next surgery cannot be started even though Patient 2 has already arrived. Therefore the surgeon is idle and Patient 2 has to wait during the period when Patient 1 is blocked. After finishing all surgeries, the surgeon in OR 2 leaves even when Patient 2 is blocked in the OR after surgery.

In the PACU, since the last patient is discharged after the end of regular time, PACU overtime is penalized.

3.2 Continuous-time Model

In this section, the scheduling problem under study is formulated as an optimization model based on a Discrete Event Dynamic System (DEDS). We first define the DEDS model of the scheduling process with a given set of scheduled start times (SST) and a scenario of random surgical durations and LOS in the PACU. Then Perturbation Analysis will be conducted to estimate the gradient of the sample cost function at a given set of SST, which is later used in a sample-gradient-based algorithm to solve the problem.

3.2.1 Discrete Event Dynamic System

First, we would like to define the notations in our DEDS. There are in total orr ORs and the corresponding OR index set is $J = \{0, 1, \ldots, orr - 1\}$. SR_j patients are scheduled in OR $j \in J$ and their indices are $K_j = \{0, 1, \ldots, SR_j - 1\}$. We use y_{jk} to indicate patient or surgery k in OR j, $j \in J, k \in K_j$ and set Y contains all patients or surgeries in ORs. Scenarios are indexed by ω and the index set of all scenarios is Ω. Surgery durations and LOS in the PACU of patient y_{jk} in scenario ω are denoted as d^{ω}_{jk} and p^{ω}_{jk} , respectively.

The time horizon we study is HT, within which all activities can be finished; and regular work time is MT. The total number of PACU beds is *pcap*. Parameters C_{PW}, C_I, C_B, C_O and C_{PO} represent the unit cost of patient waiting time, surgeon idle time, OR blocking time, OR overtime and PACU overtime, respectively. For simplicity, we assume C_{PW} and C_B are identical for different patients, and C_I and C_O are identical for different surgeons and ORs. It is also assumed that ORs and the PACU share the same regular work hours. Our model can, however, be modified to accommodate distinct costs for patients, surgeons and ORs and different work hours for ORs and the PACU.

We denote the set of scheduled start times of all patients as $\overline{SST}$, $\overline{SST} = \left[\overline{SST_{00}}, \ldots, \overline{SST_{jk}}, \ldots\right]$ and $\overline{SST_{jk}}$ is the scheduled start time of patient $y_{jk} \in Y$. The actual surgery start time and PACU admission time of patient y_{jk} in scenario $\omega \in \Omega$ are represented by $\overline{AST^{\omega}_{jk}}$ and $\overline{APT^{\omega}_{jk}}$, respectively. Both AST and APT are determined on the day of surgery in DEDS, as illustrated in Figure 1. We also define PF^{ω} to indicate the time when the last patient leaves the PACU in scenario ω and $PF^{\omega} = \max_{\forall j \in J, k \in K_j} (\overline{APT^{\omega}_{jk}} + p^{\omega}_{jk})$.

Patient waiting time PW^{ω}_{jk} and OR blocking time B^{ω}_{jk} of patient y_{jk} in scenario ω can be calculated as: $PW^{\omega}_{jk} = \overline{AST^{\omega}_{jk}} - \overline{SST_{jk}}$ and $B^{\omega}_{jk} = \overline{APT^{\omega}_{jk}} - \overline{AST^{\omega}_{jk}} - d^{\omega}_{jk}$, where $y_{jk} \in Y$ and $\omega \in \Omega$. Surgeon idle time before y_{jk} in scenario ω is defined as $I^{\omega}_{jk} = \overline{AST^{\omega}_{jk}} - \overline{AST^{\omega}_{j(k-1)}} - d^{\omega}_{j(k-1)}$.

OR overtime is penalized if the last patient in the OR leaves for the PACU after regular hours, and hence overtime of OR j in scenario ω is $O_j^\omega = \max(\overline{APT_{j(SR_j-1)}^\omega} - MT, 0)$. PACU overtime is incurred in scenario ω if the last patient leaves after regular hours and thus $PO^\omega = \max(PF^\omega - MT, 0)$. .

Given a set of SST and a scenario ω, the day of surgery can be formulated as a Discrete Event Dynamic System (DEDS). An event in this system is defined to be a patient's admission into the PACU or release from the PACU. A state in our DEDS is described by sets of patients in different conditions: patients whose predecessors have not entered the PACU, patients who have not entered the PACU but whose predecessors in the OR have been admitted into the PACU, and patients in the PACU. Also every patient is associated with a time stamp in each state to reflect his/her SST, surgery finish time or the time when he/she is released from the PACU.

The states can be updated following the patient flow in ORs and the PACU. The first surgery in each OR is started at the beginning of the day and other surgeries are started in the given sequence throughout the day. A surgery is finished when the patient has spent the length of time in the OR equal to the surgical duration in the given scenario. After surgery, a patient is sent to the PACU if a PACU bed is available; otherwise the patient is blocked in the OR and waits for a released PACU bed. If more than one patient is competing for a PACU bed, they are admitted in the same sequence as their surgery finish times (smallest-index rule to break the tie). When a patient is admitted into the PACU, the surgery following him/her (if any) will be started as soon as possible: if the following patient has arrived, his/her surgery will be started immediately; otherwise the surgery is started at the time he/she arrives, i.e. at his/her scheduled start time. Every patient spends LOS in the PACU in the given scenario before he/she is discharged. The DEDS is terminated when all patients are discharged from the PACU. The mathematical formulation of this DEDS can be found in Appendix A.

Throughout the DEDS, all the patient-related times are determined and hence the sample cost function can be expressed as:

$$C(\overline{SST}, \omega) = C_{PW} \sum_{\substack{j \in J \\ k \in K_j}} PW_{jk}^\omega + C_I \sum_{\substack{j \in J \\ k \in K_j, k \geq 1}} I_{jk}^\omega + C_B \sum_{\substack{j \in J \\ k \in K_j}} B_{jk}^\omega + C_O \sum_{j \in J} O_j^\omega + C_{PO} PO^\omega$$

Our objective is to minimize the expected cost over set Θ,

$$\min_{\overline{SST} \in \Theta} J(\overline{SST}) = \min_{\overline{SST} \in \Theta} E_\omega \left[C(\overline{SST}, \omega) \right] \tag{3.2.1}$$

where Θ is the set of all feasible $\overline{SST}$,

$$\Theta = \left\{ \overline{SST} \in \Re^S, S = \sum_{j \in J} SR_j \,\middle|\, 0 = \overline{SST_{j0}} \leq \overline{SST_{jk}} \leq \cdots \leq \overline{SST_{j(SR_j-1)}} \leq MT, \forall j \in J \right\}.$$

3.2.2 Differentiability of the Sample Cost Function

In this section, the behavior of sample cost function $C(\overline{SST}, \omega)$ is studied using ideas from Perturbation Analysis (PA)(Ho & Cao, 1991; Glasserman, 1991; Fu & Hu, 1997; Fu, 2015). By observing the system behavior for the given $\overline{SST}$, we use PA to analyze the change in the DEDS after a small perturbation is added to $\overline{SST}$. We would like to investigate whether the perturbation to a patient's scheduled start time will bring changes to his/her actual surgery start time and PACU admission time. It is also important to study whether such a perturbation will propagate to other patients in the system. The sample gradients can be derived based on PA. This analysis is not applicable when our DEDS has different event sequences before and after the perturbation, or an overtime cost is triggered after the perturbation. Thus, we need to find out the cases of event sequence change and overtime being triggered, and show that their occurrence has null probability for a given $\overline{SST}$ before analyzing the sample gradients using Perturbation Analysis (PA).

First, note that all performance measures can be determined in terms of event times and the given $\overline{SST}$, and that the sample cost function $C(\overline{SST}, \omega)$ is a linear function of patient waiting time, OR idle time, OR blocking time, OR overtime and PACU overtime. Therefore the sample cost function is a linear function of event times and $\overline{SST}$. It is not difficult to see that every event time is determined by a series of plus operations, min and max functions of terms $\overline{SST_{jk}}$, d^{ω}_{jk} and p^{ω}_{jk}. Consequently, the sample cost function can be written as a piecewise linear function of $\overline{SST_{jk}}$, d^{ω}_{jk} and p^{ω}_{jk}.

Due to this piecewise linearity, a small perturbation on $\overline{SST}$ will change $C(\overline{SST}, \omega)$ linearly in a small enough neighborhood of any $\overline{SST} \in \Theta$ for almost any scenario ω unless a "turning point", a point of nondifferentiability or discontinuity, is hit. Those "turning points" are encountered when perturbing a given $\overline{SST}$ changes the results of the min or max functions involved in the DEDS in some scenarios. It is observed that non-differentiability and discontinuity occur in the case of event sequence change or overtime cost being incurred. Similar observations have been made by Zhang and Xie in scheduling patients into multiple ORs without PACU constraints(Zhang & Xie, 2015). The cases corresponding to "turning points" are written mathematically in Appendix B, a descriptive summary of which is listed as follows:

- Ω_1: If y_{ab} is the patient with the earliest surgery completion time among all patients who have not entered the PACU and y_{lm} is the patient with the earliest PACU discharge among patients in the fully-occupied PACU, then a "turning point" occurs when $\overline{AST_{ab}} + d_{ab} = \overline{APT_{lm}} + p_{lm}$. A perturbation on $\overline{SST}$ may result in $\overline{AST_{ab}} + d_{ab} < \overline{APT_{lm}} + p_{lm}$ and $\overline{AST_{ab}} + d_{ab} > \overline{APT_{lm}} + p_{lm}$, which correspond to two different sequences of states. Patient y_{ab} is blocked until y_{lm} is released from the PACU in the former case, while y_{ab} is directly moved into the PACU after surgery in the latter case, which results in unequal left and right derivatives. Since the sequence of remaining events is recovered after y_{ab} gets into the PACU and y_{lm} is released, the current condition will not result in a discontinuity in the sample function. So Ω_1 corresponds to a nondifferentiable but continuous case for a given $\overline{SST}$.
- Ω_2: This represents the condition where a patient is blocked waiting for a PACU bed and two patients y_{ab} and y_{lm} finish their recovery at the same time, i.e., $\overline{APT_{ab}} + p_{ab} = \overline{APT_{lm}} + p_{lm}$. A perturbation on $\overline{SST}$ may reduce $\overline{APT_{ab}} + p_{ab}$, which reduces the blocking time for the blocked patients. OR blocking time, however, remains the same when $\overline{APT_{ab}} + p_{ab}$ is increased by perturbing $\overline{SST}$. In spite of unequal one-sided derivatives, sequences of remaining events are the same in the perturbed and nominal sample paths after both patients are admitted into the PACU. So Ω_2 describes a nondifferentiable but continuous case for a given $\overline{SST}$.
- Ω_3: This condition describes a case where SST of a patient is the same as his/her predecessor's PACU admission time. Increasing the patient' SST will not change his/her waiting time while reducing the SST will increase his/her waiting time. So a perturbation on $\overline{SST}$ may lead to unequal one-sided derivatives. Since the sequence of events recovered after the patient has his/her surgery started, Ω_3 outlines a continuous but nondifferentiable case.
- Ω_4: This describes a case in which two surgeries y_{ab} and y_{lm} are finished at the same time and compete for a PACU spot, i.e., $\overline{AST_{ab}} + d_{ab} = \overline{AST_{lm}} + d_{lm}$. A perturbation on $\overline{SST}$ may change the sequence of patients into the PACU. For example, patient y_{lm} is blocked and y_{ab} gets into the PACU in the nominal path while y_{ab} is blocked and y_{lm} is admitted into the PACU after perturbation. The change in PACU admission sequence could result

in a different sequence of the remaining events, because the blocked patient cannot be admitted until the next PACU release event. In such a case, an infinitesimal perturbation on $\overline{SST}$ triggers finite differences in patient-related times and the sample cost. Therefore Ω_4 represents a discontinuous case.

- Ω_5: In this condition, a patient is transferred into the PACU at the end of regular work hours. Increasing his/her SST could trigger OR overtime cost while reducing the SST does not incur an overtime penalty, which results in unequal left and right derivatives. Ω_5 does not create discontinuity in the sample cost function, since the sequence of the events is not changed.
- Ω_6: This is a case in which the last patient leaves the PACU at the end of regular work hours. Similar to Ω_5, this is a continuous but nondifferentiable case.

Although we only summarize the conditions in which two patients finish their surgeries or recovery simultaneously in Ω_2 and Ω_4, one can easily generalize them to cases with more than two patients involved.

As shown, local nondifferentiability occurs in the sample path function $C(\overline{SST}, \omega)$ at a given $\overline{SST}$ when $\omega \in \bigcup_{i=1,2,3,5,6} \Omega_i$ and discontinuity occurs when $\omega \in \Omega_4$. In all cases discussed, there are requirements on random surgery durations d^{ω}_{jk} and LOS in the PACU p^{ω}_{jk} of different patients. For example in Ω_1, nondifferentiability occurs at a given $\overline{SST}$ when $\overline{AST_{jk}} + d_{jk} = \overline{APT_{lm}} + p_{lm}$ with some extra conditions. Since it is assumed that surgery durations and LOS in the PACU of different patients follow mutually independent random distributions and they are independent from the choice of SST, Ω_1 occurs with null probability for any given $\overline{SST}$. Similar arguments can be made for other cases to show their null probability of occurrence. Thus $\widetilde{\Omega} = \{\Omega_1 \cup \Omega_2 \cup \Omega_3 \cup \Omega_4 \cup \Omega_5 \cup \Omega_6\}$ happens with probability zero. In $\omega \in \Omega \backslash \widetilde{\Omega}$, there always exists a small enough neighborhood of a given $\overline{SST}$ where the sequence of events is unchanged and every surgery or recovery finishes before or after regular work hours. Consequently, $C(\overline{SST}, \omega)$ is continuous and differentiable at any $\overline{SST} \in \Theta$ with probability one (w.p.1).

Proposition 1. *$C(\overline{SST}, \omega)$ is a.s. differentiable at any $\overline{SST} \in \Theta$.*

In the second part of this section, perturbation analysis (PA) is used to determine the sample gradient at a given $\overline{SST}$. It is known that:

$$\nabla_{\overline{SST}} C(\overline{SST}, \omega) = (\frac{\partial C}{\partial \overline{SST_{00}}}, \ldots, \frac{\partial C}{\partial \overline{SST_{jk}}}, \ldots, \frac{\partial C}{\partial \overline{SST_{(orr-1)(SR_{(orr-1)}-1)}}}) \tag{3.2.2}$$

where $\frac{\partial C}{\partial \overline{SST_{jk}}}$ calculates the sample partial derivative in sample ω.

We will only study the performance of the perturbed system with $(\overline{SST} + \Delta\overline{SST_{jk}})$ and derive right derivatives because of the almost-sure differentiability of $C(\overline{SST}, \omega)$ at any $\overline{SST}$. The choice of perturbation $\Delta\overline{SST_{jk}}$ is small enough so that the sample cost function maintains local linearity.

To calculate partial derivatives of $C(\overline{SST}, \omega)$ w.r.t. $\overline{SST}$, we need to examine how a perturbation on SST of a single patient impacts the whole system. The basic idea is as follows: when a patient's SST is delayed, we determine whether his/her AST is changed accordingly. If his/her AST is affected, the impact on his/her APT is then determined. In the case that a patient's APT is delayed, we need to inspect whether another patient waiting for a PACU bed is affected and also whether his/her following surgery in the OR is delayed. This calculation is conducted in a recursive pattern until the perturbation stops propagation. Accordingly, two rates of change are defined and used in the recursive calculation:

- $\frac{\Delta C}{\Delta \overline{AST_{jk}}}$ demonstrates how much the sample cost will be influenced, if the actual start time (AST) of a patient y_{jk} is postponed without considering his/her SST change.
- $\frac{\Delta C}{\Delta \overline{APT_{jk}}}$ shows the effect of delaying PACU admission time (APT) of y_{jk} on the sample cost without considering the adjustments on his/her SST and AST.

Note that similar to $\frac{\partial C}{\partial \overline{SST_{jk}}}$, we only investigate the effect of a positive perturbation for $\frac{\Delta C}{\Delta \overline{AST_{jk}}}$ and $\frac{\Delta C}{\Delta \overline{APT_{jk}}}$.

To calculate $\frac{\partial C}{\partial \overline{SST_{jk}}}$, one needs to check whether a perturbation on SST of patient y_{jk} will delay his/her AST. If y_{jk} is started later than the SST $(\overline{AST_{jk}} > \overline{SST_{jk}})$, a change in SST will not affect his/her AST but only reduces his/her patient waiting time. If y_{jk} is started punctually as scheduled $(\overline{AST_{jk}} = \overline{SST_{jk}})$, a perturbation on SST will impact his/her AST by the same amount. For example, in Figure 1, a sufficiently small perturbation on SST of Patient 2 in OR 1 will affect his/her AST. A further calculation is needed to examine the impact on the whole system by pushing off his/her AST, denoted as $\frac{\Delta C}{\Delta \overline{AST_{12}}}$. In contrast, perturbation on SST of Patient 3 in OR 1 will not affect his/her AST because his/her surgery cannot be started until the previous surgery is finished.

Therefore $\frac{\partial C}{\partial \overline{SST_{jk}}}$ can be written mathematically as:

$$\frac{\partial C}{\partial \overline{SST_{jk}}} = \begin{cases} -C_{PW} & \text{if } \overline{AST_{jk}} > \overline{SST_{jk}} \\ \dfrac{\Delta C}{\Delta \overline{AST_{jk}}} - C_{PW} & \text{if } \overline{AST_{jk}} = \overline{SST_{jk}} \end{cases}$$

$\frac{\Delta C}{\Delta \overline{AST_{jk}}}$ demonstrates how much the sample cost is influenced, if AST of y_{jk} is delayed without his/her SST being changed. In addition to an increase in the patient waiting time, extra costs can originate from change on two parallel aspects: APT and surgeon idle time.

(A) Whether his/her PACU admission time (APT) will be affected.

(A1) If y_{jk} is currently blocked ($\overline{APT_{jk}} > \overline{AST_{jk}} + d_{jk}$), a small change on AST will change the OR blocking time, but not his/her APT. For example, in Figure 1, a sufficiently small change on AST of Patient 2 in OR 2 will not affect his/her APT because he/she is blocked in the OR after surgery.

(A2) If y_{jk} is not blocked ($\overline{APT_{jk}} = \overline{AST_{jk}} + d_{jk}$), OR blocking time stays zero but APT is changed after the perturbation. Further calculation is needed to examine the impact on the whole system by pushing off his/her APT, i.e. $\frac{\Delta C}{\Delta \overline{APT_{jk}}}$. In Figure 1, perturbation on AST of Patient 2 in OR 1 will affect his/her APT since the APT depends on the time when his/her surgery is finished. We need to examine $\frac{\Delta C}{\Delta \overline{APT_{12}}}$.

(B) Whether surgeon idle time is changed

(B1) If y_{jk} is the last patient in his/her OR ($k = SR_j - 1$), in addition to the impact on APT, AST adjustment will also change his/her surgery finish time and hence alter the surgeon idle time. In Figure 1, a perturbation on AST of Patient 3 in OR 1 will change the time when the surgeon can leave and hence the surgeon idle time.

(B2) If y_{jk} is not the last patient ($k < SR_j - 1$), no extra penalty is applied.

By enumerating all possible pairs of conditions between (A1, A2) and (B1, B2), $\frac{\Delta C}{\Delta \overline{AST_{jk}}}$ can be written as the follows:

$$\frac{\Delta C}{\Delta \overline{AST_{jk}}} = \begin{cases} C_{PW} - C_B & \text{if } (A1) + (B2) \\ C_{PW} - C_B + C_I & \text{if } (A1) + (B1) \\ C_{PW} - C_B + \dfrac{\Delta C}{\Delta \overline{APT_{jk}}} & \text{if } (A2) + (B2) \\ C_{PW} - C_B + C_I + \dfrac{\Delta C}{\Delta \overline{APT_{jk}}} & \text{if } (A2) + (B1) \end{cases}$$

Similarly, $\frac{\Delta C}{\Delta \overline{APT_{jk}}}$ shows the effect of delaying the PACU admission time (APT) of y_{jk} without adjusting his/her SST and AST. OR blocking time will be increased no matter if y_{jk} is blocked or not. Additional cost will be calculated in two parallel aspects: impact on PACU admission and PACU overtime, and impact on surgery start and OR overtime.

- Impact on PACU admission and overtime

(a) If y_{jk} is the last patient discharged from the PACU ($\overline{APT_{jk}} + p_{jk} = PF$), one needs to determine whether a delay on his/her APT will incur the PACU overtime penalty.

(a1) If there is currently PACU overtime ($\overline{APT_{jk}} + p_{jk} = PF \geq MT$), a delay on APT of y_{jk} will increase the PACU overtime penalty. For example, delaying APT of Patient 3 in OR 1 in Figure 1 will result in extra penalty on PACU overtime.

(a2) If there is currently no PACU overtime ($\overline{APT_{jk}} + p_{jk} = PF < MT$), a delay on APT will not trigger the PACU overtime.

(b) If y_{jk} is not the last patient discharged from the PACU ($\overline{APT_{jk}} + p_{jk} < PF$), perturbation on his/her APT might impact another patient waiting for the PACU bed.

(b1) If y_{jk} has a close follower y_{mn} in the PACU queue who cannot enter the PACU until y_{jk} is discharged, ($\overline{APT_{jk}} + p_{jk} < PF$, $\exists m \in J, n \in K_m, s.t.\ \overline{APT_{mn}} = \overline{APT_{jk}} + p_{jk}$). Delaying APT of y_{jk} will postpone APT of y_{mn} and thus the perturbation is propagated. For example, perturbation on APT of Patient 2 in OR 2 in Figure 1 will result in change on APT of Patient 3 in OR 1, who can not enter the PACU until the discharge of Patient 2 in OR 2.

(b2) If y_{jk} has no close follower in the PACU queue ($\overline{APT_{jk}} + p_{jk} < PF$, $\nexists m \in J, n \in K_m, s.t.\ \overline{APT_{mn}} = \overline{APT_{jk}} + p_{jk}$), no extra penalty is applied in the aspect of PACU admission. For example, perturbation on APT of Patient 1 in OR 2 in Figure 1 will not cause changes on other patients' PACU admission.

- Impact on surgery start and OR overtime

(c) If y_{jk} is the last patient in his/her OR ($k = SR_j - 1$), delaying his/her APT may increase OR overtime.

(c1) If surgery of y_{jk} enters the PACU after regular hours ($k = SR_j - 1$, $\overline{APT_{jk}} \geq MT$), OR overtime is extended by increasing his/her APT. For example, a sufficiently small delay on APT of Patient 3 in OR 1 in Figure 1 will cause an increase in OR overtime.

(c2) If surgery of y_{jk} enters the PACU before regular hours ($k = SR_j - 1$, $\overline{APT_{jk}} < MT$), there is no extra OR overtime triggered. For example, perturbation on APT of Patient 2 in OR 2 in Figure 1 will not cause extra OR overtime.

(d) If y_{jk} is not the last patient in his/her OR ($k < SR_j - 1$) , perturbation on his/her APT might impact his/her following patient.

(d1) If y_{jk} has a close follower $y_{j(k+1)}$ in his/her OR whose surgery cannot be started until y_{jk} is transferred into the PACU ($k < SR_j - 1$, $\overline{AST_{j(k+1)}} = \overline{APT_{jk}}$), AST of $y_{j(k+1)}$ is changed accordingly if APT of y_{jk} is perturbed. For example, perturbation on APT of Patient 2 in OR 1 in Figure 1 will cause change to AST of Patient 3 in OR 1.

(d2) If y_{jk} has no close follower in his/her OR ($k < SR_j - 1$, $\overline{AST_{j(k+1)}} > \overline{APT_{jk}}$), the perturbation is not propagated to his/her following patients.

With all possible combinations between (a1, a2, b1, b2) and (c1, c2, d1, d2) included, $\frac{\Delta C}{\Delta \overline{APT_{jk}}}$ is written as follows: (note that (a2)+(c1) is not feasible since (a2) requires $\overline{APT_{jk}} + p_{jk} = PF < MT$ while (c1) requires $\overline{APT_{jk}} \geq MT$)

$$\frac{\Delta C}{\Delta \overline{APT}_{jk}} = \begin{cases} C_O + C_{PO} + C_B & \text{if } (a1) + (c1) \\ C_{PO} + C_B & \text{if } (a1) + (c2) \\ C_{PO} + \dfrac{\Delta C}{\Delta \overline{AST}_{j(k+1)}} + C_B & \text{if } (a1) + (d1) \\ C_{PO} + C_B & \text{if } (a1) + (d2) \\ C_B & \text{if } (a2) + (c2) \\ \dfrac{\Delta C}{\Delta \overline{AST}_{j(k+1)}} + C_B & \text{if } (a2) + (d1) \\ C_B & \text{if } (a2) + (d2) \\ C_O + \dfrac{\Delta C}{\Delta \overline{APT}_{mn}} + C_B & \text{if } (b1) + (c1) \\ \dfrac{\Delta C}{\Delta \overline{APT}_{mn}} + C_B & \text{if } (b1) + (c2) \\ \dfrac{\Delta C}{\Delta \overline{APT}_{mn}} + \dfrac{\Delta C}{\Delta \overline{AST}_{j(k+1)}} + C_B & \text{if } (b1) + (d1) \\ \dfrac{\Delta C}{\Delta \overline{APT}_{mn}} + C_B & \text{if } (b1) + (d2) \\ C_O + C_B & \text{if } (b2) + (c1) \\ C_B & \text{if } (b2) + (c2) \\ \dfrac{\Delta C}{\Delta \overline{AST}_{j(k+1)}} + C_B & \text{if } (b2) + (d1) \\ C_B & \text{if } (b2) + (d2) \end{cases}$$

Based on $\frac{\Delta C}{\Delta \overline{AST}_{jk}}$ and $\frac{\Delta C}{\Delta \overline{APT}_{jk}}$, $\frac{\partial C}{\partial \overline{SST}_{jk}}$ and $\nabla_{\overline{SST}} C(\overline{SST}, \omega)$ in equation (3.2.2) can be calculated recursively. From the derivation of the sample gradients, the local linear performance of the sample cost function is further confirmed. Also, the sample gradients are consistent with our earlier findings that derivative changes when nondifferentiable and discontinuous cases occur with $\omega \in \widetilde{\Omega}$ for a given $\overline{SST}$.

4 SAA-Gradient Descent Algorithm

4.1 Sample Average Approximation

As two of the most widely used stochastic optimization methods, stochastic approximation (SA) and sample average approximation (SAA) are within our consideration. The SA method can be viewed as a stochastic version of the steepest descent algorithm (Chau et al., 2014). An estimator of the gradient of the expected cost function is derived based on a new set of scenarios generated in every iteration and used to find the descent direction(van Ryzin & Vulcano, 2008). It has been extensively studied and applied in various operations research fields. Two recent examples are van Ryzin and Vulcano's study (2008) in network revenue management and Zhang and Xie's research (2015) in surgery appointment scheduling.

SAA converts a stochastic problem into a deterministic counterpart by taking a finite number of samples. It has been shown to be most effective when the expected cost function is continuous(Kim et al., 2015). The theory of SAA is comprehensively covered in (Kleywegt et al., 2002; Ahmed et al., 2002) and it has been widely used in solving OR scheduling problems, including two articles we have discussed (Denton et al., 2007; Mancilla & Storer, 2012).

Typically with sample gradient information obtainable, one would show in SA that sample gradients are unbiased estimators of the gradients of $J(\overline{SST})$. Due to the discontinuity of the sample cost function $C(\overline{SST}, \omega)$, however, it is not possible for us to prove the result. Kim et al.(2015), Fu (1997) and Ho(1991) all give sets of sufficient conditions under which the gradient of a discontinuous sample function is an unbiased estimator, but none of them can be effectively applied in our problem. All of their conditions need information about the probability of discontinuity occurring in $\left[\overline{SST}, \overline{SST} + \Delta\right]$, which would require us to write a complex convolution of multivariate distributions.

In contrast, sample gradients can be directly used in SAA to solve the deterministic problem. We are able to prove the continuity of our expected cost function in Proposition 2 in Appendix C, in which condition SAA is effective(Kim et al., 2015). In addition, SAA is shown to outperform the SA method in our numerical experiments in Section 5.2. Therefore, SAA is adopted to fully exploit the efficiency of perturbation analysis and the continuity of $J(\overline{SST})$.

In SAA, n independent samples $\omega_1, \omega_2, \ldots, \omega_n$ are generated from the distribution of ω and let

$$J_n(\overline{SST}) = \frac{1}{n}\sum_{i=1}^{n} C(\overline{SST}, \omega_i)$$

Then deterministic optimization algorithms can be applied to solve the SAA formulation of the stochastic scheduling problem:

$$\min_{\overline{SST} \in \Theta} J_n(\overline{SST}) \tag{4.1.1}$$

We are able to show the consistency of the SAA estimators of the optimal value and the optimal solutions in Appendix C. This ensures that, when the number of scenarios is sufficiently large, the minimizer or a solution with a near-infimum value of the SAA problem will converge to a minimizer of the stochastic problem w.p.1. In other words, if solutions with near-infimum value could be found in the SAA problem with a large enough number of scenarios, they are a.s.

near-optimal solutions to the stochastic problem. In the next section, an SAA-gradient descent algorithm with random restarts (SAA-GDR) is proposed to find solutions with near-infimum value to our SAA formulation.

4.2 SAA-Gradient Descent Algorithm with Random Restarts (SAA-GDR)

In SAA-GDR (Algorithm 1), a backtracking line search scheme (Nocedal & Wright, 2006) is implemented. Sample gradients are calculated by PA and used to determine the improving direction in each iteration. Step size is dynamically adapted based on initial step size ρ and step size updating factor α. A sufficient decrease percentage requirement is enforced in the line search, i.e., a step is taken only when the improvement in the objective value is no less than a threshold percentage, c. This rule is shown in the computational experiments to be as good as the Armijo rule(Nocedal & Wright, 2006) in the quality of solutions but more efficient in terms of running time. The iteration limit M is selected sufficiently large so that the gradient descent is not terminated while the objective is improving. The search is randomly restarted when a potential local minimum is identified to explore more broadly in the feasible region. The number of scenarios N and the number of random restarts K need to be specified in the algorithm, selection of which will be discussed in Section 5.1.

A move in the steepest descent direction may result in an infeasible solution, in which case Proj($\overline{SST}$) will project the infeasible SST into the feasible region Θ. With all the patients examined in the predetermined sequence in each OR, Proj($\overline{SST}$) will identify and adjust the scheduled start times that violate the predetermined sequence or exceed the regular work hours in Function Proj.

```
for $j = 0$ to $orr - 1$ do
    for $k = 1$ to $SR_j - 1$ do
        if $\overline{SST_{jk}} < \overline{SST_{j(k-1)}}$ then
            $\overline{SST_{jk}} = \overline{SST_{j(k-1)}}$
        else if $\overline{SST_{jk}} > MT$ then
            $\overline{SST_{jk}} = MT$
        else
            $\overline{SST_{jk}} = \overline{SST_{jk}}$
        end
    end
end
```

Function Proj($\overline{SST}$)

It is not possible for us to theoretically prove that the SAA-GDR algorithm identifies optimal solutions in the stochastic problem, considering the discontinuity and nondifferentiability of the sample cost function. Instead, extensive computational results are presented in Section

```
Data: Scenarios $\omega_1, \dots \omega_N$, percentage threshold $c$, initial step size $\rho$, step size updating
      factor $\alpha$, step size threshold $\gamma_s$, iteration limit $M$
Result: Best solution $\overline{SST^*}$ and objective value $J_N^*$
begin
    $J_N^* \leftarrow \infty$; $\overline{SST^*} \leftarrow \mathbf{0}$;
    for $n$=1 to $K$ do
        $\gamma_0 = \rho$;
        Randomly choose $\overline{SST_0} \in \Theta$. Run DEDS with all N scenarios and obtain
            $$J_N(\overline{SST_0}) = \frac{1}{N}\sum_{i=1}^{N} C(\overline{SST_0}, \omega_i)$$
        for $m$=1 to $M$ do
            Calculate the sample gradient at the current SST by Perturbation Analysis:
                $$\nabla J_N(\overline{SST_{m-1}}) = \frac{1}{N}\sum_{i=1}^{N} \nabla C(\overline{SST_{m-1}}, \omega_i)$$
            Update the SST by moving along the steepest descent direction.
                $$\overline{SST_m} = \text{Proj}(\overline{SST_{m-1}} - \gamma_m \frac{\nabla J_N(\overline{SST_{m-1}})}{\left\|\nabla J_N(\overline{SST_{m-1}})\right\|})$$
              where Proj($\overline{SST}$) projects SST into the feasible region $\Theta$;
            Run DEDS and obtain $J_N(\overline{SST_m})$ ;
            Enforce the sufficient decrease percentage requirement;
            if $J_N(\overline{SST_m}) < J_N(\overline{SST_{m-1}}) * (1-c)$ then
                $\gamma_m = \alpha\gamma_{m-1}$ ;
            else
                $J_N(\overline{SST_m}) = J_N(\overline{SST_{m-1}}), \overline{SST_m} = \overline{SST_{m-1}}, \gamma_m = \frac{\gamma_{m-1}}{\alpha}$;
            end
            if $\gamma_m < \gamma_s$ then
                break;
            end
        end
        if $J_N(\overline{SST_M}) < J_N^*$ then
            $J_N^* = J_N(\overline{SST_M})$ and $\overline{SST^*} = \overline{SST_M}$
        end
    end
end
```

Algorithm 1: SAA-gradient descent algorithm with random restarts (SAA-GDR)

5.1 to demonstrate the performance of the SAA-GDR algorithm and verify that it converges to a set of near-optimal points regardless of where the search starts.

5 Numerical Results

The DEDS and SAA-GDR algorithm are implemented in C++ and tested on a compute node with 800MHz AMD Opteron$^{\text{TM}}$ processor 6128 and 32GB memory In our experiments, patients' surgery durations and LOS in the PACU are modeled as independent truncated lognormal distributions. The lognormal distribution has been shown to be effective in surgical process modeling and used in papers including (Marcon & Dexter, 2006; Marcon et al., 2003; Min & Yih, 2010a; Zhang & Xie, 2015; Mancilla & Storer, 2012). Truncated random distributions, as discussed in Section 3.1, are also widely adopted in modeling medical procedures.

Test data in Lee and Yih(2014) is adopted in our experiments, which includes 15 ORs and two to six patients in a given sequence in each OR. Their data is originally given in the form of triangular fuzzy numbers and we convert them to truncated lognormal distributions using an approximation method described in their paper. Each patient has unique surgical duration and LOS in the PACU distributions. After conversion, the mean surgical durations range from 0.5 hours to over 5 hours and the average LOS in the PACU from 0.5 to 2.5 hours. In our experiments, the number of ORs are randomly generated and ORs are arbitrarily drawn from the pool of 15 ORs. Scenarios of surgical durations and LOS in the PACU are generated for patients using the corresponding means and standard deviations in these selected ORs. The number of PACU beds is determined so that the ratios of PACU beds to ORs range from 0.6 to 0.75, similar to the the recommended ratio of 0.7 in (Lee & Yih, 2014).

We conduct experiments with a wide range of cost parameters given the fact that hospitals vary in costs. According to (Raphael et al., 2014), surgery time cost is about $3-3.5$ times more than PACU time in total knee arthroplasty and total hip arthroplasty. Kapur (2001) reports that PACU expenses may be more than 35% of OR costs while expenditure on ORs is 8.9 times more than that on the PACU in the hospital studied by (Macario et al., 1995). Costs associated to ORs are also reported in previous studies: $\frac{C_O}{C_I} = 2$ and $\frac{C_{PW}}{C_I} = 0.25$ in (Glowacka et al., 2009); $\frac{C_O}{C_I} = 1.5$ and $\frac{C_{PW}}{C_I} \in (0.02, 1)$ in (Cayirli et al., 2012); $\frac{C_O}{C_I} = 1.5$ and $\frac{C_{PW}}{C_I} \in (0.01, 1)$ in (Robinson & Chen, 2010; Chen & Robinson, 2014); and $\frac{C_O}{C_{PW}} = 33$ in (Berg et al., 2014). Though we were not able to find literature about OR blocking cost, it is mentioned that OR blocking could be costly to patients and service providers (Gordon et al., 1988; Jonnalagadda et al., 2005; Lee & Yih, 2014) and thus we select its cost within a wide range. Based on the cost analysis in previous literature, cost parameters in our problem are generated randomly as follows:

$$C_I = 1 \text{ (After standardization)}$$

$$C_B = C_I x_2, \text{where } x_2 \sim U(0, 10) \qquad C_O = C_I x_3, \text{where } x_3 \sim U(1.5, 4)$$

$$C_{PO} = C_O x_4, \text{where } x_3 \sim U(0.1, 1) \qquad C_{PW} = C_I x_1, \text{where } x_1 \sim U(0.01, 1)$$

We also assume that ORs and the PACU have the same 8-hour regular work time (MT), though our method could be easily modified to accommodate different length of regular work time for ORs and the PACU as described in Section 3.2.1. Algorithmic parameters are selected in our pilot tests and used in all our numerical experiments: percentage threshold $c = 10^{-5}$, initial step size $\rho = 1$ and step size updating factor $\alpha = 2$.

5.1 SAA-GDR Algorithm

Our first step is to conduct a set of experiments in a similar way as Linderoth et al. (2006), which demonstrates the quality of solutions obtained in our SAA-GDR algorithm. The key idea is to derive an estimate of the optimality gap based on (Mak et al., 1999):

$$E\left[J_N^*\right] \leq J^* \leq J(\overline{SST})$$

where J_N^* is the infimum of the SAA problem with the number of scenario N, $J(\overline{SST})$ is the objective value of a feasible solution $\overline{SST}$ and J^* is the optimal value in the original stochastic problem. In Linderoth et al.'s (2006) experiment, the optimality gap $J(\overline{SST}) - J^*$ is estimated by an over-estimator $J(\overline{SST}) - E\left[J_N^*\right]$, the bias of which can be reduced by increasing the number of scenarios N. Therefore different numbers of scenarios are tested in the experiments. $J(\overline{SST})$ is estimated by the Monte-Carlo sampling method with batches of sufficiently large samples. An estimate of $E\left[J_N^*\right]$ can be derived statistically if we can estimate the infimum J_N^* of the SAA problems. If J_{NK} is the best objective value obtained with K random restarts, the gap between J_N^* and J_{NK} can be narrowed by increasing K. Therefore, we will estimate J_N^* using J_{NK} with a sufficiently large number of random restarts K and further use $E\left[J_{NK}\right]$ as an estimate of $E\left[J_N^*\right]$. The details of this experiment are given in Appendix D.

Random test instances with N scenarios are generated and solved using SAA-GDR with K random restarts. Estimates of the upper bound $J(\overline{SST})$ and the lower bound $E\left[J_N^*\right]$ are derived and part of the results are presented in Tables 1 to 3 with corresponding setup in Appendix E. In all test problems, the relative gap is very small (<0.5%) with $N = 5000$ and $K = 5000$. When the number of scenarios N is increased and the number of random restarts K is unchanged, we observe an increase in the lower bound estimates, which is consistent with our expectation that $J^* - E\left[J_N^*\right]$ vanishes as N increases. When K is increased and N is unchanged, a decrease in the lower bound estimates is observed, as we expect that J_{NK} decreases to J_N^* as K increases. Additionally, we have a similar observation as noted by Linderoth et al. (2006) that the upper bound estimate is almost unaffected by the choice of N and K, suggesting that the solutions obtained with $N = 50$ and $K = 50$ are of similar quality to those with $N = 5000$ and $K = 5000$. This property is particularly helpful when solutions are needed in a limited time frame.

Secondly, we study the convergence of the approximate solutions in the same way as Linderoth et al. did. The results are shown in Appendix F. Similar to Linderoth et al.'s conjecture, it is likely that in our problem, there is a "feasible neighborhood of the solution set with which the objective value is not much different from the optimal value" J^* (Linderoth et al., 2006).

5.2 Comparisons with Other Algorithms

We would like to compare our SAA-GDR algorithm with four current algorithms in terms of running time and solution quality.

The first method is a time-index sample-average-approximation integer formulation (TIndex-SAA) for the problem under study. By defining the length of a basic time interval, the time horizon can be discretized into a finite number of time buckets. If surgery durations and LOS in the PACU are rounded to integer multiples of the time bucket, this problem can be represented as a two-stage time-index model, which can be approximated by TIndex-SAA. TIndex-SAA is difficult to solve and it takes CPLEX more than 10^5 seconds on average to solve a 5-scenario TIndex-SAA with 5-minute time bucket, 5 ORs of in total 19 patients and 3 PACU beds. Various algorithms for solving two-stage stochastic binary problems have also been extensively tested by us on the TIndex-SAA model, including Lagrangian relaxation, L-shaped method and their variants. None of them are able to handle the problem efficiently. In this comparison test, we choose to solve TIndex-SAA with 3 scenarios and 10-minute time buckets, which can be solved by CPLEX within 1 hour in most test instances.

The second algorithm included is the mean-value method (MV), which assumes that all random variables take their mean values. The MV method in our comparison test is a single-scenario TIndex-SAA formulation that takes mean values of surgery durations and LOS in the PACU with 10-minute time buckets.

The third method is scheduling the surgeries without considering PACU constraints (NoPACU). This method can be formulated as a stochastic linear programming problem and approximated by an SAA formulation. The model we use is similar to the one in (Robinson & Chen, 2003) but with overtime cost of ORs. The number of scenarios can be selected using Linderoth et al. (2006)'s experiment. We choose 1000 scenarios based on our judgment in this comparison test.

The fourth algorithm used for comparison is stochastic approximation (SA) with random restarts. We understand that choices of parameters can impact the performance of SA, but tuning parameters for SA is beyond the scope of this paper. Instead, we implement SA similar to the one in Zhang and Xie (2015). 10 scenarios are generated in each iteration and the total number of iterations is $K = 10^5$. Step sizes ρ^k are updated by $\rho^k = a/k$ where k is the current iteration number. Different values of $a = 0.1, 0.5, 1, 5$, used by Zhang and Xie (2015), have been tested in our problem and solutions of the best quality are obtained with $a = 5$, which is thus adopted in our comparisons. Random restarts are included in this method for the sake of a fair comparison. A point is randomly picked in the feasible region Θ to start a new round of SA after the iteration number achieves $K = 10^5$. After 50 random restarts, the solution with the best objective value among 50 is used for comparison.

This comparison test is conducted on random test problems. Service times and mean durations are rounded to integer multiples of 10 minutes in TIndex-SAA and the MV method, respectively. NoPACU uses 1000 scenarios to find a schedule without PACU constraints. We choose $N = 1000$ and $K = 50$ in SAA-GDR based on our judgment.

The performance of different methods are compared in terms of solution time and solution quality. In random test problems, each algorithm is run 20 times on independent batches of scenarios and hence 20 sets of scheduled start times are obtained by each algorithm. Scheduled start times are evaluated by their average performance in DEDS simulations with 50 independent batches of 20000 scenarios. The mean objective value of 20 solutions, $\overline{J}$ are used as the criteria to compare solution quality of different algorithms. Solution quality comparison is shown in Figure 2 and solution times are presented in Table 6 (details in Table 9 in Appendix G) with corresponding setup in Appendix E.

In our comparison tests, SAA-GDR outperforms other methods on average solution quality in problems of different sizes and various cost parameters. Meanwhile, solution time is within 50 seconds for a medium-sized problem with 6 ORs of in total 21 patients and 4 PACU beds.

5.3 Impact of PACU Constraints

We would like to show the impact of PACU constraints on the system relative to schedules that ignore PACU constraints. We obtain 20 schedules from SAA-GDR and 20 schedules from the NoPACU method in each of Test Problem 4 to 12 (setup in Appendix E) and evaluate their performance in DEDS simulations with 50 independent batches of 20000 scenarios. We calculate, for each schedule, the total OR blocking time, total patient waiting time, total surgeon idle time, total OR overtime in all ORs and total PACU overtime. We then report the mean differences between SAA-GDR and NoPACU solutions.

It is observed that after considering PACU constraints, total OR blocking times and total patient waiting times are decreased while total surgeon idle time, total OR overtime and total PACU overtime are increased in all test problems. Patients' scheduled start times in SAA-GDR solutions tend to have more idle time added between surgeries. A possible explanation for these observations is: PACU constraints will trigger OR blocking events that result in delays of surgeries. To reduce the chance of blocking, patients' scheduled start times in SAA-GDR solutions tend to be more spread out, which results in an increased SST span after including the PACU constraints. Consequently, OR blocking time is decreased and surgeries are more likely to be started punctually, i.e., patient waiting time is reduced. Since more idle time added between scheduled start times after considering PACU constraints, surgeon idle time, OR and PACU overtime may be increased as well.

Scheduling with PACU constraints have two-fold benefits. First, by comparison with methods that do not consider PACU constraints, we show in Figure 2 that considerable cost savings are possible in the surgical suites where PACU beds are a constraint. Second, schedules considering PACU constraints maintain a better balance between patients and service providers. From the patients' perspective, waiting times are significantly reduced and blocking after surgeries are also decreased. These changes are good for patients, since excessive patient waiting time results in patient dissatisfaction (Cayirli & Veral, 2003) and OR blocking is critical to patients' health (Lee & Yih, 2014). From the service providers' perspective, schedules considering PACU constraints can be beneficial as well. Though OR and PACU overtimes and surgeon idle times are increased, the amounts of increase are much smaller compared with the significant reductions in patient waiting times and OR blocking times. Moreover, according to (Hall, 2008), improved patient satisfaction can lead to better work efficiency and increased revenue for service providers. Therefore, surgery scheduling that considers PACU capacity can also help the service provider financially in the long term.

6 Conclusions

This paper addresses an open challenge in the field of surgery scheduling: stochastic surgery scheduling in multiple ORs with PACU constraints. The objective is to minimize the expected cost of patient waiting time, surgeon idle time, OR blocking time, OR overtime and PACU overtime. With the surgery sequence predetermined in each OR, this problem is formulated as a stochastic optimization model based on a Discrete Event Dynamic System (DEDS). With sample gradients derived by Perturbation Analysis, an SAA-gradient descent algorithm with random restarts (SAA-GDR) is proposed to solve the SAA of our stochastic optimization model. Numerical experiments are conducted to select sample size and the number of random restarts. It is demonstrated that SAA-GDR with 1000 scenarios and 50 random restarts could identify a near-optimal schedule for more than 20 patients within 1 minute. SAA-GDR is shown to outperform other methods, including the time-indexed model and stochastic approximation method, in terms of solution quality and running time. Lastly, we present the change in the schedule after including PACU constraints and demonstrate that considerable cost savings are possible in the many hospitals where PACU beds are a constraint.

Although the First-Come-First-Served (FCFS) rule is applied in the PACU admission process in our study, problems with other priority-based rules, as discussed in Section 3.1, could also be solved by the SAA-GDR with modifications in the sample-gradient derivation. The proposed algorithm can also solve a broader range of stochastic optimization problems that can be modeled as a DEDS. Typical examples include appointment scheduling problems with multi-stage resource constraints since the customer/patient flow can be modeled as a DEDS. With modifications, SAA-GDR can address surgery scheduling problems with capacity constraints in upstream preparatory rooms and downstream Step-down Units, Intensive Care Units and wards. SAA-GDR can be also useful in production planning. For example, a stochastic flow shop can be modeled as a DEDS and thus some production planning problems in a flowshop can be solved by our methodology.

References

Ahmed, S., Shapiro, A., & Shapiro, E. (2002). The sample average approximation method for stochastic programs with integer recourse. *Submitted for publication*, 1–24.

Augusto, V., Xie, X., & Perdomo, V. (2010). Operating theatre scheduling with patient recovery in both operating rooms and recovery beds. *Computers & Industrial Engineering*, *58*(2), 231–238.

Batun, S., Denton, B. T., Huschka, T. R., & Schaefer, A. J. (2011). Operating room pooling and parallel surgery processing under uncertainty. *INFORMS Journal on Computing*, *23*(2), 220-237.

Begen, M. A., Levi, R., & Queyranne, M. (2012). Technical Note–A Sampling-Based Approach to Appointment Scheduling. *Operations Research*, *60*(3), 675–681.

Begen, M. A., & Queyranne, M. (2011). Appointment Scheduling with Discrete Random Durations. *Mathematics of Operations Research*, *36*(2), 240–257.

Berg, B. P., Denton, B. T., Ayca Erdogan, S., Rohleder, T., & Huschka, T. (2014, apr). Optimal booking and scheduling in outpatient procedure centers. *Computers & Operations Research.*

Cardoen, B., Demeulemeester, E., & Beliën, J. (2010). Operating room planning and scheduling: A literature review. *European Journal of Operational Research*, *201*(3), 921–932.

Cayirli, T., & Veral, E. (2003). Outpatient scheduling in health care: A review of literature. *Production and Operations Management*, *12*(4), 519–549.

Cayirli, T., Yang, K. K., & Quek, S. A. (2012, jul). A Universal Appointment Rule in the Presence of No-Shows and Walk-Ins. *Production and Operations Management*, *21*(4), 682–697.

Chau, M., Fu, M., Qu, H., & Ryzhov, I. (2014). Simulation optimization: A tutorial overview and recent developments in gradient-based methods. In *Simulation conference (wsc), 2014 winter* (p. 21-35).

Chen, R. R., & Robinson, L. W. (2014, sep). Sequencing and Scheduling Appointments with Potential Call-In Patients. *Production and Operations Management*, *23*(9), 1522–1538.

Denton, B. T., & Gupta, D. (2003). A sequential bounding approach for optimal appointment scheduling. *IIE Transactions*, *35*(11), 1003-1016.

Denton, B. T., Miller, A. J., Balasubramanian, H. J., & Huschka, T. R. (2010). Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations Research*, *58*(4-part-1), 802-816.

Denton, B. T., Viapiano, J., & Vogl, A. (2007). Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health care management science*, 13–24.

Erdogan, S. A., & Denton, B. T. (2010). Surgery planning and scheduling. In J. J. Cochran, L. A. Cox, P. Keskinocak, J. P. Kharoufeh, & J. C. Smith (Eds.), *Wiley encyclopedia of operations research and management science.* John Wiley & Sons, Inc.

Fu, M. (2015). Stochastic gradient estimation. In M. C. Fu (Ed.), *Handbook of simulation optimization* (Vol. 216, p. 105-147). Springer New York.

Fu, M., & Hu, J. Q. (1997). *Conditional Monte Carlo: gradient estimation and optimization applications.* Boston: Kluwer Academic.

Glasserman, P. (1991). *Gradient estimation via perturbation analysis.* Boston: Kluwer.

Glowacka, K. J., Henry, R. M., & May, J. H. (2009, mar). A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling. *Journal of the Operational*

Research Society, *60*(8), 1056–1068.

Gordon, T., Paul, S., Lyles, A., & Fountain, J. (1988). Surgical unit time utilization review: Resource utilization and management implications. *Journal of Medical Systems*, *12*(3), 169-179.

Gupta, D. (2007). Surgical suites' operations management. *Production and Operations Management*, *16*(6), 689–700.

Gupta, D., & Denton, B. T. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, *40*(9), 800–819.

Hall, M. F. (2008). Looking to improve financial results? start by listening to patients. *Healthcare Financial Management*, *10*(62), 76-80.

Ho, Y.-C., & Cao, X.-R. (1991). *Perturbation analysis of discrete event dynamic systems.* Boston: Kluwer Academic Publishers.

Hsu, V. N., de Matta, R., & Lee, C.-Y. (2003). Scheduling patients in an ambulatory surgical center. *Naval Research Logistics*, *50*(3), 218–238.

Iser, J. H., Denton, B. T., & King, R. E. (2008). Heuristics for balancing operating room and post-anesthesia resources under uncertainty. In *Proceedings of the 40th conference on winter simulation* (pp. 1601–1608). Winter Simulation Conference.

Jonnalagadda, R., Walrond, E., Hariharan, S., Walrond, M., & Prasad, C. (2005). Evaluation of the reasons for cancellations and delays of surgical procedures in a developing country. *International Journal of Clinical Practice*, *59*(6), 716–720.

Kapur, P. a. (2001). Postanesthesia Care Unit Challenges. *Anesthesia and Analgesia*, 64–69.

Kim, S., & Henderson, S. G. (2008). Mathematics of continuous-variable simulation Optimization. In *Proceedings - winter simulation conference* (pp. 122–132).

Kim, S., Pasupathy, R., & Henderson, S. (2015). A guide to sample average approximation. In M. C. Fu (Ed.), *Handbook of simulation optimization* (Vol. 216, p. 207-243). Springer New York.

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

Lamiri, M., Xie, X., & Zhang, S. (2008). Column generation approach to operating theater planning with elective and emergency patients. *IIE Transactions*, *40*(9), 838–852.

Lee, S., & Yih, Y. (2012). Surgery Scheduling of Multiple Operating Rooms under Uncertainty and Resource Constraints of Post-Anesthesia Care Units. In *Proceedings of the 2012* *industrial and systems engineering research conference.*

Lee, S., & Yih, Y. (2014). Reducing Patient-Flow Delays In Surgical Suites Through Determining Start-Times Of Surgical Cases. *European Journal of Operational Research*(2014).

Linderoth, J., Shapiro, A., & Wright, S. (2006). The empirical behavior of sampling methods for stochastic programming. *Annals of Operations Research*, *142*(1), 215–241.

Macario, A., Vitez, T. S., Dunn, B., & McDonald, T. (1995). Where are the costs in perioperative care? Analysis of hospital costs and charges for inpatient surgical care. *Anesthesiology*, *83*(6), 1138–1144.

Mak, W.-K., Morton, D. P., & Wood, R. (1999). Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters*, *24*(1-2), 47–56.

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

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

Marcon, E., Kharraja, S., Smolski, N., Luquet, B., & Viale, J. P. (2003). Determining the Number of Beds in the Postanesthesia Care Unit: A Computer Simulation Flow Approach. *Anesthesia & Analgesia*(4), 1415–1423.

May, J. H., Spangler, W. E., Strum, D. P., & Vargas, L. G. (2011). The surgical scheduling problem: Current research and future opportunities. *Production and Operations Management*, *20*(3), 392–405.

Min, D., & Yih, Y. (2010a). An elective surgery scheduling problem considering patient priority. *Computers & Operations Research*, *37*(6), 1091–1099.

Min, D., & Yih, Y. (2010b). Scheduling elective surgery under uncertainty and downstream capacity constraints. *European Journal of Operational Research*, *206*(3), 642–652.

Nocedal, J., & Wright, S. J. (2006). *Numerical optimization* (2nd ed.). New York: Springer.

Pham, D.-N., & Klinkert, A. (2008). Surgical case scheduling as a generalized job shop scheduling problem. *European Journal of Operational Research*, *185*(3), 1011–1025.

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

Raphael, D. R., Cannesson, M., Schwarzkopf, R., Garson, L. M., Vakharia, S. B., Gupta, R., &

Kain, Z. N. (2014). Total joint Perioperative Surgical Home: an observational financial review. *Perioperative Medicine*, *3*(1), 6.

Robinson, L. W., & Chen, R. R. (2003). Scheduling doctors' appointments: optimal and empirically-based heuristic policies. *IIE Transactions*, *35*(3), 295-307.

Robinson, L. W., & Chen, R. R. (2010, apr). A Comparison of Traditional and Open-Access Policies for Appointment Scheduling. *Manufacturing & Service Operations Management*, *12*(2), 330–346.

Shapiro, A., Dentcheva, D., & Ruszczynski, A. (2009). *Lectures on stochastic programming.* Society for Industrial and Applied Mathematics.

van Ryzin, G., & Vulcano, G. (2008). Simulation-Based Optimization of Virtual Nesting Controls for Network Revenue Management. *Operations Research*, *56*(4), 865–880.

Zhang, Z., & Xie, X. (2015). Simulation-based optimization for surgery appointment scheduling of multiple operating rooms. *IIE Transactions*.

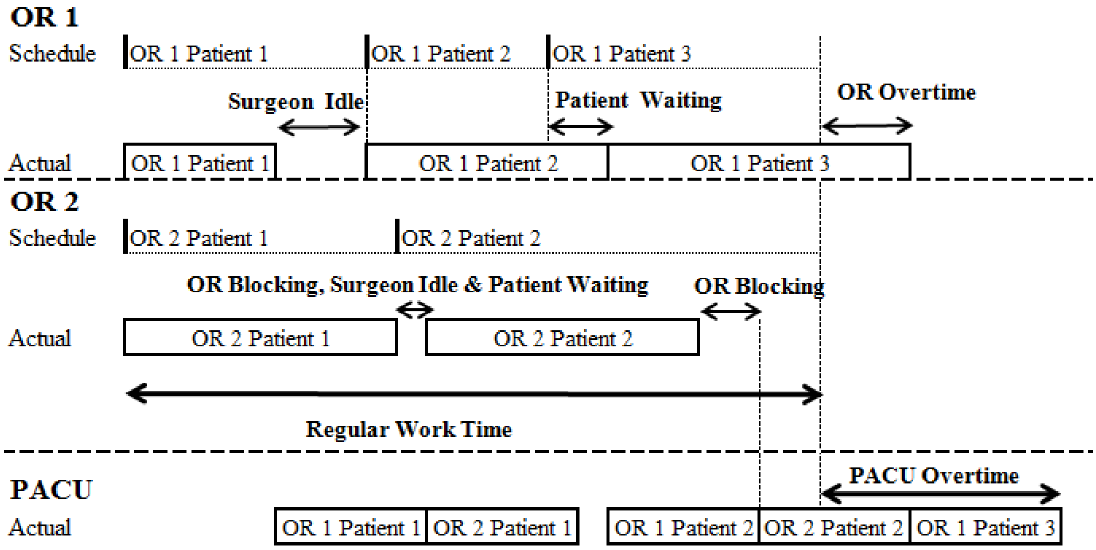

Figure 1: An Example of 2 ORs and 1 PACU bed

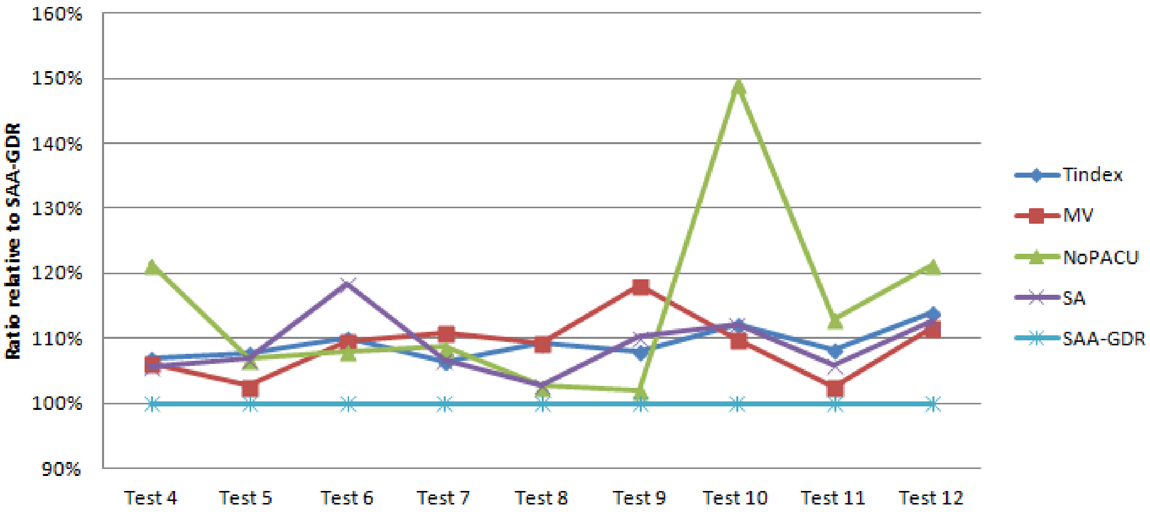

Figure 2: Comparisons on average solution quality in Test Problem 4 to 12

Sample size	Random restarts	Lower bound estimates	Upper bound estimates
N	K	95% confidence interval	95% confidence interval
50	50	9.753 ± 0.197	10.143 ± 0.006
50	100	9.749 ± 0.199	10.152 ± 0.006
50	500	9.740 ± 0.197	10.159 ± 0.006
50	5000	9.727 ± 0.197	10.154 ± 0.007
500	50	10.116 ± 0.049	10.109 ± 0.006
500	100	10.114 ± 0.049	10.116 ± 0.006
500	500	10.110 ± 0.048	10.108 ± 0.006
500	5000	10.106 ± 0.048	10.104 ± 0.006
1000	50	10.086 ± 0.050	10.107 ± 0.007
1000	100	10.085 ± 0.050	10.107 ± 0.006
1000	500	10.082 ± 0.050	10.112 ± 0.006
1000	5000	10.080 ± 0.049	10.106 ± 0.005
5000	50	10.101 ± 0.021	10.103 ± 0.006
5000	100	10.101 ± 0.021	10.105 ± 0.005
5000	500	10.100 ± 0.021	10.102 ± 0.007
5000	5000	10.099 ± 0.021	10.100 ± 0.006

Table 1: Lower and Upper bound estimates for J^* in Test Problem 1

Sample size N	Random restarts K	Lower bound estimates 95% confidence interval	Upper bound estimates 95% confidence interval
50	50	24.927 ± 0.404	26.074 ± 0.013
50	100	24.742 ± 0.370	26.258 ± 0.012
50	500	24.584 ± 0.362	25.959 ± 0.014
50	5000	24.416 ± 0.358	26.072 ± 0.012
500	50	25.684 ± 0.158	25.883 ± 0.014
500	100	25.659 ± 0.173	25.856 ± 0.014
500	500	25.621 ± 0.158	25.855 ± 0.013
500	5000	25.558 ± 0.155	25.838 ± 0.015
1000	50	25.711 ± 0.066	25.910 ± 0.012
1000	100	25.684 ± 0.073	25.887 ± 0.010
1000	500	25.657 ± 0.070	25.877 ± 0.012
1000	5000	25.619 ± 0.072	25.825 ± 0.017
5000	50	25.883 ± 0.052	25.889 ± 0.013
5000	100	25.876 ± 0.049	25.895 ± 0.012
5000	500	25.853 ± 0.046	25.868 ± 0.016
5000	5000	25.834 ± 0.046	25.825 ± 0.014

Table 2: Lower and Upper bound estimates for J^* in Test Problem 2

Sample size	Random restarts	Lower bound estimates	Upper bound estimates
N	K	95% confidence interval	95% confidence interval
50	50	43.057 ± 0.393	43.430 ± 0.013
50	100	42.686 ± 0.376	43.288 ± 0.015
50	500	42.455 ± 0.416	43.067 ± 0.015
50	5000	42.133 ± 0.410	43.057 ± 0.019
500	50	43.374 ± 0.183	42.891 ± 0.015
500	100	43.351 ± 0.191	43.067 ± 0.014
500	500	43.044 ± 0.165	43.084 ± 0.014
500	5000	42.835 ± 0.159	42.801 ± 0.017
1000	50	43.606 ± 0.202	43.256 ± 0.020
1000	100	43.344 ± 0.192	43.157 ± 0.016
1000	500	43.153 ± 0.159	43.133 ± 0.015
1000	5000	43.010 ± 0.144	42.794 ± 0.014
5000	50	43.479 ± 0.150	43.202 ± 0.014
5000	100	43.402 ± 0.131	43.165 ± 0.016
5000	500	43.173 ± 0.061	43.097 ± 0.015
5000	5000	43.082 ± 0.052	42.847 ± 0.016

Table 3: Lower and Upper bound estimates for J^* in Test Problem 3

Methods	Average Solution Time (sec)	Maximal Solution Time (sec)
Tindex	1067.24	2982.06
MV	15.41	45.09
NoPACU	8.76	15.24
SA	985.47	1278.57
SAA-GDR	25.32	48.29

Table 4: Solution times in Test Problem 4 to 12

Appendices

A Discrete Event Dynamic System

Given a set of $\overline{SST}$ and a scenario ω, the day of surgery can be formulated as a Discrete Event Dynamic System (DEDS). An event in this system is defined to be a patient' admission into the PACU or release from the PACU. S denotes the state space of our system, s_n is the n$^{\text{th}}$ state that our DEDS visits and τ_n is the time of the n$^{\text{th}}$ state transition. A state in our DEDS is described by sets of patients in different conditions and time stamps associated with every patient. More specifically, $s_n = (O_{jn}, \forall j \in J, U_n, R_n, c_n)$. O_{jn} is the set of patients in OR $j \in J$ whose predecessors have not entered the PACU in the n$^{\text{th}}$ state. U_n is the set of patients in all ORs who have not entered the PACU but whose predecessors in the OR have been admitted into the PACU in the n$^{\text{th}}$ state. R_n is the set of patients in the PACU in the n$^{\text{th}}$ state. $c_n(y_{jk})$ is the time stamp of patient y_{jk} in the n$^{\text{th}}$ state. The time stamp of a patient in O_{jn} reflects his/her SST; the time stamp of a patient in U_n tells his/her surgery finish time; and the time stamp of a patient in R_n is the time when he/she is released from the PACU.

The flow chart in Figure 3 demonstrates the basic logic of the DEDS and we would like to further explain it with more mathematical details.

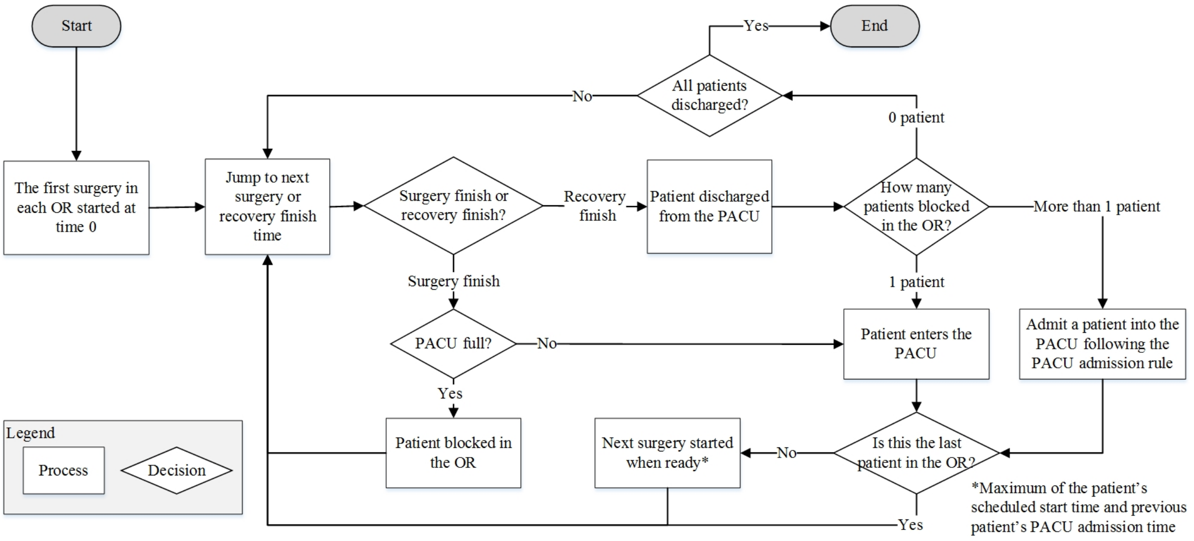
Start
End
Yes
All patients discharged?
No
0 patient
The first surgery in each OR started at time 0
Jump to next surgery or recovery finish time
Surgery finish or recovery finish?
Recovery finish
Patient discharged from the PACU
How many patients blocked in the OR?
More than 1 patient
Surgery finish
1 patient
PACU full?
No
Patient enters the PACU
Admit a patient into the PACU following the PACU admission rule
Yes
Legend
Process
Decision
Patient blocked in the OR
Next surgery started when ready*
No
Is this the last patient in the OR?
Yes
*Maximum of the patient's scheduled start time and previous patient's PACU admission time

Figure 3: Flow chart of the DEDS

The initial state s_0 is defined according to the assumptions we make. The first surgery in each OR is scheduled and started at time 0, that is $\overline{SST_{j0}} = \overline{AST_{j0}} = 0, \forall j \in J$. Therefore at time $\tau_0 = 0$, the system visits state s_0, where $s_0 = (O_{j0}, \forall j \in J, U_0, R_0, c_0)$.

$$\begin{aligned} O_{j0} &= \{y_{jk}, \forall k \in K_j : k > 0\} \\ U_0 &= \{y_{j0}, \forall j \in J\} \\ R_0 &= \phi \end{aligned} \qquad c_0(y_{jk}) = \begin{cases} \overline{SST_{jk}}, & \forall y_{jk} \in O_{jn} \\ d_{jk}, & \forall y_{jk} \in U_n \end{cases} \tag{A.0.1}$$

Functions Φ and Γ update states and transition times of this system by $s_{n+1} = \Phi(s_n)$ and $\tau_{n+1} = \Gamma(\tau_n, s_n)$. To write functions Φ and Γ, two functions need to be defined first:

$$t_1(s_n) = \min\{c_n(y_{jk}) : y_{jk} \in U_n\} \qquad t_2(s_n) = \min\{c_n(y_{jk}) : y_{jk} \in R_n\}$$

where min() breaks tie by choosing the patient from the smallest-indexed OR, as described in the FCFS rule. $t_1(s_n)$ returns the earliest time when a patient in any OR has his/her surgery finished. Such a patient may have his/her surgery finished just at $t_1(s_n)$ or may have been blocked for a certain time. $t_2(s_n)$ returns the earliest time when a patient in the PACU is ready for release.

By comparing $t_1(s_n)$ and $t_2(s_n)$, the next event and the next state of the system can be determined. Therefore the state-updating function Φ and transition-time-updating function Γ can be written in the following conditions:

1. When $t_2(s_n) \le t_1(s_n)$, the next event is a patient's departure from the PACU. This patient is eliminated from the set of patients in the PACU R_n and the time of the next event is the same as the time stamp of him/her. The same results also work in the case when $t_2(s_n) > t_1(s_n)$, but no PACU admission is allowed because of a fully-occupied PACU.

   $$\underline{\text{if } t_2(s_n) \le t_1(s_n) \text{ or } \{t_2(s_n) > t_1(s_n) \text{ and } |R_n| = pcap\}}$$

   $$\begin{aligned} \tau_{n+1} &= t_2(s_n) = \Gamma(\tau_n, s_n) \\ y_{rq} &= \arg\min_{y_{jk}} \{c_n(y_{jk}) : y_{jk} \in R_n\} \\ O_{j(n+1)} &= O_{jn}, \forall j \in J \\ U_{n+1} &= U_n \end{aligned} \qquad \begin{aligned} R_{n+1} &= R_n \setminus \{y_{rq}\} \\ c_{n+1}(y_{jk}) &= \begin{cases} HT, & \text{if } y_{jk} = y_{rq} \\ c_n(y_{jk}) & \text{otherwise} \end{cases} \end{aligned}$$

   $$s_{n+1} = \Phi(s_n) = (O_{j(n+1)}, \forall j \in J, U_{n+1}, R_{n+1}, c_{n+1}) \tag{A.0.2}$$

   where $|R_n|$ is the cardinality of set R_n and y_{rq} is the patient discharged from the PACU. If $O_{j(n+1)} = U_{n+1} = R_{n+1} = \phi, \forall j \in J$, then the system is terminated, and $PF = \tau_{n+1}$.

2. When $t_2(s_n) > t_1(s_n)$ and $|R_n| < pcap$, the next event is transferring a patient into the PACU. The newly-admitted patient will be removed from set U_n and added to set R_n. If

this patient has a follower in his/her OR, i.e.$|O_{rn}| > 0$, this follower will be moved from O_{rn} into U_n. One should note that the time stamp of a patient in set U_n can be smaller than τ_n if he/she is blocked in the OR. Therefore the time of the next event should be determined by the maximum between $t_1(s_n)$ and τ_n.

$$\underline{\text{if } t_2(s_n) > t_1(s_n), |R_n| < pcap \text{ and } |O_{rn}| > 0}$$

$$y_{rq} = \arg\min_{y_{jk}} \{c_n(y_{jk}) : y_{jk} \in U_n\} \qquad O_{r(n+1)} = O_{rn} \setminus \{y_{r(q+1)}\}$$

$$\tau_{n+1} = \max(\tau_n, t_1(s_n)) = \Gamma(\tau_n, s_n) \qquad U_{n+1} = U_n \setminus \{y_{rq}\} \cup \{y_{r(q+1)}\}$$

$$O_{j(n+1)} = O_{jn}, \forall j \in J, j \neq r \qquad R_{n+1} = R_n \cup \{y_{rq}\}$$

$$c_{n+1}(y_{jk}) = \begin{cases} \tau_{n+1} + p_{rq}, & \text{if } y_{jk} = y_{rq} \\ \max(\overline{SST_{r(q+1)}}, \tau_{n+1}) + d_{r(q+1)}, & \text{if } y_{jk} = y_{r(q+1)} \\ c_n(y_{jk}), & \text{otherwise} \end{cases}$$

$$s_{n+1} = (O_{j(n+1)}, \forall j \in J, U_{n+1}, R_{n+1}, c_{n+1}) \tag{A.0.3}$$

where patient y_{rq} is transferred from the OR to the PACU. Note that the time stamp of patient y_{rq} after update indicates his/her PACU discharge time. The new time stamp of $y_{r(q+1)}$ reflects his/her surgery finish time, which is determined by his/her actual start time and surgical duration $d_{r(q+1)}$. The actual start time of $y_{r(q+1)}$ is determined by his/her SST and τ_{n+1}, the time when y_{rq} enters the PACU. Time stamps of other patients are unchanged. Under this condition, it can also be determined that

$$\overline{APT_{rq}} = \tau_{n+1} \qquad \overline{AST_{r(q+1)}} = \max(\overline{SST_{r(q+1)}}, \tau_{n+1})$$

3. When $t_2(s_n) > t_1(s_n)$ and $|R_n| < pcap$, a patient is moved into the PACU, but he/she is the last patient scheduled in his/her OR.

$$\underline{\text{if } t_2(s_n) > t_1(s_n), |R_n| < pcap \text{ and } |O_{rn}| = 0}$$

$$y_{rq} = \arg\min_{y_{jk}} \{c_n(y_{jk}) : y_{jk} \in U_n\}$$

$$\tau_{n+1} = \max(\tau_n, t_1(s_n)) = \Gamma(\tau_n, s_n) \qquad c_{n+1}(y_{jk}) = \begin{cases} \tau_{n+1} + p_{rq}, & \text{if } y_{jk} = y_{rq} \\ c_n(y_{jk}), & \text{otherwise} \end{cases}$$

$$O_{j(n+1)} = O_{jn}, \forall j \in J \qquad R_{n+1} = R_n \cup \{y_{rq}\}$$

$$U_{n+1} = U_n \setminus \{y_{rq}\}$$

$$s_{n+1} = (O_{j(n+1)}, \forall j \in J, U_{n+1}, R_{n+1}, c_{n+1}) \tag{A.0.4}$$

Under this condition, one can determine that $\overline{APT_{rq}} = \tau_{n+1}$.

Combining all conditions, the DEDS can be written as follows:

```
**Input**: $\overline{SST_{jk}}$, $d^{\omega}_{jk}$ and $p^{\omega}_{jk}$ for all $y_{jk} \in Y$ in scenario $\omega$
**Output**: $\overline{AST^{\omega}_{jk}}$, $\overline{APT^{\omega}_{jk}}$ and $PF^{\omega}$ for $y_{jk} \in Y$ in scenario $\omega$
begin
    Initialize $s_0 = (O_{j0}, \forall j \in J, U_0, R_0, c_0)$ based on equations (A.0.1);
    $\tau_0 \leftarrow 0$, $n \leftarrow 1$;
    while $(O_{jn} = U_n = R_n = \phi, \forall j \in J)$ false do
        Update using Equations (A.0.2) to (A.0.4), $s_n = \Phi(s_{n-1})$, $\tau_n = \Gamma(\tau_{n-1}, s_{n-1})$;
        $n \leftarrow n + 1$;
    end
end
```

Procedure Discrete Event Dynamic System

B Nondifferentiability and Discontinuity in the Sample Cost Function

- $\Omega_1 = \{\omega \in \Omega : t_1(s_n) = t_2(s_n), |R_n| = pcap\}$.

  Based on the way AST and APT are determined in DEDS, one can also write this condition as $t_1(s_n) = c_n(y_{ab}) = \overline{AST_{ab}} + d_{ab} = t_2(s_n) = c_n(y_{lm}) = \overline{APT_{lm}} + p_{lm}$ where $y_{ab} \in R_n, y_{lm} \in U_n$. In this condition, the earliest surgery completion time in U_n is the same as the earliest patient release time in R_n when the PACU is full.

- $\Omega_2 = \{\omega \in \Omega : t_2(s_n) = c_n(y_{ab}) = c_n(y_{lm}) > t_1(s_n), |R_n| = pcap, y_{ab}, y_{lm} \in R_n\}$.

  Written as $c_n(y_{ab}) = \overline{APT_{ab}} + p_{ab}$ and $c_n(y_{lm}) = \overline{APT_{lm}} + p_{lm}$, it represents the condition where a patient is blocked waiting for a PACU bed and two patients finish their recovery at the same time.

- $\Omega_3 = \Big\{\omega \in \Omega : c_{n+1}(y_{r(q+1)}) - d_{r(q+1)} = \tau_{n+1} = \overline{SST_{r(q+1)}}, t_2(s_n) > t_1(s_n), |R_n| < pcap,$ $|O_{rq}| > 0, y_{rq} \in U_n\}$

  Written in patient-associated time as $\tau_{n+1} = \overline{APT_{rq}} = \overline{SST_{r(q+1)}}$, the current condition describes the case where SST of a patient is the same as his/her predecessor's PACU admission time. Please note that although patients' arrivals are not explicitly modeled as events in our DEDS, this case is included when event sequence change is discussed.

- $\Omega_4 = \{\omega \in \Omega : t_1(s_n) = c_n(y_{ab}) = c_n(y_{lm}) < t_2(s_n), |R_n| \geq pcap - 1, y_{ab}, y_{lm} \in U_n\}$.

  This describes the case in which two surgeries are finished at the same time and compete for a PACU spot, if we write $c_n(y_{ab}) = \overline{AST_{ab}} + d_{ab}$ and $c_n(y_{lm}) = \overline{AST_{lm}} + d_{lm}$.

- $\Omega_5 = \{\omega \in \Omega : \tau_{n+1} = MT, t_2(s_n) > t_1(s_n), |R_n| < pcap, |O_{rn}| = 0\}$

  Written as $\tau_{n+1} = \overline{APT_{rq}}, y_{rq} \in U_n$, it can be seen that a patient is transferred into the PACU at the end of regular work hours in this condition.

- $\Omega_6 = \Big\{\omega \in \Omega : \tau_{n+1} = MT, O_{j(n+1)} = U_{n+1} = R_{n+1} = \phi, \forall j \in J\Big\}$

  Since the last event in the system is a patient's release from the PACU, one can write $\tau_n = \overline{APT_{rq}} + p_{rq} = MT, y_{rq} \in R_n$.

C Consistency of the SAA estimators

In the following part, we would like to show the consistency of the SAA estimators of the optimal value and the optimal solutions. First, we will show the expected cost function $J(\overline{SST}) = E_\omega\Big[C(\overline{SST}, \omega)\Big]$ is continuous in the feasible region of $\overline{SST}$.

Proposition 2. *The expected cost function $J(\overline{SST}) = E_\omega\Big[C(\overline{SST}, \omega)\Big]$ is continuous in the feasible region Θ, where*

$$\Theta = \left\{\overline{SST} \in \Re^N, N = \sum_{j \in J} SR_j \,\Big|\, 0 = \overline{SST_{j0}} \leq \overline{SST_{jk}} \leq \cdots \leq \overline{SST_{j(SR_j-1)}} \leq MT, \forall j \in J\right\}$$

Proof. The proof is similar to those for Proposition 1 in (Kim et al., 2015) and Proposition 5 in (Kim & Henderson, 2008).

It has been shown that $C(\overline{SST}, \omega)$ is a.s. continuous at a given $\overline{SST}$, that is, $C(\overline{SST} + \Delta, \omega) \to C(\overline{SST}, \omega)$ a.s. as $\Delta \to 0$. Since all activities can be finished within a large enough period HT, $C(\overline{SST}, \omega)$ is bounded by a large number M for any $\overline{SST} \in \Theta$ and any sample $\omega \in \Omega$, where

$$M = \max\{C_{PW}, C_B, C_I, CO, C_{PO}\} * (2\,|Y| + 2 * orr + pcap) * HT$$

and $|Y|$, orr and $pcap$ are the number of patients, ORs and PACU beds, respectively. Because of this boundedness, the Dominated Convergence Theorem (DCT) (c.f. (Kim & Henderson, 2008; Kim et al., 2015)) can be applied to the family of random variables $\Big\{C(\overline{SST} + \Delta, \omega) - C(\overline{SST}, \omega)\Big\}$.

For any $\overline{SST}, \overline{SST} + \Delta \in \Theta$, we have

$$\begin{aligned}
&\lim_{\Delta \to 0}\Big[J(\overline{SST} + \Delta) - J(\overline{SST})\Big] \\
=&\lim_{\Delta \to 0}\Big\{E_\omega\Big[C(\overline{SST} + \Delta, \omega) - C(\overline{SST}, \omega)\Big]\Big\} \\
=&E_\omega\Big[\lim_{\Delta \to 0}\Big\{C(\overline{SST} + \Delta, \omega) - C(\overline{SST}, \omega)\Big\}\Big] = 0
\end{aligned}$$

where the interchange of limit and expectation is justified by the DCT.

Since the result above can be applied to any $\overline{SST} \in \Theta$, $J(\overline{SST})$ is continuous in Θ. □

Based on the continuity of $J(\overline{SST})$, we are able to follow the proof procedure of Shapiro et. al.'s Theorem 7.48 in (Shapiro et al., 2009) to show the objective function of the SAA formulation a.s. uniformly converges to the expected cost function.

Proposition 3. $\left\{J_n(\overline{SST})\right\} \rightarrow J(\overline{SST})$ *uniformly on* Θ*, a.s. as* $n \rightarrow \infty$

Since Θ is bounded and the function $J(\overline{SST})$ is continuous, the minimum can be attained in the feasible region. Let J^* be the optimal objective value of the stochastic problem (3.2.1) and π^* denote the set of corresponding optimal solutions. In the SAA problem (4.1.1), the objective function $J_n(\overline{SST})$ may be discontinuous and might not attain the minimum in Θ. Instead, we define $J_n^* = \inf\left[J_n(\overline{SST})\right]$. Due to the boundedness of Θ, one could always find a sequence $\left\{\overline{SST_k}\right\}$ in Θ that $\lim_{k\rightarrow\infty} J_n(\overline{SST_k}) = J_n^*$. Since Θ is compact, $\lim_{k\rightarrow\infty} \overline{SST_k} = \overline{SST_0}$.

Then we borrow the methods in proving Theorem 9 in (Kim et al., 2015) and Theorem 5.3 in (Shapiro et al., 2009) to show the consistency of the SAA estimators of the optimal value and the optimal solutions based on the previous two propositions.

Proposition 4. $J_n^* \rightarrow J^*$ *and* $d(\overline{SST_0}, \pi^*) \rightarrow 0$ *a.s. as* $n \rightarrow \infty$*. where* $d(x, B) = \inf_{y\in B} |x - y|$.

D Lower and Upper Bounds Derivation in Section 5.1

We first generate $M = 20$ independent batches of scenarios, $\omega_1^j, \omega_2^j, \ldots, \omega_N^j$, $j = 1, 2, \ldots, M$, each of size N and solve $M = 20$ SAA problems

$$\min_{\overline{SST}\in\Theta} \left\{ J_N^j(\overline{SST}) = \frac{1}{N}\sum_{i=1}^{N} C(\overline{SST}, \omega_i^j) \right\}, j = 1, 2, \ldots, M \tag{D.0.1}$$

by SAA-GDR with K random restarts. Let J_{NK}^j be the best objective value by solving (D.0.1) with K random restarts and sst_{NK}^j be the corresponding solution in each of the $M = 20$ SAA problems. According to (Linderoth et al., 2006), if J_{NK}^j is optimal to (D.0.1), a 95% confidence interval for LB_N, a lower bound on the optimal value J^* of the stochastic problem (3.2.1) can be found by

$$\left[L_{NK} - \frac{t_{\alpha/2,M-1}s_{NK}^{lb}}{\sqrt{M}}, L_{NK} + \frac{t_{\alpha/2,M-1}s_{NK}^{lb}}{\sqrt{M}}\right]$$

where

$$L_{NK} = \frac{1}{M}\sum_{j=1}^{M} J_{NK}^j \qquad \text{and} \qquad s_{NK}^{lb} = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}(J_{NK}^j - L_{NK})^2}$$

Upper bound estimates on the optimal value J^* are also obtained as described in (Linderoth et al., 2006). For each solution sst^j_{NK} in the lower bound derivation, one could calculate

$$U_{\overline{N}}(sst^j_{NK}) = \frac{1}{T}\sum_{i=1}^{T} J^i_{\overline{N}}(sst^j_{NK}) = \frac{1}{T}\sum_{i=1}^{T}\sum_{l=1}^{\overline{N}} C(sst^j_{NK}, \omega^i_l)$$

by sampling $T = 50$ independent batches of $\overline{N} = 20000$ scenarios. If we select the sst^j_{NK} with the smallest $U_{\overline{N}}(sst^j_{NK})$ value and generate a new and independent set of $T = 50$ batches of $\overline{N} = 20000$ scenarios, a new estimate of $U_{\overline{N}}(sst^j_{NK})$ is obtained. Then a 95% confidence interval for upper bound estimate $UB_{\overline{N}}$ is

$$\left[U_{\overline{N}}(sst^j_{NK}) - \frac{t_{\alpha/2,T-1}s^{ub}_{\overline{N}}}{\sqrt{T}}, U_{\overline{N}}(sst^j_{NK}) + \frac{t_{\alpha/2,T-1}s^{ub}_{\overline{N}}}{\sqrt{T}}\right]$$

where

$$s^{ub}_{\overline{N}} = \sqrt{\frac{1}{T-1}\sum_{i=1}^{T}(J^i_{\overline{N}}(sst^j_{NK}) - U_{\overline{N}}(sst^j_{NK}))^2}$$

E Setup of the Experiments in Section 5.1 and 5.2

Problem	No. of ORs	Total Patients	No. of PACU beds	C_{PW}	C_I	C_B	C_O	C_{PO}
Test 1	4	15	3	0.3	0.8	2	1.5	1.2
Test 2	6	21	4	0.2	0.5	5	2	1.7
Test 3	5	23	3	0.1	1.5	2	2	1.7
Test 4	5	19	3	0.3	0.8	2	1.5	1.2
Test 5	3	11	2	0.3	0.8	2	1.5	1.2
Test 6	6	21	4	0.3	0.8	2	1.5	1.2
Test 7	Same as Test 4			0.1	1.5	2	2	1.7
Test 8	Same as Test 5			0.1	1.5	2	2	1.7
Test 9	Same as Test 6			0.1	1.5	2	2	1.7
Test 10	Same as Test 4			0.1	2	1	2.5	2
Test 11	Same as Test 5			0.1	2	1	2.5	2
Test 12	Same as Test 6			0.1	2	1	2.5	2

Table 5: Setup of the Experiments

F Convergence Test

We generate 5 batches of $N = 5000$ scenarios and solve the problem by SAA-GDR with $K = 5000$ for Test 1 to 3. Five sets of scheduled start times are obtained and their pairwise Euclidean distances are calculated and presented in Tables 6 to 8. In all tests, the pairwise distances are not sufficiently small to declare the convergence of solutions. To rule out the possibility of insufficient number of scenarios and random restarts, we conduct additional tests by taking $N = 20000$ and $K = 20000$, and observe similar pairwise distances in all test instances.

Similar to Linderoth et al.'s conjecture, it is likely that in our problem, there is a "feasible neighborhood of the solution set with which the objective value is not much different from the optimal value" J^* (Linderoth et al., 2006).

0.000	0.147	0.098	0.051	0.085
0.147	0.000	0.108	0.124	0.161
0.098	0.108	0.000	0.076	0.092
0.051	0.124	0.076	0.000	0.065
0.085	0.161	0.092	0.065	0.000

Table 6: Distance Matrix of 5 solutions in Test 1 with $N = K = 5000$

0.000	2.056	1.203	2.714	2.164
2.056	0.000	1.339	2.079	0.527
1.203	1.339	0.000	2.355	1.342
2.714	2.079	2.355	0.000	1.855
2.164	0.527	1.342	1.855	0.000

Table 7: Distance Matrix of 5 solutions in Test 2 with $N = K = 5000$

0.000	0.661	0.893	0.878	1.022
0.661	0.000	0.484	0.552	0.679
0.893	0.484	0.000	0.410	0.562
0.878	0.552	0.410	0.000	0.787
1.022	0.679	0.562	0.787	0.000

Table 8: Distance Matrix of 5 solutions in Test 3 with $N = K = 5000$

G Comparison of Results

Method	Test	$\overline{J}$	Time (s)	Test	$\overline{J}$	Time (s)	Test	$\overline{J}$	Time (s)
Tindex-SAA	4	33.5	2356.9	5	15.5	16.5	6	22.1	446.2
MV		33.3	41.0		14.8	0.5		22.0	4.5
NoPACU		38.1	14.2		15.4	4.1		21.7	15.2
SA		33.1	1105.6		15.4	589.5		23.8	1278.6
SAA-GDR		31.3	24.6		14.4	8.5		20.1	45.6
Tindex-SAA	7	43.7	2524.4	8	19.5	13.5	9	26.8	426.4
MV		45.5	45.1		19.5	0.5		29.3	7.8
NoPACU		44.6	5.7		18.3	1.5		25.3	6.4
SA		43.8	1094.4		18.4	577.2		27.3	1262.9
SAA-GDR		41.0	19.5		17.8	7.5		24.8	40.6
Tindex-SAA	10	47.2	2982.1	11	22.9	32.5	12	32.1	806.8
MV		46.2	32.7		21.7	0.6		31.5	6.1
NoPACU		62.7	14.5		23.9	3.4		34.2	13.8
SA		47.2	1104.8		22.4	584.6		31.7	1271.8
SAA-GDR		42.1	24.7		21.2	8.6		28.2	48.3

Table 9: Comparison with Other Methods

H Time-index SAA Model

In this section, a two-stage time-index sample-average-approximation binary integer programming model (TIndex-SAA) is presented.

Indices, Sets and Parameters

orr: Number of ORs.

j: Index of ORs.

J: Index set of ORs, $J = \{0, 1, \ldots, orr - 1\}$

SR_j: Number of surgeries in OR j.

k: Index of patients

K_j: Index set of patients in OR j, $K_j = \{0, 1, \ldots, SR_j - 1\}$.

N: Total number of scenarios

w: Index of scenarios of the random surgery durations and LOS in the PACU.

W: Index set of all scenarios, $W = \{0, 1, \ldots, N - 1\}$

HT: Total number of basic time intervals under study and corresponds to the large time period within which all medical activities can be finished in all scenarios.

t: Index of time intervals

T: Index set of time intervals. $T = \{0, 1, \ldots, HT - 1\}$

d^w_{jk}: Surgery duration of patient k in OR j in scenario w. $j \in J, k \in K_j, w \in W$

p^w_{jk}: Length of PACU stay for patient k in OR j in scenario w. $j \in J, k \in K_j, w \in W$

MT: Length of regular work time

C_{PW}: Patient waiting cost per time unit

C_I: surgeon idle cost per time unit

C_B: OR blocking cost per time unit

C_O: OR overtime cost per time unit

C_{PO}: PACU overtime cost per time unit

$pcap$: Total number of available spots in the PACU

M: big M

Variables

In this time-indexed model, event times are represented by arrays of binary variables and the length of each array is equal to HT. Every array is non-increasing element-wise and the sum over its elements is the time of an event's occurrence. For example, if the scheduled start time

(SST) of patient k in OR j is at the beginning of the 7th time interval, then $\overline{SST_{jkt}} = 1, \forall t \leq 6$ and $\overline{SST_{jkt}} = 0, \forall t > 6$.

$\overline{SST_{jkt}}$: binary variable, $\overline{SST_{jkt}} = 1$ if patient k in OR j is scheduled later than time t, 0 otherwise. $j \in J, k \in K_j, t \in T$

$\overline{AST^w_{jkt}}$: binary variable, $\overline{AST^w_{jkt}} = 1$ if the actual surgery start time (AST) of patient k in OR j is later than t in scenario w, 0 otherwise. $j \in J, k \in K_j, t \in T, w \in W$

ORR^w_{jkt}: binary variable, $ORR^w_{jkt} = 1$ if the surgery finish time of patient k in OR j in scenario w is later than t, 0 otherwise. $j \in J, k \in K_j, t \in T, w \in W$

$\overline{APT^w_{jkt}}$: binary variable, $\overline{APT^w_{jkt}} = 1$ if the time when patient k in OR j is admitted into the PACU in scenario w is later than t, 0 otherwise. $j \in J, k \in K_j, t \in T, w \in W$

$PACU^w_{jkt}$: binary variable, $PACU^w_{jkt} = 1$ if the time when patient k in OR j is discharged from the PACU in scenario w is later than t, 0 otherwise. $j \in J, k \in K_j, t \in T, w \in W$

z^w_{jkmn}: binary variable for implementing FCFS in "big M" method, $j, m \in J, k \in K_j, n \in K_m, w \in W : m > j$

PO^w_t: binary variable, $PO^w_t = 1$ if PACU is closed later than t in scenario w, 0 otherwise. $t \in T, w \in W$

O^w_{jt}: binary variable, $O^w_{jt} = 1$ if OR j is closed later than t in scenario w, 0 otherwise. $j \in J, t \in T, w \in W$

Model

$$\min \frac{1}{N} \Bigg[C_{PW} \sum_{j \in J} \sum_{k \in K_j} \sum_{w \in W} \sum_{t \in T} (\overline{AST^w_{jkt}} - \overline{SST_{jkt}}) \tag{H.0.1}$$

$$+ C_I \sum_{j \in J} \sum_{\substack{k \in K_j, \\ k \geq 1}} \sum_{w \in W} \sum_{t \in T} (\overline{AST^w_{jkt}} - ORR^w_{j(k-1)t}) \tag{H.0.2}$$

$$+ C_B \sum_{j \in J} \sum_{k \in K_j} \sum_{w \in W} \sum_{t \in T} (\overline{APT^w_{jkt}} - ORR^w_{jkt}) \tag{H.0.3}$$

$$+ C_O \sum_{j \in J} \sum_{w \in W} \sum_{\substack{t \in T, \\ t \geq MT}} O^w_{jt} + C_{PO} \sum_{w \in W} \sum_{\substack{t \in T, \\ t \geq MT}} PO^w_t \Bigg] \tag{H.0.4}$$

subject to

$$\overline{SST_{j00}} = 0 \tag{H.0.5}$$

$$\forall j \in J$$

$$\overline{SST_{jkt}} \geq \overline{SST_{j(k-1)t}} \tag{H.0.6}$$

$$\forall j \in J, k \in K_j, t \in T : k > 0$$

$$\overline{SST_{jkt}} \leq \overline{SST_{jk(t-1)}} \tag{H.0.7}$$

$$\forall j \in J, k \in K_j, t \in T : t > 0$$

$$\overline{SST_{jkt}} = 0 \tag{H.0.8}$$

$$\forall j \in J, k \in K_j, t \in T : t \geq MT$$

$$\overline{AST^w_{j00}} = 0 \tag{H.0.9}$$

$$\forall j \in J, w \in W$$

$$\overline{AST^w_{jkt}} \leq \overline{AST^w_{jk(t-1)}} \tag{H.0.10}$$

$$\forall j \in J, k \in K_j, w \in W, t \in T : t > 0$$

$$\overline{AST^w_{jkt}} \geq \overline{SST_{jkt}} \tag{H.0.11}$$

$$\forall j \in J, k \in K_j, w \in W, t \in T$$

$$\overline{APT^w_{jkt}} \leq \overline{APT^w_{jk(t-1)}} \tag{H.0.12}$$

$$\forall j \in J, k \in K_j, w \in W, t \in T : t > 0$$

$$\overline{APT^w_{jkt}} \geq ORR^w_{jkt} \tag{H.0.13}$$

$$\forall j \in J, k \in K_j, w \in W, t \in T$$

$$\overline{AST^w_{jkt}} \geq \overline{APT^w_{j(k-1)t}} \tag{H.0.14}$$

$$\overline{AST^w_{jkt}} \leq \overline{SST_{jkt}} + \overline{APT^w_{j(k-1)t}} \tag{H.0.15}$$

$$\forall j \in J, k \in K_j, w \in W, t \in T : k > 0$$

$$ORR^w_{jkt} = 1 \tag{H.0.16}$$

$$\forall j \in J, k \in K_j, w \in W, t \in T : t < d^w_{jk}$$

$$ORR^w_{jkt} = \overline{AST^w_{jk(t-d^w_{jk})}} \tag{H.0.17}$$

$$\forall j \in J, k \in K_j, w \in W, t \in T : t \geq d^w_{jk}$$

$$PACU^w_{jkt} = 1 \tag{H.0.18}$$

$$\forall j \in J, k \in K_j, w \in W, t \in T : t < p^w_{jk}$$

$$PACU^w_{jkt} = \overline{APT^w_{jk(t-p^w_{jk})}} \tag{H.0.19}$$

$$\forall j \in J, k \in K_j, w \in W, t \in T : t \geq p^w_{jk}$$

$$\sum_{j \in J} \sum_{k \in K_j} (PACU^w_{jkt} - \overline{APT^w_{jkt}}) \leq pcap \tag{H.0.20}$$

$$\forall w \in W, t \in T$$

$$\overline{APT^{w}_{j(SR_j-1)t}} \leq O^{w}_{jt} \tag{H.0.21}$$

$$\forall j \in J, w \in W, t \in T : t \geq MT$$

$$PACU^{w}_{jkt} \leq PO^{w}_{t} \tag{H.0.22}$$

$$\forall j \in J, k \in K_j, w \in W, t \in T : t \geq MT$$

$$\sum_{t \in T} ORR^{w}_{jkt} - \sum_{t \in T} ORR^{w}_{mnt} \leq M * z^{w}_{jkmn} \tag{H.0.23}$$

$$\sum_{t \in T} ORR^{w}_{mnt} - \sum_{t \in T} ORR^{w}_{jkt} + 1 \leq M * (1 - z^{w}_{jkmn}) \tag{H.0.24}$$

$$\sum_{t \in T} \overline{APT^{w}_{jkt}} - \sum_{t \in T} \overline{APT^{w}_{mnt}} \geq M * (z^{w}_{jkmn} - 1) \tag{H.0.25}$$

$$\sum_{t \in T} \overline{APT^{w}_{jkt}} - \sum_{t \in T} \overline{APT^{w}_{mnt}} \leq M * z^{w}_{jkmn} \tag{H.0.26}$$

$$\forall j, m \in J, k \in K_j, n \in K_m, w \in W : m > j$$

$$\overline{APT^{w}_{jkt}} \leq ORR^{w}_{jkt} + \frac{1}{pcap} \sum_{j \in J} \sum_{k \in K_j} (PACU^{w}_{jkt} - \overline{APT^{w}_{jkt}}) \tag{H.0.27}$$

$$\forall j \in J, k \in K_j, w \in W, t \in T$$

Explanation:

H.0.1 to H.0.4 The objective function is the weighted cost of patient waiting time, surgeon idle time, OR blocking time and OR and PACU overtime.

H.0.5 to H.0.8 These constraints define the SST for surgeries. As mentioned, SST is represented by a series of binary variables with non-increasing values in Constraint (H.0.7). The first surgery in each OR is scheduled at the beginning of the day in Constraint (H.0.5). No surgery can be scheduled after regular hours as required in Constraint (H.0.8). Constraint (H.0.6) enforces that SST of surgeries in an OR should follow the predefined sequence.

H.0.9 to H.0.11 Similar to SST, the AST of a surgery is represented by a series of binary variables with non-increasing values as seen in Constraint (H.0.10). In Constraint (H.0.9), the first patient in every OR is started at time 0 according to the first FCFS rule. AST cannot be earlier than the SST of a patient as described in Constraint (H.0.11).

H.0.12 to H.0.13 Each PACU admission time is a non-increasing series of binary variables in Constraint (H.0.12). Constraint (H.0.13) makes sure that no patient can be transferred to the PACU before his/her surgery is finished.

H.0.14 to H.0.15 Constraint (H.0.14) restricts each patient to be operated on only after the previous patient has been moved to the PACU. Grouped with Constraint (H.0.11) and (H.0.14), Constraint (H.0.15) can enforce that a surgery is started as soon as the patient, the surgeon and the OR are available and no intentional delay is allowed. More specifically, for patient k in OR j in scenario w, $\overline{APT^{w}_{j(k-1)t}}$ is 0 when his/her previous patient has already left OR at time t and $\overline{SST_{jkt}}$ is 0 when the current time t is no earlier than his/her SST. If the patient, the surgeon and the OR are all available, i.e. both $\overline{APT^{w}_{j(k-1)t}}$ and $\overline{SST_{jkt}}$ are 0 at t, $\overline{AST^{w}_{jkt}}$ has to be 0 at t as well, which indicates that surgery k has been started at time t.

H.0.16 to H.0.17 The surgery finish time should be the sum of AST and the surgery duration (d^{w}_{jk}). In the current model, the number of 1's in variables $ORR^{w}_{jkt}\,(t \in T)$ should be d^{w}_{jk} more than that in $\overline{AST^{w}_{jkt}}\,(t \in T)$. To maintain the monotonicity, these extra 1's will be in the front of the series $ORR^{w}_{jkt}\,(t \in T)$ and the remaining part of the series should be the same as $\overline{AST^{w}_{jk(t-d^{w}_{jk})}}\left(t \geq d^{w}_{jk}\right)$

H.0.18 to H.0.19 Similar to Constraint (H.0.16) and (H.0.17), the time when a patient is discharged from the PACU should be the sum of PACU admission time and LOS in the PACU (p^{w}_{jk}).

H.0.20 Constraint (H.0.20) is the PACU capacity constraint. In scenario w, $PACU^{w}_{jkt} = 1$ indicates that patient k in OR j has not been discharged from the PACU at time t and $\overline{APT^{w}_{jkt}} = 0$ indicates that he/she has entered the PACU at time t. Therefore $PACU^{w}_{jkt} - \overline{APT^{w}_{jkt}} = 1$ reflects that patient k from OR j is in the PACU at time t. Taking the sum over all surgeries in ORs, this constraint ensures that no PACU capacity violation occurs all across the time horizon.

H.0.21 OR overtime is accumulated if the last patient leaves the OR after regular work hours.

H.0.22 PACU overtime is penalized if the last patient leaves the PACU after regular work hours.

H.0.23 to H.0.26 These constraints enforce the third FCFS rule.

$$\sum_{t \in T} ORR^{w}_{jkt} > \sum_{t \in T} ORR^{w}_{mnt} \overset{\text{(H.0.23)}}{\rightarrow} z^{w}_{jkmn} = 1 \overset{\text{(H.0.25)}}{\rightarrow} \sum_{t \in T} \overline{APT^{w}_{jkt}} \geq \sum_{t \in T} \overline{APT^{w}_{mnt}}$$

$$\sum_{t \in T} ORR^{w}_{jkt} < \sum_{t \in T} ORR^{w}_{mnt} \overset{\text{(H.0.24)}}{\rightarrow} z^{w}_{jkmn} = 0 \overset{\text{(H.0.26)}}{\rightarrow} \sum_{t \in T} \overline{APT^{w}_{jkt}} \leq \sum_{t \in T} \overline{APT^{w}_{mnt}}$$

$$\sum_{t \in T} ORR^{w}_{jkt} = \sum_{t \in T} ORR^{w}_{mnt} \overset{\text{(H.0.24)}}{\rightarrow} z^{w}_{jkmn} = 0 \overset{\text{(H.0.26)}}{\rightarrow} \sum_{t \in T} \overline{APT^{w}_{jkt}} \leq \sum_{t \in T} \overline{APT^{w}_{mnt}}$$

When the surgery of patient k in OR j is finished at the same time as that of patient n in

OR m, i.e.,$\sum_{t\in T} ORR^w_{jkt} = \sum_{t\in T} ORR^w_{mnt}$, FCFS requires patient k in OR j to get into the PACU first, because he/she comes from an smaller-indexed OR ($j < m$).

H.0.27 This constraint makes sure that no patient is blocked in the OR when a bed is available in the PACU. In other words, a patient is sent to the PACU immediately when his/her surgery is finished and a PACU bed is available, which is part of the third FCFS rule. Grouped with Constraint (H.0.13), if surgery k in OR j has been finished at time t in scenario w, $ORR^w_{jkt} = 0$ and a spot is available in the PACU at time t, $\frac{1}{pcap}\sum_{j\in J,k\in K_j}(PACU^w_{jkt} - \overline{APT^w_{jkt}}) < 1$, this patient must have started his/her recovery in the PACU because $\overline{APT^w_{jkt}} = 0$.

I NOPACU Model

First we would like to define the notation used in NoPACU. In total orr ORs are under study and the OR index set is $J = \{0, 1, \ldots, orr - 1\}$. In OR $j \in J$, SR_j patients or surgeries (these two terms are interchangeable) are scheduled and the corresponding index set is $K_j = \{0, 1, \ldots, SR_j - 1\}$. y_{jk} indicates patient k in OR j, $j \in J, k \in K_j$ and Y is the set of all the patients in ORs. Scenarios are indexed by ω and the index set of all N scenarios is Ω. We use d^ω_{jk} to indicate the surgery duration of patient y_{jk} in scenario ω. The length of regular work time is MT time unit. C_{PW}, C_I and C_O are the cost per time unit of patient waiting time, surgeon idle time and OR overtime, respectively.

$\overline{SST_{jk}}$ is the scheduled start time of surgery $y_{jk} \in Y$. The actual start time of y_{jk} in scenario $\omega \in \Omega$ are represented by $\overline{AST^\omega_{jk}}$. We assume all SST and AST are integer multiples of the time unit. O^ω_j is overtime of OR j in scenario ω.

NoPACU:

$$\min \frac{1}{N}\Bigg[\sum_{j\in J}\sum_{k\in K_j}\sum_{\omega\in\Omega} C_{PW}(\overline{AST^\omega_{jk}} - \overline{SST_{jk}}) + C_O\sum_{j\in J}\sum_{\omega\in\Omega} O^\omega_j$$

$$+ C_I\sum_{j\in J}\sum_{\substack{k\in K_j,\\ k\geq 1}}\sum_{\omega\in\Omega}(\overline{AST^w_{jk}} - \overline{AST^w_{j(k-1)}} - d^w_{j(k-1)})\Bigg] \tag{I.0.1}$$

s.t.

$$\overline{SST_{j0}} = 0 \qquad \forall j \in J \tag{I.0.2}$$

$$\overline{SST_{jk}} \leq MT \qquad \forall j \in J, k \in K_j \tag{I.0.3}$$

$$\overline{AST^\omega_{jk}} \geq \overline{SST_{jk}} \qquad \forall j \in J, k \in K_j, \omega \in \Omega \tag{I.0.4}$$

$$\overline{AST_{jk}^{\omega}} \geq \overline{AST_{j(k-1)}^{\omega}} + d_{j(k-1)}^{\omega} \qquad \forall j \in J, k \in K_j, \omega \in \Omega : k \geq 1 \tag{I.0.5}$$

$$O_j^{\omega} \geq \overline{AST_{j(SR_j-1)}^{\omega}} + d_{j(SR_j-1)}^{\omega} - MT \qquad \forall j \in J, \omega \in \Omega \tag{I.0.6}$$

(I.0.1) The objective function is the expected cost of patient waiting time, OR overtime and surgeon idle time.

(I.0.2) The first surgery in each OR is scheduled at the beginning of the day.

(I.0.3) No surgery can be scheduled after regular work hours.

(I.0.4) A surgery is started no earlier than its scheduled start time (SST) .

(I.0.5) A surgery can not be started until its previous surgery (if any) is finished.

(I.0.6) OR overtime is accumulated if the last patient leaves the OR after regular work hours.