Adaptive Elective Surgery Planning Under Duration and Length-Of-Stay Uncertainty: A Robust Optimization Approach

Saba Neyshabouri*[1] and Bjorn Berg†[2]

[1]SEOR Department, Engr Bldg, Rm 2100, MS 4A6, George Mason University, Fairfax, VA 22030, USA

[2]Mayo Clinic, 200 First St. SW Rochester, MN 55905, USA

September 29, 2015

Abstract

Scheduling elective surgeries is a complicated task due to the coupled effect of multiple sources of uncertainty and the impact of the proposed schedule on the downstream units. In this paper, we propose an adaptive robust optimization model to address the existing uncertainty in surgery duration and length-of-stay in the surgical intensive care unit. The framework allows for the decision-maker to adjust the level of risk she/he is comfortable. We propose the adapted column-and-constraint generation method to obtain exact solutions for the proposed formulation. The performance and the quality of our proposed solution methodology are tested through computational experiments and simulation modeling.

1 Introduction

An important aspect of high quality health care delivery in a surgery department, is the assignment of appropriate post-operative care which is usually provided by specialized units such as Post-Anesthesia Care Unit (PACU), Intensive Care Unit (ICU), or Surgical Intensive Care Unit (SICU). To show the importance of these downstream resources, Jonnalagadda et al. (2005) show that 15% of the total surgery cancellation is due to the lack of an available recovery bed in the hospital they studied. Similarly, Sobolev et al. (2005) show that the Length-Of-Stay (LOS) in the ICU and the bed availability in the ICU affect the surgery schedule. This is mainly due to the existence of uncertainty in patients' LOS in such units and limited capacity of aforementioned units. In the case of lack of available capacity for patients in such units, two different policies can be employed: (1) **Cancellation** of an already planned surgery which leads to cancellation costs as well as patient discomfort, (2) **premature discharge** or **transfer** of a patient from one of these care units in order to free a bed for scheduled surgeries.

*sneyshab@gmu.edu (Corresponding author)
†berg.bjorn@mayo.edu

Each policy affects the system differently. Surgery cancellations result in prolonged stays, delayed preoperative treatments and repeated preoperative tests and treatments (Gul et al., 2012). Cancellations have been found to incur a cost of $1700-$2000 per case (Argo et al., 2009). In fact, surgical suites' operations consume around 10% of hospital budgets and deferrable surgery procedures may account for up to 52% of all hospital admissions (Gupta, 2007). Efficient management of elective procedures can result in large improvements in the overall performance of hospitals. Operating rooms (ORs) are one of the most expensive resources in hospitals and require highly skilled staff and expensive resources. Up to 70% of all hospital admissions involve a stay in the OR department (van Oostrum et al., 2008). In terms of costs, surgeries account for more than 40% of a hospital's total revenues and expenses (Erdogan et al., 2011). Considering downstream resources, Utzolino et al. (2010) show that the readmission rate to the SICU for patients with unplanned discharge from the SICU was 25.1%, which is almost 4 times that of those whom where discharged electively (8.3%). They also show that the mortality rate for patients who are readmitted to the SICU (13.3%) is almost 6 times higher than those who are not readmitted (2.28%). These statistics present the existing effects of not considering the implications of downstream capacity in surgery scheduling. However, it is important to mention that it is very difficult to quantify the costs related to mentioned risks.

This paper focuses on the decision process of assigning elective surgery patients to available surgery blocks under the *block scheduling* policy. We assume that emergency patients have a specialized unit allocated to them and do not include them in this study, as this is common in practice (Ferrand et al., 2014).

Under the *block scheduling* framework, operating room schedules are divided into multiple blocks of defined lengths and each block is assigned to a surgical team or a specific specialty (e.g., Cardiology, ENT, Neurology, etc.). Each specialty is allowed to schedule surgeries in their allocated block. Usually, surgery blocks are planned to be cyclical on a weekly or biweekly basis.

We consider OR planning for elective patients with limited downstream capacity in the SICU. We assume that the operating blocks and their cycles have been previously determined. Surgery durations and SICU LOS are modeled as uncertain parameters while the volume of surgery patients and required surgeries are known with certainty. The goal is to minimize the cost of patient admissions (based on waiting times, priority, and surgery type), overtime costs for surgery blocks, and the cost of not having enough capacity in the downstream unit. This model creates surgery plans that incorporate the uncertainty in the surgery durations and patient SICU LOS while assuring that patients receive the required care in the downstream units with high probability.

The contributions of this paper can be summarized in four parts: (1) We propose a formulation to model multi-stage care facilities that explicitly addresses intraoperative and postoperative stages in surgery planning, while it implicitly addressing the preoperative stage by trying to minimize cancellations. Therefore, we view this study as a step towards *perioperative* planning. (2) Our formulation addresses the discrete and non-convex uncertainty in surgery duration and LOS using a novel approach. To the best of our knowledge, no other studies have addressed and modeled uncertainty in LOS in such a way. (3) In our robust formulation, the definition of uncertainty is dependent on the first-stage decisions, which make the traditional solution methods invalid. We propose a new approach and adapt existing solution methodologies such that more general cases in robust optimization can be handled. (4) Based on our computational results, we present guidelines for practitioners to be used as general rules when the use of optimization techniques are not accessible.

The rest of the paper is organized as follows. Section 2 presents the relevant existing literature on the problem we are studying as well as the brief literature of the robust optimization methodology that is employed. In Section 3, we present the definitions and general formulation for this problem. In Section 4 a detailed description the structural properties and the resulting mathematical formulations are presented. The solution methodology is discussed in Section 5. Section 6 presents the results of our numerical experiments. Finally, Section 7 presents the concluding remarks and future directions.

2 Literature Review

In this section we provide a review of OR planning and scheduling research that is related to multistage care, as well as robust optimization.

2.1 Operating Room Planning

The problem of OR operations has been well studied in the literature in different categories such as block scheduling, capacity planning, and surgery sequencing to name a few. Readers are encouraged to refer to Cardoen et al. (2010), Gupta (2007), Ferrand et al. (2014), Guerriero and Guido (2011), Demeulemeester et al. (2013) for in-depth reviews of the literature related to multiple problems addressed in previous research efforts.

Thanks to the existing extensive and recent surveys on this topic, we turn our focus to research that is closely related to scheduling ORs and downstream units. Hsu et al. (2003) sets to minimize the number of nurses in the PACU (Post Anesthesia Care Unit) by determining the surgery sequences in a single day setting. They formulate the problem as a deterministic no-wait, two-stage process shop scheduling problem and solve it using a tabu search-based algorithm. Although the proposed algorithm is shown to be effective in finding near optimal solutions, uncertainty is not addressed in their setting. As shown by Marcon and Dexter (2006) through a discrete event simulation, different surgery sequencing policies have significant impact on the congestion and resource requirements in the PACU. Gupta (2007) employs a dynamic programming formulation for the elective surgery booking problem. While downstream resources are considered in this model, the multi-period nature of the demand for downstream resources has not been addressed. In fact the author states that "a tractable model of the surgery booking control problem is difficult to formulate because, following surgery, patients may require care for several days in a downstream unit and the lengths of stay are not known with certainty. Thus, each booked surgery consumes an unknown and discrete chunk of the downstream unit's resources."

In reality, the LOS for patients in the SICU can be longer than one day and is not deterministic. Truong et al. (2013) use dynamic programming to obtain optimal policies while considering the multi-period demand for downstream units. The decision is to identify the number of elective patients to serve. The authors show that localized decision rules with the focus on a single unit in the hospital can result in up to 60 % higher costs. While this formulation includes uncertainty in LOS and emergency arrivals, the model does not explicitly consider the individual capacity requirements for each surgery block in the surgery stage.

There are multiple sources of uncertainty in surgery planning and decision making under uncertainty is a much more difficult task. For example, uncertain surgery duration, which is case-dependent, can be a cause for overtime. Emergency arrivals to the ORs can lead into disruption of the planned schedule. Finally, uncertain LOS in the SICU can cause cancellation in surgeries and early discharges due to a lack available SICU beds.

Stochastic programming (SP) and robust optimization (RO) techniques have been used to address the uncertainty in surgery durations. Denton et al. (2010) provide a two-stage stochastic program as well as a robust optimization model to obtain the optimal assignment of surgery blocks to operating rooms. They show that the value of the stochastic solution is the highest when the overtime costs are high. They also show that their robust formulation provides high quality solutions very fast. Deng et al. (2014) propose a chance-constrained programming and a distributionally robust model to the surgery planning problem under uncertain surgery durations. Their formulation finds which operating rooms to open, as well as the assignment and sequencing of surgeries to the ORs. Gul et al. (2012) propose a multi-stage mixed-integer programming approach to assign surgeries to operating rooms over a finite horizon. They assume surgery durations as well as patient demand for surgery are random and aim to minimize the expected cost of cancellations, postponements, and OR overtime costs. They propose a progressive hedging algorithm as well as a heuristic to obtain solutions. They draw managerial insights regarding the existing trade-offs between postponing and canceling surgeries.

Addis et al. (2014) propose a robust optimization approach for assigning patients to surgery blocks in a block scheduling setting. They assume uncertain durations for surgeries and propose a static robust formulation to minimize a function that penalizes associated waiting time, urgency, and tardiness of patients. Although this proposes a robust formulation for the surgery planning problem, there is no focus on the downstream resources and their effect on the schedule. In addition, the formulation is static and does not include any recourse decisions. Shylo et al. (2012) propose a model for batch scheduling surgeries with uncertain duration to surgery blocks. The goal is to maximize the expected utilization of operating rooms subject to a set of probabilistic capacity constraints. The authors show that their proposed method produces significantly better schedules in terms of performance as compared to simple heuristic scheduling rules. While addressing the high-volume batch scheduling, the proposed methodology does not address the downstream effects of the proposed schedule.

Min and Yih (2010) propose a two-stage stochastic programming approach to model the elective surgery planning problem. They consider uncertainty in both surgery duration and the LOS in the SICU. They employ a sample average approximation to obtain optimal solutions to this problem. While this study addresses the uncertainty in both surgery duration and LOS, the two-stage formulation is risk-neutral.

Fügener et al. (2014) consider the master surgery scheduling (MSS) problem and its effects on the downstream units. They propose an analytical approach to calculate the exact distribution for downstream resources for a given MSS. Next they define multiple cost measures resulting from the MSS and methods to minimize these costs. They rely on the existence of empirical data for every specialty such as admission probability, LOS probability, etc., to characterize the probability distributions. Obtaining an accurate estimate for these probabilities is not always possible and requires a large amount of data.

As can be seen, the surgery planning problem with downstream resource capacity considerations under uncertainty has received relatively less attention while it is known that not having a holistic planning approach and focusing on isolated units increases the chance of suboptimal or globally infeasibe solutions (Fügener et al., 2014).

In our study, we focus on elective surgery patients with uncertain surgery duration and LOS in the

SICU. However, we model these uncertainties within the framework of a robust optimization model. The main reason for employing this approach is that in many cases, obtaining and characterizing a probability distribution can be very difficult. On the other hand, the use of distributional information may not be appropriate for construction of scenarios for short-term planning. In other words, the surgery duration and LOS distribution for a small group of heart patients on the waiting list for an upcoming week may be drastically different than the general distribution obtained from the data collected over the years.

We believe that a data-driven approach for tactical surgery planning is more suitable since it can consider patient and case specific information, which can be obtained from the physicians or subject natter experts. In addition, existence of distributions will not necessarily ensure tractable solution methods. The first difficulty is producing the number of scenarios that can be a representative sample of the multi-dimensional uncertainty which can be prohibitively large. Next, the two-stage stochastic programming approach requires the evaluation of the second-stage for each scenario, which if coupled with large number of scenarios can cause tractability issues. Further, the risk-averse nature of the RO methodology makes it an appropriate framework since it inherently caters towards the patient safety and focus on high quality care.

2.2 Robust Optimization

Robust optimization (RO) is a relatively new approach for decision making under uncertainty which is introduced by El Ghaoui et al. (1998), Ben-Tal and Nemirovski (1998), and Bertsimas and Sim (2004). Unlike the stochastic programming theory which assumes distributional information about uncertain parameters and aims to optimize an expectation measure (Birge and Louveaux, 2011), RO does not need distributional information and produces optimal solutions that are feasible for a defined set of values that uncertain parameters can take. In other words, RO seeks to optimize against the worst-case realization of uncertainty. It can also produce a probabilistic guarantee for the feasibility of the solutions. The conservatism of the solution can be controlled by the means of a defined *budget of uncertainty*. In cases where obtaining probabilistic information is not possible or infeasibility cannot be tolerated, RO offers a flexible framework for producing good solutions. Bertsimas et al. (2011) and Gabrel et al. (2014) provide surveys of the existing literature on the theory and applications of the robust optimization.

RO has been successfully applied to various applications where modeling uncertainty using traditional stochastic methods are not applicable. The use of distributional information for a short-term planning of the surgery department may not provide correct solutions, while better results can be obtained considering a data-driven approach that requires specific information for each patient. Consider the example of scheduling surgeries during a week in which all of the patients needing a heart surgery have very complicated surgeries, with high surgery durations and LOSs in the SICU coming from the tail of their distribution. Stochastic programming methods or sampling methods cannot accommodate such special cases. RO on the other hand, can include case-specific information and facilitate such short-term planning complexities. Additionally, RO methodology provides means to obtain solutions that are protected against uncertainty. This characteristic is of particular interest in healthcare settings like surgery planning since the outcome of not considering these sources of uncertainty can have a serious impact on the health outcome of the patients. RO provides decision-makers with high quality solutions that while minmizing the operating costs, guarantee a high level of safety and quality of care for patients.

Similar to two-stage stochastic programs (Birge and Louveaux, 2011), two-stage (adaptive) robust optimization addresses cases when decisions can be split into two different sets. First, there are the *here-and-now* decisions that have to be made prior to the realization of the uncertainty. After the uncertainty is realized, *wait-and-see* decisions are made. These are the recourse decisions that are made to correct for the impact of the uncertainty. In contrast to stochastic programming methods where the realization of uncertainty is a product of a stochastic process (e.g., sample from a distribution or simulation), the realization of uncertainty in the RO methodology is a product of an optimization model.

Ben-Tal et al. (2004) state that in cases where recourse actions are available, two-stage robust optimization problems produce better solutions than the static version. They show most cases of two-stage RO problems are NP-hard and propose to restrict recourse solutions to be affine functions of uncertain data and then show cases where this restriction produces tractable optimization problems. They apply the proposed approximation to an inventory problem and solve the resulting reformulation as a linear program. Later in (Atamtürk and Zhang, 2007) a two-stage robust optimization approach for a network design problem where demand is uncertain is proposed. Theoretical complexity results and special cases are also presented. The methodology is applied to various problems such as lot-sizing and location-transportation. Thiele et al. (2009) propose a general framework for two-stage robust LP problems. To solve these problems, an algorithm based on Kelley's cutting-plane algorithm (Kelley, 1960) is proposed. Gabrel et al. (2011) formulate the location-transportation problem with demand uncertainty using a two-stage robust formulation. A cutting-plane algorithm is used to obtain solutions. Zeng and Zhao (2013) propose a column-and-constraint generation method to solve two-stage robust problems and show that the solution methodology performs better than the cutting-plane methods proposed by Gabrel et al. (2011) and Thiele et al. (2009) when applied to location-transportation problems. In addition, they provide theoretical results that show their proposed algorithm has a superior worst-case complexity as compared with one proposed by Thiele et al. (2009). For more information regarding adaptability in robust optimization and its approximations, readers are referred to Bertsimas et al. (2011) and references there in.

There is not much literature on two-stage robust optimization problems with mixed-integer recourse structure. Zhao and Zeng (2012) propose a formulation, conditions, and a solution method based on their column-and-constraint generation method (Zeng and Zhao, 2013) to obtain provably optimal solutions to this class of problems. Approximation methods based on finite adaptability for two-stage and multi-stage RO problems are proposed in Bertsimas and Caramanis (2010) (see also Bertsimas et al. (2011) and references therein). Although the recourse decisions in the SICU are in the form of number of patients and bounded to be integer, thanks to our novel formulation the integrality constraints can be relaxed.

In our study, we base our theory on the work of Bertsimas and Sim (2004), Thiele et al. (2009), Gabrel et al. (2011), and Zeng and Zhao (2013). We propose a novel two-stage robust formulation to model surgery planning with downstream resource capacity. Both surgery durations and LOS in the SICU are uncertain parameters. Our formulation naturally decomposes the problem into sub-problems for ORs and downstream units. We propose a column-and-constraint generation algorithm to solve this problem. The formulation can be used in other important applications such as project management, manufacturing, etc., where downstream resources are shared by multiple entities.

3 Model Development

In this section, the model development for planning of surgery and downstream capacity under uncertainty is presented. First, for better understanding, some notation and a deterministic formulation is presented. Next, we define our model of uncertainty in detail. Finally, the robust formulation to model this problem is presented. We go in depth to study the structure of the proposed formulation and its characteristics in the subsequent section.

3.1 Definitions and Deterministic Formulation

We define set B as the set of surgery blocks in a given decision period (usually a week or two weeks long). Set S defines the set of specialties that are included in the cyclic schedule. Each block, $b \in B$, is dedicated to only one type of specialty while there can be multiple blocks of the same specialty during a cycle in the surgery schedule. B_s is used to denote the set of blocks for specialty s during the planning horizon. Set $I = \{1, ..., n\}$ represents the elective surgery patients. Set $I_s \subseteq I$ represents the set of patients that require specialty s. Note that each patient is assigned to one specialty. Therefore patient i of specialty s can be assigned to any of the blocks $b \in B_s$ during the planning horizon. We assume, without loss of generality, that the planning horizon is T days and is an integer multiple of the surgery schedule cycle length. Each surgery block b has a pre-allocated length of time which is denoted by h_b on a specific day t_b. Due to uncertainty in surgery times, a surgery block may need to use extra time to finish the scheduled surgeries. Therefore overtime cost c_s for each unit of time is incurred for specialty s. For ease of modeling, a dummy block ,$b' \in B$, is considered for patients who are not assigned to any surgery blocks during the planning horizon T and are postponed be assigned to surgery during the next planning horizon. The cost of assigning patient i to each block is defined as a_{ib}, where $a_{ib} < a_{ib'}\ \forall b \neq b'$. This represents the admission costs for each patient considering their waiting times and priorities.

Associated with each patient i, there is the length-of-stay l_i which denotes the number of consecutive days that the patient is required to stay in the SICU. In addition, there is the surgery duration d_i, which represents the time required to perform the surgery for patient i. The SICU has limited number of beds on each day, represented by r_t. In the case of lack of capacity in the SICU, a patient will be denied admission to the SICU and/or has to be transferred to units with lower level of care. The cost incurred for each day a patient is not receiving the care in the SICU is denoted by e_t.

To model the problem, we let x_{ib}^s be equal to one if patient i with specialty s is assigned to perform surgery on block b, and is set to zero otherwise. y_{it} is one if patient i is in need of a SICU bed on day t, and is zero otherwise. Continuous decision variable o_b captures the amount of overtime incurred during the surgery in block b. u_t counts the number of patients on day t that are in need of a SICU bed, but are denied admission due to the lack of capacity.

Here we provide the Deterministic Operating Room Planning with Downstream Capacity (DORP-DC) before introducing its robust counterpart.

$$\min \quad \sum_{s\in S}\sum_{i=1}^{n}\sum_{b\in B} a_{ib}x_{ib}^{s} + \sum_{s\in S}\sum_{b\in B_s\setminus\{b'\}} c_s o_b + \sum_{t=1}^{T} e_t u_t \tag{1a}$$

s.t.

$$\sum_{b\in B_s\cup\{b'\}} x_{ib}^{s} = 1 \qquad i\in I_s, s\in S \tag{1b}$$

$$\sum_{i\in I_s} d_i x_{ib}^{s} \leq h_b + o_b \qquad b\in B_s\setminus\{b'\}, s\in S \tag{1c}$$

$$y_{it} \geq x_{ib}^{s} \qquad s\in S, i\in I_s, b\in B_s, t=t_b,..,t_b+l_i-1 \tag{1d}$$

$$\sum_{i=1}^{n} y_{it} \leq r_t + u_t \qquad \forall t \tag{1e}$$

$$y_{it}, x_{ib}^{s} \in \{0,1\}, o_b \geq 0, u_t \geq 0 \qquad \forall s,i,b,t \tag{1f}$$

In DORP-DC, the objective (1a) is to minimize a measure of total costs. The first term on the left is the sum of the cost of assigning patients to surgery blocks (which includes patient priority and waiting time). The second term calculates the over time cost for performing surgeries. The third term measures the total cost of lack of SICU capacity which causes premature discharges or transfers. The first constraints, (1b), enforce the assignment of patients to blocks, requiring each patient to be assigned exactly once (including the dummy surgery block) to a block within the required specialty. The second constraints, (1c), calculate the value of overtime for each surgery block based on the assignments and surgery duration. Constraints (1d) are defined to indicate if a patient is in the SICU on any given day based on their assignment and LOS. This constraint ensures that patient i stays in the SICU for l_i consecutive days upon performing the surgery on day t_b, which is the day of surgery for block b. Constraints (1e) enforces the SICU capacity limitation and calculates denied admissions to the SICU (or transfers out of the SICU) for each day. Note that the value for the variable u_t is integer since both y_{it} and r_t are integer valued. Therefore we can drop the constraint that forces variable u to be integer. The final constraints, (1f), define the domain for the decision variables.

3.2 Robust Model

In our deterministic formulation, it is assumed that all the parameters of the problem are known with certainty. However in reality, it is very difficult, if not impossible, to predict the values for surgery duration and LOS in the SICU. Therefore, these parameters are assumed to be uncertain while belonging to a known set.

Considering the uncertainty and the decision process, this problem can be viewed as a two-stage process in which decisions to assign patients to surgery blocks (x_{ib}^{s}) are made in the first stage. Next, uncertainty in the surgery duration and LOS for each assigned patient is realized. In the second stage, the goal is to minimize the defined worst-case scenario for the overtime and denied SICU admission costs. This is an adaptive process which tries to employ the best recourse decision after the realization of the uncertainty. Considering the worst-case realization of uncertainty can be a suitable approach since the risk associated with not satisfying the SICU bed requirements can have very adverse effects on patients safety and health.

We assume that only a subset of uncertain parameters will deviate from their nominal value and try to minimize the the worst-case costs. Let us define $\tilde{d}_i \in [\overline{d}_i, \overline{d}_i + \hat{d}_i]$, in which $\overline{d}_i$ is the nominal value for surgery duration for patient i, while $\hat{d}_i$ is the total deviation from the nominal value that the duration can have. Without loss of generality, for LOS, we define $\tilde{l}_i \in \{\bar{l}_i, \ldots, \bar{l}_i + \hat{l}_i\}$ to represent the uncertainty set for LOS in the SICU for patient i. For simplicity, we further assume that the LOS for each patient is integer-valued corresponding to the number of days, however to model a finer granularity in time, shorter time windows can be considered (e.g. hours, shifts). Note that the assumption for integer-valued LOS is very close to reality as providers typically round once per day. In the case of surgery durations, $z_i = \dfrac{\tilde{d}_i - \overline{d}_i}{\hat{d}_i}$ is the normalized deviation from the nominal surgery duration for patient i and $0 \leq z_i \leq 1$. Following the notation defined by (Bertsimas and Sim, 2004), we define $\Gamma_d = (\Gamma_d^1, \ldots, \Gamma_d^{|S|})$ as the *budget of uncertainty* vector for surgery durations within each specialty. Then we enforce $\sum_{i \in I_s} z_i \leq \Gamma_d^s, \forall s$, which limits the total possible normalized deviation from the nominal value being less than the budget of uncertainty Γ_d^s. In other words, if Γ_d^s is integer-valued, only Γ_d^s of patients of specialty s can have surgery durations equal to their highest possible duration. In a simpler case, we can define Γ_d to be a single parameter that limits the deviations over all the specialties.

Different surgery types and specialties have different levels of uncertainty associated to them. For example, the possibility of deviation from the nominal time in a standard joint replacement surgery is expected to be far less than an open-heart surgery due to the inherent uncertainties and possible complicating factors. Therefore, based on the preference of management, different uncertainty budgets can be allocated to different specialties. In case the decision-maker has no preference or information on specialty-related risks, the formulation can be easily adapted by defining a single inequality for the budget of uncertainty rather than $|S|$ inequalities. The same ideas and assumptions can be applied to define the uncertainty set for the LOS in the SICU.

Following the same assumption made for surgery durations, we define $\dfrac{\tilde{l}_i - \bar{l}_i}{\hat{l}_i}, \forall i$ and enforce the budget of uncertainty $\Gamma_l = (\Gamma_l^1, \ldots, \Gamma_l^{|S|})$ for LOS in the SICU as $\sum_{i \in I_s} \dfrac{\tilde{l}_i - \bar{l}_i}{\hat{l}_i} \leq \Gamma_l^s, \forall s$. We also assume that the realization of uncertainty in surgery duration is independent from the realization in the LOS.

We define the uncertainty sets for surgery duration and LOS as follows:

$$\mathcal{U}_d = \{d \in R^n : \tilde{d}_i = \overline{d}_i + z_i \hat{d}_i, 0 \leq z_i \leq 1 \ \ \forall i, \sum_{i \in I_s} z_i \leq \Gamma_d^s \ \ \forall s\} \tag{2}$$

$$\mathcal{U}_l = \{l \in R^n : \tilde{l}_i \in \{\bar{l}_i, \ldots, \bar{l}_i + \hat{l}_i\} \ \ \forall i, \sum_{i \in I_s} \frac{\tilde{l}_i - \bar{l}_i}{\hat{l}_i} \leq \Gamma_l^s \ \ \forall s\}. \tag{3}$$

The data for the nominal values and worst-case deviations can be obtained from subject matter experts, physicians, or managers that have detailed information about each patient's health and conditions. Depending on the risk-attitude of the decision-maker, a value for the budget of uncertainty is chosen. Higher values of the uncertainty budget show that there is high chance of deviations in uncertain parameters and the resulting schedules are more conservative. The use of such uncertainty sets in health care settings and their benefits and drawbacks are considered by Addis et al. (2015). The authors address the operating room planning problem, and the nurse-to-patient assignment problem in home care services using the proposed cardinality-constrained uncertainty.

Considering the assumptions and definitions mentioned earlier, the formulation for the Robust Adaptive Surgery Planning with Downstream Capacity (RASP-DC) problem can be written as follows:

$$\min \quad \sum_{s\in S}\sum_{i=1}^{n}\sum_{b\in B} a_{ib}x_{ib}^{s} + opt[R(x,\Gamma_d,\Gamma_l)] \tag{4a}$$

$$s.t.$$

$$\sum_{b\in B_s\cup\{b'\}} x_{ib}^{s} = 1 \qquad i\in I_s, s\in S \tag{4b}$$

$$x_{ib}^{s}\in\{0,1\} \qquad \forall s,i,b \tag{4c}$$

where $opt[R(x,\Gamma_d,\Gamma_l)]$ is the optimal solution to the recourse problem, $R[(x,\Gamma_d,\Gamma_l)]$:

$$\max_{\tilde{d}\in\mathcal{U}_d,\tilde{l}\in\mathcal{U}_l} \min \quad \sum_{s\in S}\sum_{b\in B_s\setminus\{b'\}} c_s o_b + \sum_{t=1}^{T} e_t u_t \tag{5a}$$

$$s.t.$$

$$\sum_{i\in I_s} \tilde{d}_i x_{ib}^{s} \leq h_b + o_b \qquad b\in B_s\setminus\{b'\}, s\in S \tag{5b}$$

$$y_{it} \geq x_{ib}^{s} \qquad s\in S, i\in I_s, b\in B_s, t=t_b,..,t_b+\tilde{l}_i-1 \tag{5c}$$

$$\sum_{i=1}^{n} y_{it} \leq r_t + u_t \qquad \forall t \tag{5d}$$

$$y_{it}\in\{0,1\}, o_b\geq 0, u_t\geq 0 \qquad \forall s,i,b,t \tag{5e}$$

In the first stage, prior to any knowledge of the realization of uncertainty, assignment decisions are made. Patients are assigned to surgery blocks. During the operations, the value for surgery duration will be realized and the costs of overtime for each surgery block is incurred. In addition, the LOS in the SICU is realized which determines the utilization of the SICU capacity and possible denied admissions or transfers. The uncertainty exists in the technology matrix of the first constraints, (5b), and the set of indexes of the forth set of constraints, (5c). In fact, using this formulation, if the LOS is uncertain, the number of constraints of type (5c) will be uncertain which poses serious complexity issues for solving this problem. The goal of the decision-maker is to minimize the costs associated with surgery block overtimes and denied admissions to the SICU. The goal of the second-stage problem is to minimize the worst-case recourse costs based on the definition of the uncertainty sets. It is important to note that this formulation cannot be supplied to any solver in the presented form. An extensive formulation can be obtained by enumerating all the possible realizations of uncertainty. However this is prohibitive even for problems of medium size.

Note that in recourse problem $R[(x,\Gamma_d,\Gamma_l)]$, ((5a)-(5e)), x_{ib}^{s} is not a decision variable. First-stage decision variable values $\boldsymbol{x}$ are passed to the second-stage as parameters. It is clear from the formulation of recourse problem $R[(x,\Gamma_d,\Gamma_l)]$ that the variables related to the surgery block overtime, o_b, are independent from the variables capturing the status of the SICU bed capacity, y_{it}, and denied admissions, u_t. This observation helps us to decompose the recourse problem further into two different and important problems: (1) **Surgery Block Capacity (SBC)** problem which aims to calculate the worst-case minimum overtime costs due to deviations in surgery durations, (2) **Downstream Capacity (DC)** problem which aims to calculate the worst-case minimum costs of denied admissions to the SICU due to the deviations in the LOS.

Following the observation on the separability of these problems we can reformulate the robust adaptive surgery planning with downstream capacity as follows:

$$\min \quad \sum_{s\in S}\sum_{i=1}^{n}\sum_{b\in B} a_{ib}x_{ib}^{s} + opt[R_d(x,\Gamma_d)] + opt[R_l(x,\Gamma_l)] \tag{6a}$$

$$s.t.$$

$$\sum_{b\in B_s\cup\{b'\}} x_{ib}^{s} = 1 \qquad i\in I_s, s\in S \tag{6b}$$

$$x_{ib}^{s}\in\{0,1\} \qquad \forall s,i,b \tag{6c}$$

where $opt[R(x,\Gamma_d)]$ is the optimal value of the surgery block capacity recourse problem $R_d[(x,\Gamma_d)]$:

$$\max_{\tilde{d}\in\mathcal{U}_d} \quad \min \quad \sum_{s\in S}\sum_{b\in B_s\setminus\{b'\}} c_s o_b \tag{7a}$$

$$s.t.$$

$$\sum_{i\in I_s} \tilde{d}_i x_{ib}^{s} \le h_b + o_b \qquad b\in B_s\setminus\{b'\}, s\in S \tag{7b}$$

$$o_b\ge 0 \qquad \forall b \tag{7c}$$

and $opt[R_l(x,\Gamma_l)]$ is the optimal value of the downstream capacity recourse problem $R_l[(x,\Gamma_l)]$:

$$\max_{\tilde{l}\in\mathcal{U}_l} \quad \min \quad \sum_{t=1}^{T} e_t u_t \tag{8a}$$

$$s.t.$$

$$y_{it}\ge x_{ib}^{s} \qquad s\in S, i\in I_s, b\in B_s, t=t_b,..,t_b+\tilde{l}_i-1 \tag{8b}$$

$$\sum_{i=1}^{n} y_{it}\le r_t+u_t \qquad \forall t \tag{8c}$$

$$y_{it}\in\{0,1\}, u_t\ge 0. \qquad \forall i,t \tag{8d}$$

In the following sections, we study each of these problems in depth and present insightful structural properties that provides insights to each of these problems. The properties are employed to provide a mixed-integer linear programming (MILP) reformulation for each of the sub-problems that can be solved using commercial solvers.

4 Structural Properties

In this section, we study the structural properties of surgery block capacity and downstream capacity problems. These insights help us reformulate these problems to more tractable mixed-integer linear programs (MILPs) that can be solved using commercial solvers, which in turn allows us to solve the overall surgery planning problem.

4.1 Surgery Block Capacity Problem

In this section, we focus on the surgery block capacity recourse problem defined by (7a)-(7c). This problem in itself is a two-level optimization problem in which first a maximization over the uncertainty set defines the outer-level problem and seeks the worst-case scenario for a given surgery assignment. After the realization of the surgery durations, the inner-minimization problem aims to minimize the overall overtime costs. In the inner-minimization problem, the only decision variables are those that capture the overtime for each surgery block ($o_b, \forall b$), and they are continuous. Using the definition of the uncertainty set $\mathcal{U}_d$, we can substitute for the values of $\tilde{d}_i$ and have the surgery block capacity recourse problem as follows:

$$\max_{\sum_{i \in I_s} z_i \leq \Gamma_d^s, \forall s, 0 \leq z_i \leq 1, \forall i} \quad \min \quad \sum_{s \in S} \sum_{b \in B_s \setminus \{b'\}} c_s o_b \tag{9a}$$

$$s.t.$$

$$\sum_{i \in I_s} (\bar{d}_i + \hat{d}_i z_i) x_{ib}^s \leq h_b + o_b \qquad b \in B_s \setminus \{b'\}, s \in S \tag{9b}$$

$$o_b \geq 0. \qquad \forall b \tag{9c}$$

Because that the inner-minimization is a linear program, strong duality can be used to substitute the inner-minimization problem with its dual, which can be written as a single maximization problem as follows:

$$\max \quad \sum_{s \in S} \sum_{b \in B_s} \left[\sum_{i \in I_s} (\bar{d}_i + \hat{d}_i z_i) x_{ib}^s - h_b \right] \pi_b^s \tag{10a}$$

$$s.t.$$

$$\sum_{i \in I_s} z_i \leq \Gamma_d^s \qquad s \in S \tag{10b}$$

$$0 \leq \pi_b^s \leq c_s \qquad s \in S, b \in B_s \tag{10c}$$

$$0 \leq z_i \leq 1. \qquad \forall i \tag{10d}$$

(10a)-(10d) is the reformulation for the surgery block capacity recourse problem $R_d(x, \Gamma_d)$ that transforms the $\max - \min$ objective into a single maximization problem. Note that variable π_b is the dual variable associated with the capacity constraint (9b) for surgery block b. Due to the existence of the bilinear term $z_i \pi_b^s$, the second-stage problem $R_d(x, \Gamma_d)$ is the maximization of a bilinear function over linear constraints. Bilinear programming is a special case of quadratic programming and the objective function, in general, is neither convex nor concave (Gallo and Ülkücü, 1977). This is a limiting factor in using standard convex optimization solvers to obtain optimal solutions to the second-stage problem. The following propositions, based on the structure of the recourse problem, enable us to reformulate the second-stage problem $r(x, \Gamma_r)$ as a mixed-integer linear program (MILP).

Proposition 4.1 *If the components of the budget of uncertainty* Γ_d *are integer values, there exists an optimal solution* (π^*, z^*) *to the second-stage problem* $R_d(x, \Gamma_d)$ *such that* $z_i^* \in \{0, 1\}, \forall i$.

Proof Let us define the feasible region for the second-stage problem by the following polyhedra $\Pi = \{\pi \in \mathbb{R}^m | 0 \le \pi_b^s \le c_s, \forall s, b \in B_s\}$ and $\mathcal{Z}(\Gamma_d) = \{z \in \mathbb{R}^n | \sum_{i \in I_s} z_i \le \Gamma_d^s, \forall s, 0 \le z_i \le 1, \forall i\}$. Note that both sets Π and $\mathcal{Z}(\Gamma_d)$ are bounded (all variables are bounded) and therefore an optimal solution (π^*, z^*) exists such that π^* is an extreme point of Π and z^* is an extreme point of $\mathcal{Z}(\Gamma_d)$ (Gallo and Ülkücü, 1977). This implies that when Γ_d is composed of all integer values, there exists an optimal solution such that $z^* \in \{0, 1\}^n$ (also see (Gabrel et al., 2011)). □

Proposition 4.2 *For any* $s \in S, b \in B_s$ *in the second-stage, the optimal solution* $\pi_b^{s*} \in \{0, c_s\}$.

Proof Due to the structure of the objective function one of the following cases is true for any $\pi_b^s, s \in S, b \in B_s$:

- First, consider the case that for a given block b with specialty s, assignment vector x, and deviation vector z, $\sum_{i \in I_s} (\overline{d}_i + \hat{d}_i z_i) x_{ib}^s - h_b > 0$. In this case, due to the maximization of the objective, π_b^s will assume its upperbound c_s.
- Second, consider the case that for a given block b with specialty s, assignment x, and deviation vector z, $\sum_{i \in I_s} (\overline{d}_i + \hat{d}_i z_i) x_{ib}^s - h_b \le 0$. This means that given the assignment and the deviations for resource consumption parameters, resource consumption will not exceed the available capacity h_b. In this case, due to the maximization of the objective, $\pi_b^s = 0$.□

As a result of 4.1 and 4.2 we can reformulate and solve (9a)-(9c) as a MILP. Next we consider the downstream capacity problem and its structural properties.

4.2 Downstream Capacity Problem

In this section, we turn our focus to the downstream capacity recourse problem. The formulation is presented by (8a)-(8d). Note that the maximization over the uncertain parameter $\tilde{l}$ is impossible in the current formulation since $\tilde{l}$ is in the set of indexes of the constraint (8b) and not in the inequalities. In other words, the maximization has to decide the number of constraints of type (8b) as decision variables. There is no existing method to address problems of this structure. This motivates the need for a new formulation such that we can transfer the decision variable $\tilde{l}$ into the equations. Here, we provide a new formulation for DCP by redefining our decision variables.

As explained before, we can divide the process into three different decision-making stages. In the first stage, patients are assigned to surgery blocks. Considering the uncertainty in surgery duration and LOS and the definition of the budget of uncertainty, uncertain parameters assume their value in the second stage. We can assume the second stage decisions are made by an adversary. For the third stage, we aim to minimize the cost of recourse for the previous stages. In order to formulate this problem we need to redefine our variables as $v_{it} = 1$, if patient i enters SICU *by* day t, 0 otherwise, and $w_{it} = 1$, if patient i leaves SICU *by* day t, 0 otherwise.

It is important to note that *"by"* is used in the definition of the variables rather than *"at."* Stemming directly from the definition of the variables, the following inequalities hold:

$$v_{it} \leq v_{i,t+1} \qquad \forall i, t \tag{11a}$$

$$w_{it} \leq w_{i,t+1}. \qquad \forall i, t \tag{11b}$$

The first inequality indicates that if a patient has arrived to the SICU *by* day t ($v_{it} = 1$), then he/she has arrived by the days after t. Therefore, all the variables for those days must equal 1. The second inequality indicates the same principle for leaving the SICU.

Defining the variables in this way naturally adapts to our problem setup and stages. Variables v_{it} are automatically defined after the first stage decisions are made. Next, the worst-case second-stage recourse costs are calculated by choosing the times that patients will leave the SICU, controlled by variables w_{it}. The LOS for patient i is equal to $\sum_{t=1}^{T}(v_{it} - w_{it})$, and the number of patients in the SICU on day t is equal to $\sum_{i=1}^{n}(v_{it} - w_{it})$. This requires us to redefine the uncertainty set based on the variables that represent the uncertainty as the SICU departure time. Note that this reformulation transforms the parameter for the LOS (l) (which does not depend on the arrival to the SICU) into an arrival/departure process. While using the arrival and departure times we can simply calculate the LOS, the definition of uncertainty changes to be the time that a patient is released from the SICU.

4.2.1 Alternative Representation of Uncertainty

As discussed before, each patient i has a LOS ($\tilde{l}_i$) at the SICU that belongs to the discrete set $\{\bar{l}_i, \ldots, \bar{l}_i + \hat{l}_i\}$. We assume, without the loss of generality, that the LOS is defined to be integer which means that the LOS cannot be a fraction of a day and both $\bar{l}_i$ and $\hat{l}_i$ are also integer. We can consider fractions of a day by further dividing the steps in the time-windows. Note that the definition of our variables naturally adapts to our assumptions on the LOS in the SICU. These assumptions along with the value of first-stage variables help us fix the values of a subset of variables as follows:

$$v_{it} \geq x_{ib}^{s} \qquad t = t_b, \forall s, i \in I_s, b \in B_s \tag{12a}$$

$$v_{it} \leq 1 - x_{ib}^{s} \qquad t = 1, .., t_b - 1, \forall s, i \in I_s, b \in B_s \tag{12b}$$

$$w_{it} \geq x_{ib}^{s} \qquad t = t_b + \bar{l}_i + \hat{l}_i - 1, .., T, \forall s, i \in I_s, b \in B_s \tag{12c}$$

$$w_{it} \leq 1 - x_{ib}^{s}. \qquad t = 1, .., t_b + \bar{l}_i - 1, \forall s, i \in I_s, b \in B_s \tag{12d}$$

The first inequalities, (12a), ensure that each patient goes to the SICU on the day of surgery, while the second set of inequalities, (12b), enforce that patients cannot go to the SICU before the day of surgery. The third set of inequalities, (12c), ensure that each patient can only stay in the SICU for at most $\bar{l}_i + \hat{l}_i$ days. The forth set of inequalities, (12d), ensure that patients cannot leave the SICU before the minimum LOS in the SICU, which is $\bar{l}_i$ days.

Considering that the LOS for patient i can be written as $\tilde{l}_i = \sum_{t=1}^{T}(v_{it} - w_{it})$, the mathematical representation of the budget of uncertainty constraint can be written as follows:

$$\sum_{i\in I_s,\hat{l}_i>0,x_{ib'}=0}\left[\frac{\sum_{t=1}^{T}(v_{it}-w_{it})-\bar{l}_i}{\hat{l}_i}\right]\leq \Gamma_l^s \qquad \forall s \tag{13}$$

which is defined for the patients that have uncertainty in the LOS and are assigned to have a surgery during the planning horizon.

The set that characterizes the uncertainty in the LOS in the SICU can be redefined based on their departure time from the SICU, while their arrival is an input parameter to define the uncertainty set as follows:

$$\mathcal{U}_l(x,v)=\{w\in\{0,1\}^{n\times T}:(11b),(12c)-(12d),(13)\}. \tag{14}$$

This novel definition of the variables and reformulation of the uncertainty set allows us to incorporate the random parameter in the problem as a decision variable so it can easily adapt to the robust optimization approach. To the best of our knowledge, this is the first study such that the definition of uncertainty set depends on the first-stage variables.

4.2.2 Resulting Solvable Formulation

Considering the structural properties based on the definition of our variables, the robust adaptive surgery planning with downstream capacity Problem can be formulated as a two-stage problem. In the first stage, decisions regarding the assignment of patients to surgery blocks (x_{ib}^s) are made. The by-product of this stage is the time for each patient to enter the SICU (v_{it}) which can be obtained using inequalities (11a), (12a), and (12b). Variables v_{it} represent the arrival of patients to the SICU based on the assignment of patients to the surgery blocks and the variable definitions.

Variables v_{it} naturally belong to the first stage of the problem and do not have any impact of the objective value of the first stage nor limit the feasible region for variables x_{ib}.

The downstream capacity recourse problem, $R(\boldsymbol{x},\boldsymbol{v},\Gamma_l)$, can be written as follows:

$$\max_{w\in\mathcal{U}_l(x,v)} \quad \min \quad \sum_{t=1}^{T} e_t u_t \tag{15a}$$

$s.t.$

$$\sum_{i=1}^{n} v_{it}-w_{it}\leq r_t+u_t \qquad \forall t \tag{15b}$$

$$u_t\geq 0. \qquad \forall i,t \tag{15c}$$

The objective function (15a) is the maximization of the denied admission costs over the uncertainty set which controls the the discharge date, and consequently the LOS for each patient. The first constraint (15b) calculates the number of denied admissions to the SICU based on the arrivals to the SICU (determined by v_{it}) and the SICU discharges (determined by w_{it}) for each day. The second constraints, (15c), define the range for the number of denied admissions.

Note that this problem has two levels. The first level is finding the worst-case realization of LOS for patients. Next, the decision-maker aims to minimize the costs associated with the realized LOS and required transfer costs. In this formulation, the inner-minimization problem is a linear program and all the arrival variables (v_{it}) are decided during the first stage while departure variables (w_{it}) are decided through finding the worst-case (maximization) realization of uncertainty in the second stage. It can be seen that although the variables u_t are defined to be continuous, since r_t is integer and arrivals and departure variables are binary, the optimal value for u_t is always integer.

In order to be able to solve the downstream capacity recourse problem (15a)-(15c), we apply strong duality to reformulate the inner-minimization as a maximization problem and also substitute for the definition of the uncertainty set $\mathcal{U}_l(\boldsymbol{x}, \boldsymbol{v})$. The downstream capacity recourse problem $R_l(\boldsymbol{x}, \boldsymbol{v}, \Gamma_l)$ is presented as follows:

$$\max \quad \sum_{t=1}^{T} \left[\sum_{i=1}^{n} (v_{it} - w_{it}) - r_t \right] \lambda_t \tag{16a}$$

$$s.t.$$

$$\sum_{i \in I_s, \hat{l}_i > 0, x_{ib'} = 0} \left[\frac{\sum_{t=1}^{T} (v_{it} - w_{it}) - \bar{l}_i}{\hat{l}_i} \right] \leq \Gamma_l^s \qquad \forall s \tag{16b}$$

$$w_{it} \geq x_{ib}^s \qquad t = t_b + \bar{l}_i + \hat{l}_i - 1, .., T, \forall s, i \in I_s, b \in B_s \tag{16c}$$

$$w_{it} \leq 1 - x_{ib} \qquad t = 1, .., t_b + \bar{l}_i - 1, \forall s, i \in I_s, b \in B_s \tag{16d}$$

$$w_{it} \leq w_{i,t+1} \qquad \forall s, i \in I_s, b \in B_s, \forall t \tag{16e}$$

$$0 \leq \lambda_t \leq e_t \qquad \forall t \tag{16f}$$

$$w_{it} \in \{0, 1\}. \qquad \forall i, t \tag{16g}$$

Note that in the downstream capacity sub-problem, v_{it} are first-stage variables and have known values when solving the second stage. λ_t is the dual variable associated with the SICU capacity constraint (15b). In an economical sense, it defines the price of denied admission to the SICU. Therefore, the objective function is the maximization of the cost for denied admissions (16a). The first constraints, (16b), enforce the budget of uncertainty for maximum possible deviations in LOS for each specialty group. These constraints allow the decision-maker to be able to have different risk preferences for different specialties. The second (16c) and third (16d) set of constraints, fix the value for w_{it} such that no patient can leave the SICU before its minimum LOS ($\bar{l}_i$) has passed, and each patient cannot stay in the SICU longer than largest possible LOS ($\bar{l}_i + \hat{l}_i$). The forth constraints, (16e), are defined based on the definition of the variable w_{it}. The fifth constraints, (16f), define the range for the values of the dual variables λ_t. Finally, the last set of constraints, (16g), define the domain for the variable w_{it}.

This formulation is in fact a bilinear program which is generally non-convex. Next we exploit some of the structrural properties of the downstream capacity recourse problem $R_l(x, y, \Gamma_l)$ presented by (16a)-(16g) and propose a mixed-integer linear programming reformulation that can be solved using traditional methods.

The following proposition characterizes the optimal value for the cost of denied admissions in the downstream capacity recourse problem, which helps us reformulate DC into a MILP.

Proposition 4.3 *For any $t = 1, \ldots, T$ in the second-stage optimal solution of the $R_l(x, v, \Gamma_l)$, $\lambda_t \in \{0, e_t\}$.*

Proof Similar to the proof for the Proposition 4.2.□

In the next section, we outline the steps of our exact solution methodology to solve our problem.

5 Solution Technique

The formulations for the SBC and DC subproblems are bilinear programs and, in general, are not convex. In order to be able to address the Robust Adaptive Surgery Planning with Downstream Capacity (RASP-DC) problem, we need to be able to solve each of these problems.

In the light of Propositions 4.1 and 4.2 we can reformulate the bilinear second-stage problem $R_d(x, \Gamma_d)$ as an MILP, $R_d^{MIP}(x, \Gamma_d)$, by defining $p_{ib}^s = \pi_b^s z_i, \forall s, i \in I_s, b \in B_s$ as follows:

$$\max \quad \sum_{s\in S}\sum_{b\in B_s}\sum_{i\in I_s} \bar{d}_i x_{ib}^s + \sum_{s\in S}\sum_{b\in B_s}\sum_{i\in I_s} \hat{d}_i x_{ib}^s p_{ib}^s - \sum_{s\in S}\sum_{b\in B_s} h_b \pi_b \tag{17a}$$

$s.t.$

$$\sum_{i\in I_s} z_i \le \Gamma_d^s \qquad s \in S \tag{17b}$$

$$0 \le \pi_b^s \le c_s \qquad s \in S, b \in B_s \tag{17c}$$

$$p_{ib}^s \le c_s z_i \qquad s \in S, b \in B_s \tag{17d}$$

$$z_i \in \{0,1\}, p_{ib}^s \ge 0. \qquad \forall s, b, i \tag{17e}$$

The MILP formulation (17a)-(17e) can be solved using standard solvers to obtain a solution for the surgery block capacity recourse problem. Next, we explore the structural properties of the SBCP in order to gain deeper insight that can be employed to improve the efficiency of the solution approach.

It can be observed from the formulation proposed for the surgery block capacity recourse problem $R_d^{MIP}(x, \Gamma_d)$ that the problem of calculating the worst-case over-time cost for surgery blocks can be decomposed into $|S|$ separate and independent problems that are connected only through the first-stage decision variables x_{ib}^s. In other words, the worst-case overtime costs can be calculated separately for each specialty $s \in S$. This property can be used when the size of the recourse problem over all specialties is large and a smaller problem for each specialty can be solved. In addition, this characteristic can be employed to devise a multi-cut approach similar to the well-known multi-cut L-shaped method in stochastic programming (see Birge and Louveaux (2011) and references there in).

As for the DC subproblem, the objective function for (16a) includes a bilinear term, $w_{it}\lambda_t$, and since w_{it} is defined to be a binary decision variable, we can simply reformulate the problem using the same technique in previous sections by having $q_{it} = w_{it}\lambda_t, \forall i, t$. Keeping in mind the negative coefficient of q_{it} in the objective function, the linearization requires the addition of $q_{it} \le \lambda_t, \forall i, t$, $q_{it} \le e_t w_{it}, \forall i, t$, and $q_{it} \ge \lambda_t - e_t(1 - w_{it}), \forall i, t$ as constraints. The first inequality is redundant as a result of the Proposition 4.3.The MILP downstream capacity recourse problem $R_l^{MIP}(x, v, \Gamma_l)$ can be written as:

$$
\begin{aligned}
\max \quad & \sum_{t=1}^{T}\sum_{i=1}^{n} v_{it}\lambda_t - \sum_{t=1}^{T}\sum_{i=1}^{n} q_{it} - \sum_{t=1}^{T} r_t\lambda_t && && \text{(18a)}\\
s.t. \quad & && && \\
& \sum_{i\in I_s, \hat{l}_i>0} \left[\frac{\sum_{t=1}^{T}(v_{it}-w_{it})-\bar{l}_i}{\hat{l}_i}\right] \le \Gamma_l^s && \forall s && \text{(18b)}\\
& w_{it} \ge x_{ib}^s && t = t_b+\bar{l}_i+\hat{l}_i-1,..,T, \forall s, i\in I_s, b\in B_s && \text{(18c)}\\
& w_{it} \le 1-x_{ib}^s && t = 1,..,t_b+\bar{l}_i-1, \forall s, i\in I_s, b\in B_s && \text{(18d)}\\
& w_{it} \le w_{i,t+1} && \forall s, i\in I_s, b\in B_s, \forall t && \text{(18e)}\\
& q_{it} \ge \lambda_t - e_t(1-w_{it}) && \forall s, i\in I_s, b\in B_s, \forall t && \text{(18f)}\\
& q_{it} \le e_t w_{it} && \forall i, t && \text{(18g)}\\
& 0 \le \lambda_t \le e_t && \forall t && \text{(18h)}\\
& w_{it} \in \{0,1\}. && \forall i, t && \text{(18i)}
\end{aligned}
$$

In the recourse formulation (18a)-(18i), the subset of variables w_{it} are fixed for some values of t. More specifically, the second and third constraints fix the values for w_{it} to either 1, or 0. In fact, for each patient i, the only binary variables that are not fixed are $\{w_{it} : t = t_b + \bar{l}_i - 1, ..., t_b + \bar{l}_i + \hat{l}_i - 1, t_b : x_{ib} = 1\}$. In other words, for each patient i, only $\hat{l}_i$ of the variables w_{it} are not fixed. Therefore, the most number of fractional values at optimality for the LP-relaxation of $R_l^{MIP}(x, y, \Gamma_l)$ is $\sum_{i=1}^{n} \hat{l}_i$. With the assumption that the deviations in the LOS at the SICU are relatively small compared to the decision making horizon T, a small percentage of variables can be fractional and the number of variables to be branched on is small. This is key to improve the performance of the solution method to make sure that the solution time for each recourse problem is not prohibitive.

The linear relaxation of this formulation does not necessarily yield optimal solutions which all the variables w_{it} are binary. The reason for non-integer values are solely due to the existence of the first constraint (18b) (the budget of uncertainty) and (18f) despite the fact that all the other constraints are facet defining (Bertsimas and Patterson, 1998).

Unlike the surgery block capacity problem, the downstream capacity recourse problem $R_l^{MIP}(x, y, \Gamma_l)$ cannot be decomposed into independent sub-problems for each specialty. The main reason is that patients from different specialties are sharing common resources in the SICU.

In Section 5.1, we explain why the existing methods in the literature, namely cutting-plane method based on Kelley's cutting-plane (CP) algorithm (Kelley, 1960) or L-shaped method (Van Slyke and Wets, 1969),(Thiele et al., 2009), and the column-and-constraint generation method (C&CG) introduced by Zeng and Zhao (2013) cannot be directly employed to solve the RASP-DC.

Next, in Section 5.2 we propose an adapted (C&CG) algorithm based on Zeng and Zhao (2013) to solve the RASP-DC and two-stage robust problems of similar structure, specifically stage-dependent uncertainty sets.

5.1 Deficiencies of Previously Developed Methods

The first approach proposed to solve the two-stage robust optimization (2SRO) problems is introduced by Thiele et al. (2009) and has its roots in the cutting-plane method (Kelley (1960) and Van Slyke and Wets (1969)). In their setting, the definition of the uncertainty set does not depend on the first-stage variables and the use of the proposed method produces optimal solutions. In the case of the CP method, a 2SRO problem is decomposed into a master and a recourse problem. The master problem generates first-stage decisions (in our case a surgery schedule is the first-stage set of decisions) and the recourse problem identifies the worst-case outcome of uncertainty and its associated cost for a given first-stage decision. In other words, the recourse problem in a 2SRO problem is a scenario generation problem that creates worst-case scenarios of uncertain parameters. This method relies on introducing the dual variables of the recourse problem into the master problem.

In the case of applying the CP method to the RASP-DC, the cut that is passed to the master problem from the DC subproblem has the following form:

$$\theta_l \geq \sum_{t=1}^{T} \left[\sum_{i=1}^{n} (v_{it} - w_{it}^k) - r_t \right] \lambda_t^k, k = 1, .., K, \tag{19}$$

where k is the number of constraints generated. Constraints (19), generated from the the DC subproblem, are invalid cuts and can potentially cut off the optimal solution. The reason is, in fact, due to the dependency between the first-stage variables v_{it} (arrival to the SICU) and uncertainty variables w_{it} (departure from the SICU). Note that the LOS for each patient is bounded such that $\bar{l}_i \leq \tilde{l}_i \leq \bar{l}_i + \hat{l}_i$. While these restrictions are considered in the DC subproblem using the constraints (16c) and (16d), constraints (19) do not explicitly consider the earliest possible arrival time associated with a given departure time.

As illustration, consider the case in which the arrival day for patient i is the second day ($v_{i,1} = 0$ and $v_{it} = 1, t = 2, ...T$) and $\bar{l}_i = 2$ and $\hat{l}_i = 1$. The maximum LOS for patient i is three days. Assume a case in which the optimal solution to the DC subproblem chooses the departure day to be day five ($w_{it} = 0, t < 5$ and $w_{it} = 1, t \geq 5$), therefore $l_i = \sum_{t=1}^{T} v_{it} - w_{it} = 3$. In addition, let let us assume that $\lambda_5 > 0$, which means that on day five the number of patients in need of a SICU bed has exceeded the SICU capacity. Since there is not enough capacity in the SICU for $t = 5$, the proposed constraint (19) will remove $v_{i1} = 1$ as a solution that can incur high transfer costs and the optimization will not allow the variable v_{i1} to be equal to one anymore. This means that patient i cannot be scheduled for a surgery on $t = 1$. However, this is an invalid removal of a feasible solution, since for patient i and his/her departure scenario on day $t = 5$, arrival on the first day is not a valid arrival (note that the maximum LOS for patient i is three days), but this constraint does not consider the restriction on the arrival times that should be in effect by the definition of the uncertainty set.

Column-and-constraint generation (C&CG) is proposed by Zeng and Zhao (2013) and introduces a large-scale deterministic equivalent formulation for the 2SRO that relies on identifying all the scenarios for the uncertain parameters. Unlike the CP methods, C&CG does not include the dual variables associated to the recourse problem in the master problem. The master problem includes constraints in the form of the deterministic problem for every realization of uncertainty. However, as the number of uncertain parameters increases, the number of scenarios for uncertainty increases exponentially. The authors present an iterative approach to address this issue. Since our solution approach is closely related to the C&CG, we include an explanation of the methodology in Appendix A.

Note that if we apply the proposed C&CG algorithm to our problem, specifically considering the downstream capacity subproblem, at each iteration k, we solve the DC subproblem and introduce this invalid constraint into the master problem:

$$\sum_{i=1}^{n} v_{it} - w_{it}^{k} \leq r_t + u_t^k, \ \forall t \tag{20}$$

in which w^k corresponds to the vector of worst-case departure times with respect to the first-stage surgery assignments in iteration k. u^k is the vector of recourse variables corresponding to the specific first-stage decisions and realization of uncertainty.

The reason that the constraint (20) is not a valid cut is similar to the argument made for the CP method, discussed earlier this section. In essence, this constraint does not consider the restrictions on the LOS for patients when it is passed to the master problem. It is important to note that if we do not consider the downstream capacity and only focus on the uncertainty in surgery duration, both the CP and C&CG method can be applied to solve the problem since the uncertainty set is not dependent on the first-stage decisions.

In the next subsection, we propose an adapted C&CG, such that it addresses the dependence of the uncertainty set and the first-stage decisions.

5.2 Adapted-C&CG Method

We adapt the C&CG method so it can address the issue of a first-stage-dependent uncertainty set as defined in our formulation. Our adapted column-and-constraint generation (A-C&CG) algorithm employs the deterministic formulation DORD-DC to address the decision making in the master problem, while the uncertainty is realized through solving the SBC and DC subproblems. By doing so, the uncertainty defined in the master problem for the LOS is an independent parameter. In the DC subproblem, we define the problem such that the definition of the uncertainty changes to the time the patient is released from the SICU given their admission time. Algorithm 1 presents the A-C&CG to solve the RASP-DC.

The proposed A-C&CG algorithm resembles the original C&CG while it allows for the use of more sophisticated uncertainty sets. In addition, it shows great flexibility in modeling a problem as a robust optimization problem. The second-stage formulations serve as a scenario-generation step and our algorithm allows us to use a different formulation, rather than the one based on the deterministic equivalent to generate the scenarios.

In the formulation for the master problem in the A-C&CG method, the original capacity constraints are employed. However, since the value for the uncertain parameters (deviations in surgery duration denoted by $\boldsymbol{z}$ and departure from the SICU denoted by $\boldsymbol{w}$) is not known in advance, at each iteration, we solve the master problem and obtain the optimal first-stage decisions. Then the worst-case realization of the uncertain parameters for the given first-stage decision variables is obtained by solving the recourse formulations. The information is passed back to the master problem by introducing new variables and constraints and then the master is solved again.

Note that A-C&CG has a better worst-case performance than the CP methods. It can be seen that in the C&CG method, only the extreme points of the uncertainty sets are required in the formulation of the problem. However for the CP method, the extreme points of both the uncertainty sets and the dual

Initialization;
Set $LB = -\infty, UB = +\infty, K = 0, O = \emptyset$.;
Master: Solve the following master problem.

$$\min \quad \sum_{s\in S}\sum_{i=1}^{n}\sum_{b\in B} a_{ib}x_{ib}^{s} + \theta \tag{21a}$$

$s.t.$

$$\theta \geq \sum_{s\in S}\sum_{b\in B_s\setminus\{b'\}} c_s o_b^k + \sum_{t=1}^{T} e_t u_t^k \qquad \forall k\in O \tag{21b}$$

$$\sum_{b\in B_s\cup\{b'\}} x_{ib}^{s} = 1 \qquad i\in I_s, s\in S \tag{21c}$$

$$\sum_{i\in I_s} d_i^k x_{ib}^{s} \leq h_b + o_b^k \qquad \forall k\leq K, b\in B_s\setminus\{b'\}, s\in S \tag{21d}$$

$$y_{it}^k \geq x_{ib}^{s} \qquad s\in S, i\in I_s, b\in B_s, t=t_b,..,t_b+l_i^k-1, \forall k\leq K \tag{21e}$$

$$\sum_{i=1}^{n} y_{it}^k \leq r_t + u_t^k \qquad \forall t, \forall k\leq K \tag{21f}$$

$$x_{ib}^{s}, y_{it}^k \in \{0,1\}, o_b^k\geq 0, u_t^k\geq 0 \qquad \forall s,i,b,t,\forall k\leq K \tag{21g}$$

Obtain the optimal solution $(x_{K+1}^*, \theta_{K+1}^*, y^{1*},..,y^{K*}, o^{1*},..,o^{K*}, u^{1*},..,u^{K*})$ and set $LB = c^T x_{K+1}^* + \theta_{K+1}^*$.;
Recourse:

- **Step 1-** Create arrival parameter v_{it} such that, if $x_{ib}^s = 1$, $v_{it} = 1, \forall t\geq t_b$, and $v_{it} = 0, \forall t < t_b$, else if $x_{ib'}^s = 1$ (assignment to the dummy block), set $v_{it} = 0, \forall t$.
- **Step 2-** Use parameter v_{it} to construct the DC subproblem (18a)-(18i).

Solve the SBC subproblem (17a)-(17e) with objective value S_{K+1}^*. Solve the DC subproblem (18a)-(18i) with objective value D_{K+1}^*. Update $UB = \min\{UB, c^T x_{K+1}^* + S_{K+1}^* + D_{K+1}^*\}$.;
Set $U \leftarrow \min\{U, \sum_{s\in S}\sum_{i\in I_s}\sum_{b\in B_s\cup\{b'\}} a_{ib}x_{ib}^{sk} + \sum_{s\in S}\sum_{b\in B_s}\left[\sum_{i\in I_s}(\bar{d}_i + \hat{d}_i z_i^k)x_{ib}^s - h_b\right]\pi_b^{sk} + \sum_{t=1}^{T}\left[\sum_{i=1}^{n}(y_{it} - w_{it}^k) - r_t\right]\lambda_t^k\}$;
if $U - L \leq \epsilon$ **then**
| The optimal solution for RASP-DC is found;
else
| go to Add-Cut routine;
end
Add-Cut:

- **Step 1** Using the results of DC subproblem and departure variables $w_{it}, \forall i, t$, calculate the LOS for each patient at iteration $K+1$ such that $l_i^{K+1} = \sum_{t=1}^{T} v_{it} - w_{it}^*, \forall i$.
- **Step 2** Add variables $o_b^{K+1}, \forall b$, $y_{it}^{K+1}, \forall i, t$, and $u_t^{K+1}, \forall t$ and the following constraints to the master problem:

$$\theta \geq \sum_{s\in S}\sum_{b\in B_s\setminus\{b'\}} c_s o_b^{K+1} + \sum_{t=1}^{T} e_t u_t^{K+1} \tag{22a}$$

$$\sum_{i\in I_s} d_i^{K+1} x_{ib}^{s} \leq h_b + o_b^{K+1} \qquad b\in B_s\setminus\{b'\}, s\in S \tag{22b}$$

$$y_{it}^{K+1} \geq x_{ib}^{s} \qquad s\in S, i\in I_s, b\in B_s, t=t_b,..,t_b+l_i^{K+1}-1 \tag{22c}$$

$$\sum_{t=1}^{T} y_{it}^{K+1} \leq r_t + u_t^{K+1} \qquad \forall t \tag{22d}$$

where d^{k+1} is the optimal solution (worst-case scenario for surgery duration) obtained by solving the SBC subproblem and l^{K+1} is obtained from the optimal solution of the DC subproblem. Update $K \leftarrow K+1, O \leftarrow O\cup\{K+1\}$ and go to Master routine;

Algorithm 1: A-C&CG Algorithm for RASP-DC.

variables of the inner-minimization problems are needed (for detailed proof see Zeng and Zhao (2013) and its electronic companion).

As it can be seen from the outline of the A-C&CG algorithm, at each iteration $|B|+(n+1)T$ new variables are added to the master problem. Furthermore, at each iteration k, $|B|+T+1+\sum_{i=1}^{n} l_i^k$ new constraints are added to the master problem. Note that l_i^k is the realization of the LOS for patient i at iteration k. Therefore, it is likely that for problems with a high number of patients and high levels of uncertainty (large values for $\hat{l}$), the size of the master problem will grow very fast.

6 Computational Experiments

6.1 Data and Problem Setting

The numerical results in this section are based on the practice configuration presented by Min and Yih (2010). There are 10 ORs and 32 available surgical blocks per week. The assignment of ORs to surgical blocks is shown in Table 1 and we use this example to produce a weekly block schedule. There are 9 different surgical groups that perform surgeries. Each group has at least one surgical block during the week, while some have multiple blocks. A patient can be scheduled for surgery in one of the available blocks. We assume each block is 8 hours long.

Table 1: Block schedule structure.

OR Room	Monday	Tuesday	Wednesday	Thursday	Friday
OR 1	ENT	ENT	ENT		
OR 2			ENT	ENT	ENT
OR 3	OBGYN		OBGYN		OBGYN
OR 4	ORTHO	ORTHO		ORTHO	ORTHO
OR 5		ORTHO		NEURO	
OR 6	GEN	GEN	GEN	GEN	
OR 7		GEN	GEN	GEN	GEN
OR 8	OPHTH	OPHTH		OPHTH	OPHTH
OR 9	VASCULAR		CARDIAC		VASCULAR
OR 10	UROLOGY		ORTHO		

Each specialty s has a mean surgery duration μ_d^s and a standard deviation σ_d^s. We use these statistics to generate patient specific surgery duration parameters. The nominal surgery duration for patient i, $\overline{d}_i$, is randomly generated from a lognormal distribution with the mean and standard deviation for the patient's required specialty (Strum et al., 2000). The worst case deviation in surgery duration for patient i, $\hat{d}_i$, is chosen as $\hat{d}_i = \alpha\sigma_d^{s_i}$ where $\alpha \sim U(0.5, 1.5)$.

We assume the average LOS in the SICU of ENT (Ear, Nose, and Throat), OBGYN (Obstetrics and Gynecology), ORTHO (Orthopedic), NEURO (Neurosurgery), GEN (General surgery), OPHTH (Ophthalmology), VASCULAR, CARDIAC and UROLOGY are 0.1 day, 2 days, 1.5 days, 2 days, 0.05 day, 0.05 day, 3.5 days, 2 days and 0.8 day, respectively. We further assume that the standard deviation for LOS in the SICU for each specialty is equal to its mean. The nominal LOS for patient i, $\bar{l}_i$, is then generated from a log normal distribution with proposed mean and standard deviations and rounded down to obtain an integer number. The worst case deviation in LOS for each patient i, $\hat{l}_i$, is generated from a uniform integer distribution $U_{int}(1, 4)$. Note that these numbers may be far from the reality, specifically considering that we are assigning uncertainty to all patients. However, the proposed setting for generating data allows us to test our proposed formulation to understand its behavior.

Each specialty has a relative importance factor which is used as a multiplier to address the relative importance of specialty over another. For example, heart surgeries are relatively more important than elective knee surgeries. This feature enables the decision-maker to specify the existing relative importance among different specialties. In addition, each patient has a specific multiplier (which in reality can be provided by her/his physician or a subject matter expert) which identifies the importance or the level of emergency of her/his case with respect to other patients in the same specialty. To generate cases, each patient receives a random integer from $U_{int}(1, 10)$. With this assumption, the optimization gives the emergency cases more priority in scheduling surgeries.

The percentage of patients in need of a specific surgery is used to generate lists of patients randomly. Overtime costs incurred to each block are chosen to reflect the relative importance of the specialty. Therefore, for each specialty the overtime cost is chosen to be 100 times the relative importance multiplier per hour. Table 2 provides detailed information on the surgery duration statistics, case mix, and the relative importance of each specialty.

Table 2: Statistics for surgery duration based on surgery type.

Surgical group	μ_d^s (minute)	σ_d^s (minute)	Percentage (% of surgeries)	Relative importance
ENT	74	37	21.34	1
OBGYN	86	40	9.26	2
ORTHO	107	44	23.26	2
NEURO	160	77	5.04	5
GEN	93	49	22.12	1
OPHTH	38	19	2.98	2
VASCULAR	120	61	8.2	4
CARDIAC	240	103	2.44	5
UROLOGY	64	52	5.36	3

For the patients that are assigned to the dummy block, we assume that they have to wait until the beginning of the next week for the next round of assignments. The waiting time is calculated by subtracting the assignment day from the initial day when the patient is added to the list. The cost for having one denied admission to the SICU or patient transfer (for each day that a patient is out of SICU) is set to be 100.

6.2 Performance Analysis

In order to evaluate the proposed solution methodology, tests were generated with different numbers of patients. We used a single budget of uncertainty to limit the deviations in LOS (Γ_l) and one for surgery duration (Γ_d) rather than a vector. In other words, we assume that the decision-maker does not have specialty-specific risk behavior. The value for each of these budgets can vary from zero to the number of patients. Therefore, for a problem with 10 patients, there are 100 combinations to be solved. To reduce the computational burden, we limit the solution time to be 1000 seconds for each specific choice of parameters Γ_d and Γ_l. All the computational experiments are coded in Python programming language and Gurobi solver is used as the optimization solver. All the tests are run on a Windows machine with an Intel Core i5, 3.20GHz

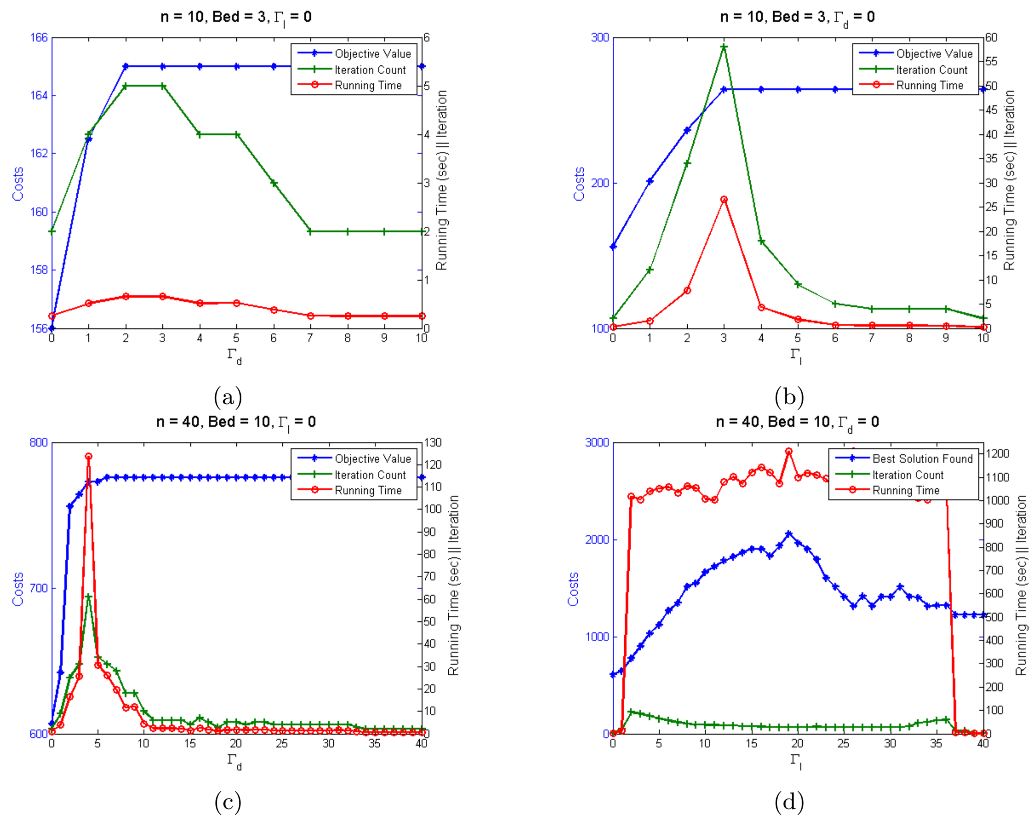


Figure 1: Plots for the performance of the algorithm when the source of uncertainty is changed.

CPU, and 4 GB of RAM.

In order to isolate the impact of uncertainty in surgery duration, we keep $\Gamma_l = 0$ and varied the value of Γ_d. The reverse is done to study the impact of uncertainty in LOS. Figure 1, illustrates an example with 10 patients and three SICU beds, and another example with 40 patients and 10 SICU beds.

Figure 1 illustrates the performance of the solution methodology for different variations of Γ_l and Γ_d. Figure (1a) shows an instance with 10 patients and three SICU beds, where $\Gamma_l = 0$, and Γ_d is increasing from zero to 10. It can be seen that for all choices of Γ_d, the instances are all solved to optimality under one second. On the other hand, Figure (1b) is the same instance when Γ_d is held constant at zero and Γ_l is increasing. Note that the time required to solve this instance when Γ_l is changing is significantly higher (with a maximum running time close to 30 seconds) than the case when $\Gamma_l = 0$. Note that solution times and iteration counts are shown on the same axis and scale. It can also be seen that for all values of Γ_l the instances are solved to optimality, but the objective value is higher compared to the case when the only source of uncertainty is in the surgery duration. While this behavior is very dependent on the structure of the objective and cost functions, there is a logical reason for this behavior under the given assumptions. Under the assumption of uncertain LOS, in the case of not having enough capacity in the SICU on a specific case, the optimization either moves the surgery of the patient to another day or assigns the patient to the waitlist for the following weeks. Therefore the uncertainty in the LOS coupled with congestion in the SICU, reduces the congestion in the operating rooms.

For the case of 40 patients, Figure (1c) shows that for all values of Γ_d when $\Gamma_l = 0$, the problem is solved to optimality. However, it can be seen in Figure (1d) that most of the instances went over the 1000 second time limit.

Depending on the value of Γ and size of the uncertainty set (possible realizations of uncertainty) the number of iterations required to solve each instance can be different. Through testing, we observed that the computational burden of the case when LOS uncertainty is much higher than the case when surgery durations are uncertain. In addition the case of uncertain LOS has not been studied as extensive as the case of uncertain surgery durations. From here on out, in all of our instances we fix $\Gamma_d = 0$ and only change the value of Γ_l. This helps us turn our focus to the impact of the surgery schedule on the downstream units and reduce the complexity of the results presented.

To evaluate the performance of the algorithm, problem instances with size $n = 5, 10, 15$ patients are generated randomly. The number of beds in the SICU is fixed to three beds. 10 instances of each size are generated and the objective value or best solution found is recorded. The best solution found for the recourse costs is calculated. The number of iterations and running time of the algorithm for each value of Γ_l is also recorded. The number of patients that could not be assigned for a surgery during the planning horizon and have to be postponed are also calculated. Table 1 in Appendix B of electronic companion includes these detailed results. Figure 2 compares the average objective value, running time, iteration count and the average number of patients that are postponed as Γ_l increases for these instances. It can be seen that for $n = 5$ and 10 all instances are solved to optimality. When $n = 15$ some of the instances could not be solved to optimality when $\Gamma_l \in \{2, 3, 4, 5\}$ within the 1000 second time limit. In our setting, as the number of patients (n) increases, the complexity of the problem increases exponentially. Since the uncertainty for LOS is chosen to be uniform random integers between one and four, the average number of possible realizations of uncertainty when $\Gamma_l = n$ is 2.5^n. It can be seen that the running time for the case of $n = 15$ is much larger than the smaller cases. Our assumption in defining the deviations in LOS for patients are for illustrative purposes. In the case of *more routine* surgeries where the uncertainty in LOS is insignificant, many patients will have $\hat{l} = 0$, thus significantly reducing the size of the uncertainty set, $\mathcal{U}_l$.

To better understand the quality of the solution provided by the our proposed model, we need to quantify the impact of a proposed surgery schedule on operations efficiency and quality of service.

6.3 Analyzing The Solution Quality

In order to understand and quantify the value of the proposed surgery schedule, we focus on the utilization rate of the SICU beds as an operational metric, where the decision-maker's aim is to keep the utilization rate of these expensive resources high to increase the efficiency of the system.

To analyze the quality of service and risks we focus on two important metrics. First, the probability of not having enough SICU capacity (which is equivalent to the probability of having to transfer a patient to a unit with lower level of care) is calculated. This is similar to probabilistic constraints which guarantee a probability for feasibility of one or a set of constraints (see Birge and Louveaux (2011), chapter 2.7). This is an important metric that helps decision-makers understand the chances of having to reduce the quality of service. In other words, it calculates the *risk* associated with a surgery schedule.

On the other hand, risk alone does not provide enough insight into the recourse actions that are required to be taken. Therefore, it is important to have an idea of how many transfers to lower quality units may be

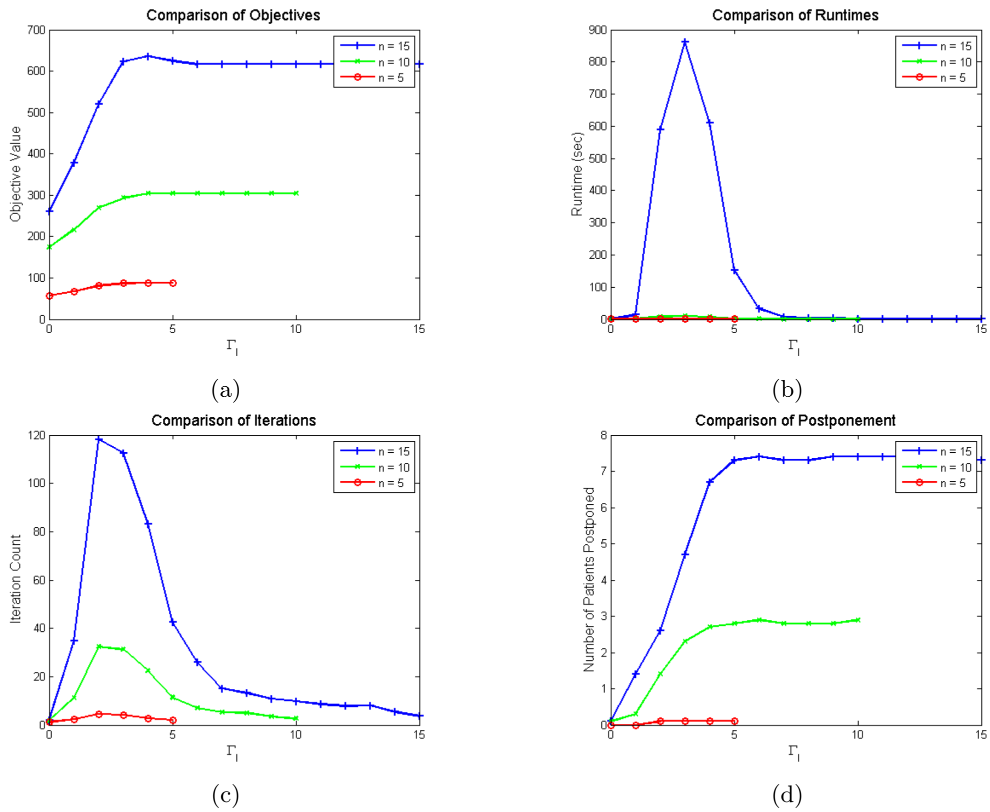


Figure 2: Aggregate comparison between the objective value (top left), running time (top right), iterations (bottom left), and the number of postponed patients (bottom right) for $n = 5, 10, 15$.

required. In many health care settings, the capacity of the SICU is determined by the number of nurses that are assigned to it. For a decision-maker, adding one nurse to the schedule to cover the expected number of transfers is much more desirable than having to hire five nurses. This metric calculates the *magnitude of risk* associated with a surgery schedule.

In order to calculate the proposed operational and risk measures, a simulation model is developed. The decision-maker has to decide the value of Γ_l before the realization of uncertainty. In reality, the number of patients whose LOS can deviate from the nominal value can be different from Γ_l. For each fixed value of Γ_l, $n + 1$ different cases of deviations can happen. Obviously, if the number of patients that have deviations is less than or equal to the Γ_l the probability of having a transfer is zero.

The simulation model randomly selects $k \in \{0, \dots, n\}$ patients and generates LOSs that are uniformly distributed between $\bar{l}$ and $\bar{l} + \hat{l}$. Using the proposed surgery schedule from the optimization step, we calculate if a transfer is required and the number of transfers required for that realization. In addition, we calculate the utilization rate for the SICU resources. This process is performed for 200 replication for each value of k (the number of patients that actually deviate from their nominal value). Then the average probability of transfer, average number of transfers required, and average utilization rate for the SICU is calculated.

Figure 3 illustrates the simulation results of previously generated instances of size $n = 5, 10, 15$. All plots show the uncertainty, as the number of patients who deviate from their nominal LOS on the X-axis. The left

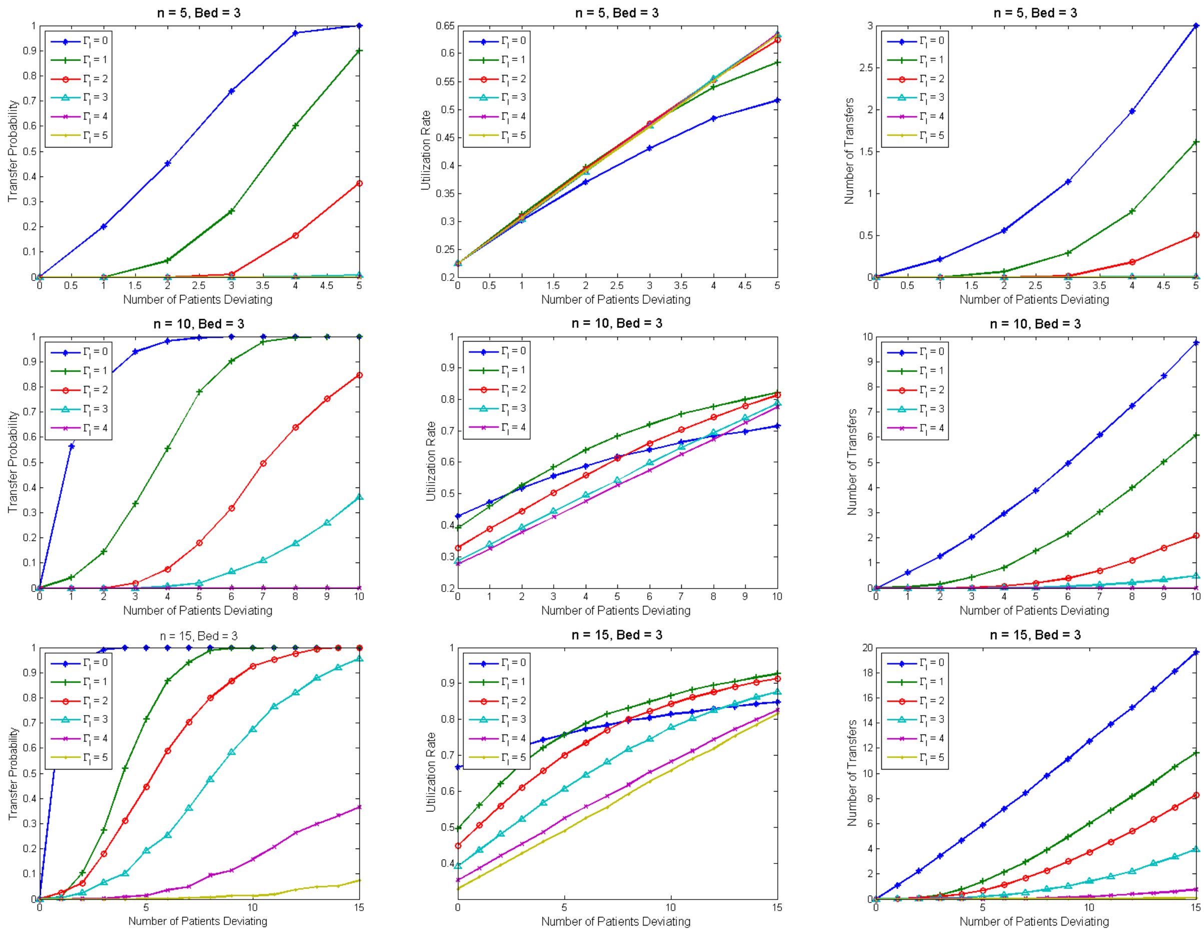


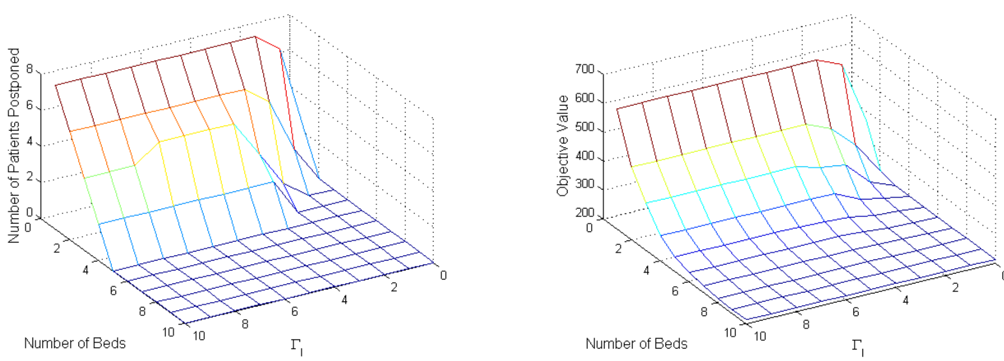
Figure 3: Simulation results: Impact of Γ_l and uncertainty on transfer probability, utilization rate, and required transfers.

column of figures, show the probability of having a transfer on the Y-axis and each line is associated with a surgery schedule with specific value for Γ_l. It can be seen that as uncertainty increases, meaning that more patients deviate from their nominal LOS, the probability of having a transfer increases. On the other hand, as we increase Γ_l, the probability of not having enough SICU capacity decreases, such that at $\Gamma_l = 3$ for the case of $n = 5$, this probability is almost zero for all cases of deviation. This behavior is repeated for cases with $n = 10$ and 15.

The middle column of Figure 3, illustrates the utilization rate of the SICU beds. It can be seen that as the uncertainty increases (moving to the right of the X-axis), the number of patients who deviate from their nominal LOS and have longer LOSs which translates into higher utilization rate. For $\Gamma_l > 0$, as Γ_l increases, the decision-maker becomes more conservative, thus creating schedules with larger slacks to accommodate uncertainty, which results in lower utilization rate. For the case of $\Gamma_l = 0$, as the uncertainty increases, the increase in the utilization rate is not as fast. This is due to the fact that by choosing $\Gamma_l = 0$, the decision-maker assumes no uncertainty in LOS, thus creating a tight schedule that assigns patients to the days at the beginning of the week while the rest of the week is empty. As uncertainty increases, due to the lack of available beds in the first few days, many transfers are required but the available capacity towards the end of the week remains untouched.

The right column of Figure 3 illustrates the average number of transfers that is required for each schedule in the presence of uncertainty. As uncertainty increases, there is greater chance for not having enough SICU beds, thus increasing the number of transfers. However increasing the value of Γ_l can greatly reduce the number of transfers that are required. For example, in the case of $n = 15$ and $\Gamma_l = 4$, even if all patients deviate from their nominal LOS, there will be on average four transfers required.

As mentioned before, one of the important factors to determine the capacity of the SICU is the number of available nurses. Note that this can be decided by the manager by designing the nurse schedules. Using our proposed methodology, the decision-maker can determine the staffing levels required to meet the throughput requirement. In other words, the decision-maker can identify the impact of extra resources on scheduling decisions. As an example, we generate an instance with $n = 10$ patients. By changing the number of beds from one to 10, optimal solutions for all values of Γ_l are obtained. To understand the throughput of each schedule, the number of patients that are postponed to have the surgery in future weeks is shown in Figure 4.



(a) (b)

Figure 4: Impact of SICU capacity on the throughput and cost.

It can be seen in Figure 4a that as the value of Γ_l increases, more conservative schedules are built, thus more patients get postponed to perform their surgeries later. Increasing the number of beds from one to three will reduce the postponements from eight down to four even in cases where the decision-maker decides to be very conservative (high Γ_l). Looking at the costs of surgery schedules in Figure 4b, the decision-maker can decide the best level of staffing for the SICU. In addition, using this method, the decision-maker is capable of understanding the trade-off between the SICU resource level and the throughput of the operating rooms.

We have assumed that the cost of postponing a surgery is equal to the cost of waiting to have a surgery during the next week. A patient that is postponed may not receive a surgery appointment at the beginning of, or during, the next week. Therefore, the decision-maker may require to incur higher costs for patients that are postponed to perform their surgeries later. In order to accommodate this feature, we introduced a multiplier $\gamma \in \{1, \ldots, 10\}$. This multiplier is multiplied by the waiting cost for the patients that are postponed. As the multiplier increases, the cost of postponing a patient increases. Depending on the cost structure, the postponing cost can surpass the cost of having a transfer out of the SICU, in which case, the optimization decides to risk having a transfer. This feature is important for health care providers as postponing a surgery can be very costly at destination medical centers.

As an example, an instance with $n = 10$ patients and three SICU beds is generated. We changed the value of γ from 1 to 10. The case of $\gamma = 1$ is equivalent to our original assumption on the cost structure. We solved this instance for all values of Γ_l and Figure 5 shows how the optimization risks having transfers as the cost of postponing increases. It can be seen that in the case of low downstream capacity, the optimization schedules patients with large values of $\hat{l}$ far away form each other. This is done to minimize the chance of overlapping stays in the SICU which can reduce the number of SICU beds for a long period. This can be used as a rule of thumb for practitioners to mitigate the impact of uncertainty in LOS while scheduling surgeries.

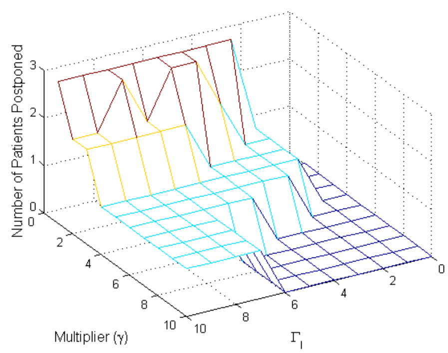


Figure 5: Impact of increasing the postponement cost on number of postponements.

We aim to propose a modeling approach that considers the uncertainty in surgery duration and LOS and enables the decision-maker to adjust for her/his risk preferences. In addition, the optimization model, coupled with our simulation model, can provide the decision-maker with variety of trade offs that can be made in managing the surgery planning process. Our approach can be used as a dashboard for decision-makers to provide them with different alternatives and the characteristics of each schedule. This will greatly improve the decision making process by assisting the managers with making well-informed and educated decisions. Figure 6 provides an example based on the tests we ran for the case with $n = 15$ patients and three SICU beds.

It can be seen in Figure 6, that the probability of having transfers (left Y-axis) changes for different risk preferences $\Gamma_l = 2$ and $\Gamma_l = 4$ in the face uncertainty. If all patients have deviations in their LOS, the probability of requiring a transfer is one at $\Gamma_l = 2$, while it is less than 0.4 if $\Gamma_l = 4$. In terms of the number of transfers required when all patients have deviations (right Y-axis), at $\Gamma_l = 2$, more than eight transfers are needed while at $\Gamma = 4$, the required number of transfers is less than one. As for the utilization rate, it can be seen that when all patients have deviation in their LOS, both values for Γ_l have utilization rates greater than 0.8. This is a great example of how our proposed model can help decision-makers choose well-informed alternatives.

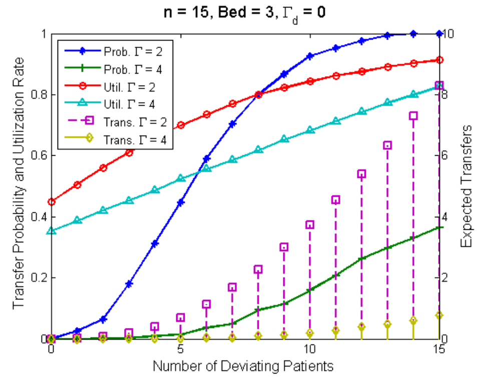


Figure 6: Comparing operational and risk metrics between $\Gamma_l = 2$ and $\Gamma_l = 4$.

7 Conclusion

In this paper, we have proposed a formulation for surgery scheduling while considering the downstream units. We apply two-stage robust optimization to address the inherent uncertainty in surgery duration and length-of-stay in the downstream unit. Since the uncertainty in LOS translates into uncertainty in the number of constraints, a novel modeling approach is proposed to address the challenges in modeling this aspect. The proposed formulation can be applied to other domains with similar downstream considerations such as project scheduling. We studied the structural properties of the proposed formulations and reformulated them into solvable MILPs and proposed an exact solution algorithm to solve this problem.

Extensive computational experiments show that this model has the potential of being employed to manage multi-stage care operations. Our simulation model quantifies the impact of our robust model on the utilization of the downstream resources. Our framework, coupled with a simulation model, helps the decision-maker understand the level of risk associated with each proposed surgery schedule and the impact of her/his attitude towards risk. An important insight is that by considering the uncertainty in the LOS, the congestion in the OR can be implicitly alleviated. In addition, the existing trade-offs between different elements in this setting are shown in our computational experiments.

The proposed algorithm may not be efficient for cases with large number of patients with large uncertainty sets. Finding better lower bounds can greatly improve the running time of the proposed algorithm. Effective bounding techniques show promise by using the scenarios generated for different values of Γ can be employed to improve the performance of the algorithm. We hope to continue this development in our future work.

Appendices

A Column-and-Constraint Generation Method

Consider the following two-stage robust optimization problem:

$$\min_{\{x:Ax\geq b, x\subseteq\mathbb{R}^n_+\}} c^T x + \max_{u\in\mathcal{U}} \min_{y\in S(x,u)} d^T y \tag{23}$$

in which $S(x,u) = \{y : Wy \geq h - Ex - Mu, y \subseteq \mathbb{R}^m_+\}$. All the uppercase letters are matrices of appropriate sizes. Lowercase letters denote vectors and we assume that the dimensions match. The authors show that this problem can be reformulated as the following:

$$\min_x \quad c^T x + \theta \tag{24a}$$
$$s.t.$$
$$Ax \geq b \tag{24b}$$
$$\theta \geq d^T y^k \qquad k = 1, .., r \tag{24c}$$
$$Ex + Wy^k \geq h - Mu^k \qquad k = 1, .., r \tag{24d}$$
$$x \subseteq \mathbb{R}^n_+, y^k \subseteq \mathbb{R}^n_+ \qquad k = 1, .., r \tag{24e}$$

The recourse problem $\mathcal{Q}(x) = \{\max_{u \in \mathcal{U}} \min d^T y : Wy \geq h - Ex - Mu, y \subseteq \mathbb{R}^m_+\}$ identifies the worst-case scenario for uncertain parameter u for a given first-stage decision x. Assuming that the second-stage problem has complete recourse (no need for maintaining feasibility) and is bounded for all feasible first-stage decisions, the C&CG algorithm has the following structure:

- **Step 1-** Set $LB = -\infty, UB = +\infty, k = 0$, and $O = \emptyset$.
- **Step 2-** Solve the following master problem.

$$\min_x \quad c^T x + \theta \tag{25a}$$
$$s.t.$$
$$Ax \geq b \tag{25b}$$
$$\theta \geq d^T y^i \qquad \forall i \in O \tag{25c}$$
$$Ex + Wy^i \geq h - Mu^i \qquad \forall i \leq k \tag{25d}$$
$$x \subseteq \mathbb{R}^n_+, y^i \subseteq \mathbb{R}^n_+ \qquad \forall i \leq k \tag{25e}$$

  Obtain the optimal solution $(x^*_{k+1}, \theta^*_{k+1}, y^{1*}, .., y^{k*})$ and set $LB = c^T x^*_{k+1} + \theta^*_{k+1}$.

- **Step 3-** Solve the subproblem $\mathcal{Q}(x^*_{k+1})$ in and update $UB = \min\{UB, c^T x^*_{k+1} + \mathcal{Q}(x^*_{k+1})\}$.
- **Step 4-** If $UB - LB \leq \epsilon$, optimal solution is found, return x^*_{k+1} and terminate. Otherwise do
  - **Step 4.1** Add the new variables y^{k+1} and the following constraints to the master problem:

$$\theta \geq d^T y^{k+1} \tag{26a}$$
$$Ex + Wy^{k+1} \geq h - Mu^{k+1} \tag{26b}$$

    where u^{k+1} is the optimal solution (worst-case scenario) obtained solving $\mathcal{Q}(x^*_{k+1})$. Update $k \leftarrow k+1, O \leftarrow O \cup \{k+1\}$ and go to Step 2.

B Appendix

Table 3 presents detailed results for $n = 5, 10, 15$ patients with three beds in the SICU. The number of beds in the SICU is fixed to 3 beds. 10 instances of each size are generated and the objective value (obj) or best solution found is recorded. The best solution found for the recourse costs (ω) is calculated. The number of iterations and running time of the algorithm for each value of Γ_l is also recorded. The number of patients that could not be assigned for a surgery during the planning horizon and have to be postponed are also calculated. Table 3 in Appendix B, shows the average (μ) and standard deviation (σ) of the performance parameters.

Table 3: Aggregate results for 10 instances of problems with $n = 5, 10, 15$.

Setting	Γ_l	Obj-μ	Obj-σ	ω-μ	ω-σ	Iteration-μ	Iteration-σ	Runtime-μ	Runtime-σ	Gap-μ	Gap-σ	Postponed-μ	Postponed-σ
n = 5	0	57.1	16.50	0	0	1.1	0.31	0.09	0.03	0	0	0	0
	1	67.6	17.54	0	0	2.3	0.48	0.22	0.05	0	0	0	0
	2	80.9	22.01	0	0	4.5	2.36	0.43	0.23	0	0	0.1	0.31
	3	87.4	21.7	0	0	4.2	1.54	0.4	0.14	0	0	0.1	0.31
	4	87.6	21.78	0	0	2.8	0.63	0.26	0.06	0	0	0.1	0.31
	5	87.6	21.78	0	0	2.1	0.31	0.19	0.03	0	0	0.1	0.31
n = 10	0	173.9	144.85	0	0	1.8	0.42	0.22	0.05	0	0	0.1	0.31
	1	216.6	166.37	10	31.62	11.1	6.0	1.59	0.9	0	0	0.3	0.48
	2	269.3	182.89	0	0	32.3	21.77	8.55	8.36	0	0	1.4	1.26
	3	293.1	175.59	0	0	31.2	20.94	9.73	9.64	0	0	2.3	0.94
	4	304.7	173.16	0	0	22.4	14.82	6.64	6.62	0	0	2.7	0.67
	5	304.7	173.16	0	0	11.4	4.67	2.17	1.31	0	0	2.8	0.78
	6	304.7	173.16	0	0	7	2.7	1.1	0.59	0	0	2.9	0.73
	7	304.7	173.16	0	0	5.2	0.91	0.7	0.15	0	0	2.8	0.78
	8	304.7	173.16	0	0	5	2.05	0.7	0.4	0	0	2.8	0.78
	9	304.7	173.16	0	0	3.4	1.57	0.44	0.2	0	0	2.8	0.78
	10	304.7	173.16	0	0	2.6	1.34	0.34	0.18	0	0	2.9	0.73
n = 15	0	261.2	100.02	0	0	2	0	0.35	0.06	0	0	0.1	0.31
	1	378.6	115.84	10	31.62	34.8	12.47	13.77	9.09	0	0	1.4	1.5
	2	519	110.29	90	73.78	118.1	38.45	588.06	451.81	5.26	7.27	2.6	1.95
	3	623.2	132.08	120	91.89	112.5	40.66	860.52	348.16	6.1	4.56	4.7	1.56
	4	635.5	131.82	40	51.63	83.2	23.53	609.78	465.99	1.44	2.74	6.7	1.63
	5	624.1	126.31	10	31.62	42.5	19.3	152.55	318.08	0.26	0.85	7.3	1.49
	6	615.9	138.09	0	0	25.9	14.41	32.13	70.03	0	0	7.4	1.26
	7	615.9	138.09	0	0	15	6.73	6.62	7.63	0	0	7.3	1.33
	8	615.9	138.09	0	0	13.2	4.58	4.14	2.37	0	0	7.3	1.33
	9	615.9	138.09	0	0	10.8	3.88	2.91	1.58	0	0	7.4	1.26
	10	615.9	138.09	0	0	9.8	3.52	2.39	0.98	0	0	7.4	1.26
	11	615.9	138.09	0	0	8.5	3.65	1.96	1.05	0	0	7.4	1.26
	12	615.9	138.09	0	0	7.9	3.84	1.65	0.97	0	0	7.4	1.26
	13	615.9	138.09	0	0	8	9.78	2.02	3.01	0	0	7.3	1.3
	14	615.9	138.09	0	0	5.4	5.44	1.04	1.17	0	0	7.3	1.33
	15	615.9	138.09	0	0	3.6	5.05	0.69	1.11	0	0	7.3	1.33

References

Addis, B., Carello, G., Grosso, A., Lanzarone, E., Mattia, S., and Tànfani, E. (2015). Handling uncertainty in health care management using the cardinality-constrained approach: Advantages and remarks. *Operations Research for Health Care*, 4:1–4.

Addis, B., Carello, G., and Tànfani, E. (2014). A robust optimization approach for the operating room planning problem with uncertain surgery duration. In *Proceedings of the International Conference on Health Care Systems Engineering*, pages 175–189. Springer.

Argo, J. L., Vick, C. C., Graham, L. A., Itani, K. M., Bishop, M. J., and Hawn, M. T. (2009). Elective surgical case cancellation in the veterans health administration system: identifying areas for improvement. *The American Journal of Surgery*, 198(5):600–606.

Atamtürk, A. and Zhang, M. (2007). Two-stage robust network flow and design under demand uncertainty. *Operations Research*, 55(4):662–673.

Ben-Tal, A., Goryashko, A., Guslitzer, E., and Nemirovski, A. (2004). Adjustable robust solutions of uncertain linear programs. *Mathematical Programming*, 99(2):351–376.

Ben-Tal, A. and Nemirovski, A. (1998). Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805.

Bertsimas, D., Brown, D. B., and Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM review*, 53(3):464–501.

Bertsimas, D. and Caramanis, C. (2010). Finite adaptability in multistage linear optimization. *Automatic Control, IEEE Transactions on*, 55(12):2751–2766.

Bertsimas, D. and Patterson, S. S. (1998). The air traffic flow management problem with enroute capacities. *Operations research*, 46(3):406–422.

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

Birge, J. R. and Louveaux, F. (2011). *Introduction to stochastic programming*. Springer.

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

Demeulemeester, E., Beliën, J., Cardoen, B., and Samudra, M. (2013). Operating room planning and scheduling. In *Handbook of Healthcare Operations Management*, pages 121–152. Springer.

Deng, Y., Shen, S., and Denton, B. (2014). Chance-constrained surgery planning under uncertain or ambiguous surgery duration. *Available at SSRN 2432375*.

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

El Ghaoui, L., Oustry, F., and Lebret, H. (1998). Robust solutions to uncertain semidefinite programs. *SIAM Journal on Optimization*, 9(1):33–52.

Erdogan, S. A., Denton, B. T., Cochran, J., Cox, L., Keskinocak, P., Kharoufeh, J., and Smith, J. (2011). Surgery planning and scheduling. *Wiley Encyclopedia of operations research and management science. http://ca. wiley. com/WileyCDA/Section/id-380199. html.*

Ferrand, Y. B., Magazine, M. J., and Rao, U. S. (2014). Managing operating room efficiency and responsiveness for emergency and elective surgeries — a literature survey. *IIE Transactions on Healthcare Systems Engineering*, 4(1):49–64.

Fügener, A., Hans, E. W., Kolisch, R., Kortbeek, N., and Vanberkel, P. T. (2014). Master surgery scheduling with consideration of multiple downstream units. *European Journal of Operational Research.*

Gabrel, V., Lacroix, M., Murat, C., and Remli, N. (2011). Robust location transportation problems under uncertain demands. *Discrete Applied Mathematics.*

Gabrel, V., Murat, C., and Thiele, A. (2014). Recent advances in robust optimization: An overview. *European Journal of Operational Research*, 235(3):471–483.

Gallo, G. and Ülkücü, A. (1977). Bilinear programming: an exact algorithm. *Mathematical Programming*, 12(1):173–194.

Guerriero, F. and Guido, R. (2011). Operational research in the management of the operating theatre: a survey. *Health care management science*, 14(1):89–114.

Gul, S., Denton, B. T., and Fowler, J. W. (2012). A multi-stage stochastic integer programming model for surgery planning.

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

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

Jonnalagadda, R., Walrond, E., Hariharan, S., Walrond, M., and 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.

Kelley, Jr, J. E. (1960). The cutting-plane method for solving convex programs. *Journal of the Society for Industrial & Applied Mathematics*, 8(4):703–712.

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

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

Shylo, O. V., Prokopyev, O. A., and Schaefer, A. J. (2012). Stochastic operating room scheduling for high-volume specialties under block booking. *INFORMS Journal on Computing*, 25(4):682–692.

Sobolev, B. G., Brown, P. M., Zelt, D., and FitzGerald, M. (2005). Priority waiting lists: Is there a clinically ordered queue? *Journal of evaluation in clinical practice*, 11(4):408–410.

Strum, D. P., May, J. H., and Vargas, L. G. (2000). Modeling the uncertainty of surgical procedure times: comparison of log-normal and normal models. *Anesthesiology*, 92(4):1160–1167.

Thiele, A., Terry, T., and Epelman, M. (2009). Robust linear optimization with recourse. *Rapport technique*, pages 4–37.

Truong, V.-A., Wang, X., and Liu, N. (2013). Integrated scheduling and capacity planning with considerations for patients' length-of-stays.

Utzolino, S., Kaffarnik, M., Keck, T., Berlet, M., and Hopt, U. T. (2010). Unplanned discharges from a surgical intensive care unit: Readmissions and mortality. *Journal of critical care*, 25(3):375–381.

van Oostrum, J. M., Van Houdenhoven, M., Hurink, J. L., Hans, E. W., Wullink, G., and Kazemier, G. (2008). A master surgical scheduling approach for cyclic scheduling in operating room departments. *OR spectrum*, 30(2):355–374.

Van Slyke, R. M. and Wets, R. (1969). L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM Journal on Applied Mathematics*, 17(4):638–663.

Zeng, B. and Zhao, L. (2013). Solving two-stage robust optimization problems using a column-and-constraint generation method. *Operations Research Letters*, 41(5):457–461.

Zhao, L. and Zeng, B. (2012). An exact algorithm for two-stage robust optimization with mixed integer recourse problems. *Submitted, available in optimization-online, University of South Florida.*