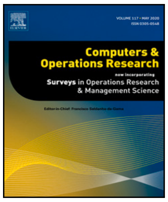
Contents lists available at ScienceDirect

Computers and Operations Research

journal homepage: www.elsevier.com/locate/cor



Survey in operations research and management science

Stochastic optimization approaches for elective surgery scheduling with downstream capacity constraints: Models, challenges, and opportunities

Karmel S. Shehadeh [a,*], Rema Padman [b]

[a] Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, USA
[b] The Heinz College of Information Systems and Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA

ARTICLE INFO

Keywords:
Surgery scheduling
Operating rooms
Downstream constraints
Stochastic optimization

ABSTRACT

Elective surgery patients have surgery in the operating room (OR), and then recover in one or more downstream recovery units for several consecutive hours or days after surgery. Upstream scheduling that focuses on OR alone or a resource-constrained scheduling approach that fails to account for the inherent uncertainty in surgery durations and postoperative downstream recovery times yield sub-optimal or infeasible schedules and, consequently, higher cost and reduced quality of care. However, modeling such uncertainties at multiple levels is challenging, especially with limited reliable data on the random parameters in the models. Moreover, sequencing of surgical and recovery activities, and the multiple conflicting objectives of all parties involved (including management, clinicians, patients), lead to *a class of complex combinatorial and multi-criteria stochastic optimization problems.* In this review, we focus on stochastic optimization (SO) approaches for elective surgery scheduling and downstream capacity planning. We describe the art of formulating and solving such a class of stochastic resource-constrained scheduling problems, provide an analysis of existing SO approaches and their challenges, and highlight areas of opportunity for developing tractable, implementable, and data-driven approaches that might be applicable within and outside healthcare operations, particularly where multiple entities/jobs share the same downstream limited resources.

> "*Each of the uncertainty veils carries a secret promise for creative mathematical art.*"
>
> [–Dr. Karmel S. Shehadeh (2021)]

> "*Problems worthy of attack prove their worth by hitting back.*"
>
> [–Piet Hein, Grooks (1966)]

1. Introduction

Surgical suites, consisting of operating rooms (OR) and their downstream recovery units such as Post-Anesthesia Care Unit (PACU) or Surgical Intensive Care Unit (SICU), are the major revenue as well as cost generating departments in hospitals. ORs alone generate about 40%–70% of revenues and incur 20%–40% of operating costs (Bovim et al., 2020; Freeman et al., 2016; Jackson, 2002; Li et al., 2016; Macario et al., 1995; Samudra et al., 2016; Viapiano and Ward, 2000; Wang et al., 2019). The SICU, for example, accounts for about 15%–40% of hospital costs (Brilli et al., 2001; Halpern et al., 2007; Kim et al., 2015b; Reis Miranda and Jegers, 2012).

According to statistics from the Agency for Healthcare Research and Quality (AHRQ), 17.2 million hospital visits in 2014 included elective surgical procedures in the US alone. Freeman et al. (2016) state that 60%–70% of all patients admitted to a hospital require some surgical intervention. In many surgical suites, patients undergo surgery in OR and then recover in one or multiple downstream units (see Fig. 1). Most recovery units have limited capacity (such as the number of recovery beds), are very expensive to operate, and often fall short of meeting the existing demand in many hospitals (Liu et al., 2019; Ouyang et al., 2020).

As an essential area for cost and patient outcomes management by improving utilization of OR resources and quality of surgical care, OR planning and surgery scheduling problems have received intense research attention over many decades (Cardoen et al., 2010; Hof et al., 2017; May et al., 2011; Samudra et al., 2016; Zhu et al., 2019). For comprehensive surveys of OR and elective surgery scheduling problems, we refer to Cardoen et al. (2010), Gartner and Padman (2019, 2017b), Hof et al. (2017), May et al. (2011), Samudra et al. (2016), and Zhu et al. (2019).

OR and surgery planning involves three decision stages: strategic, tactical, and operational (Cardoen et al., 2010; Zhu et al., 2019). The strategic level consists of long-term planning problems such as deciding

[*] Corresponding author.
E-mail addresses: kas720@lehigh.edu, kshehadeh@lehigh.edu (K.S. Shehadeh), rpadman@cmu.edu (R. Padman).

https://doi.org/10.1016/j.cor.2021.105523
Received 15 June 2020; Received in revised form 18 July 2021; Accepted 16 August 2021
Available online 26 August 2021

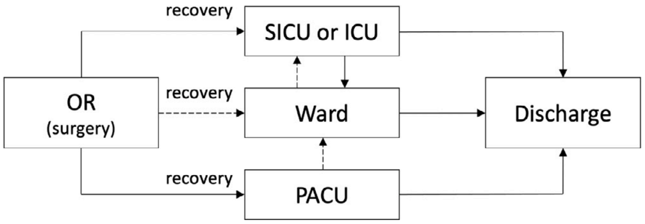


Fig. 1. Potential pathways of elective surgery patients after surgery. Dashed paths are less likely.

the size and number of ORs, number and specialties of surgeries to be planned, the number of resources required (i.e., capacity allocation and planning), etc. At the tactical level, studies often focus on developing a Master Surgery Schedule (MSS), specifying the pre-assignments of surgical blocks or OR time to surgical specialties, usually under the block scheduling strategy. Finally, several studies decompose the operational level of OR and surgery scheduling into two steps: advance scheduling and allocation scheduling (Magerlein and Martin, 1978; Samudra et al., 2016; Zhu et al., 2019). Advance scheduling is the process of assigning patients to specific surgical blocks (i.e., assigning a date for each surgery). Allocation scheduling is the process of scheduling a starting time for each surgery, i.e., the sequence of surgery start times on each OR day.

Strategic and tactical problems are long-term (months, a year, or longer) planning and scheduling problems that have been extensively studied in the literature. Advance and allocation scheduling of elective surgeries under the limited capacity of downstream units and associated uncertainty are very pressing and frequently observed (i.e., short-term) problems in practice. In this review, we focus on advance and allocation scheduling of elective surgery with consideration of downstream capacity and uncertainty limitations.

Stochasticity is an intrinsic property of the advance and allocation surgery scheduling problems since surgical and recovery activities are subject to multiple sources of uncertainties. As we discuss in Section 2, surgery duration and postoperative length-of-stay (LOS) in the downstream recovery units are, by far, two of the primary (patient-related) sources of disruption to the daily schedule of surgery and patient flow in the downstream recovery units. Most of the existing literature on elective surgery scheduling, however, focuses on OR rooms alone (ignoring the capacity of the subsequent downstream recovery units) and are either deterministic or consider the variability of surgery duration as the primary source of disruption to the OR schedule (see, e.g., Chang et al. (2014), Denton et al. (2007, 2010), Hsu et al. (2003), Shehadeh et al. (2019), Pham and Klinkert (2008) to name few).

Within the limited literature that considers the subsequent postoperative stay and the capacity of recovery units, most studies assume that surgery duration and postoperative LOS follow fully known unimodal distributions, typically lognormal (Bai et al., 2017; Min and Yih, 2010; Varmazyar et al., 2020; Zhang et al., 2019). In reality, it is challenging, if not impossible, to predict surgery durations and LOS in any of the downstream units in advance, particularly at the time when the surgery is scheduled (Shehadeh and Padman, 2021; Wang et al., 2019). In addition, given the diversity in patient characteristics and surgery contexts, the probability distributions of these random parameters may be multi-modal and ambiguous (unknown) rather than unimodal and fully known. Here, we use the term "*multi-modal*" in a slightly informal way to refer to the tendency of a random parameter to display, or belong to, several spatially distinct distributions; the terms *multi-modality*, *bimodal, bimodality* are to be interpreted analogously.

Indeed, upstream scheduling that focuses on OR alone or a resource-constrained scheduling approach that fails to account for uncertainty (i.e., variability) and ambiguity (i.e., lack of distributional information) of surgery duration and LOS in recovery units often leads to sub-optimal or infeasible schedules, delays, unpredictable availability of recovery beds, and inefficient use of capacity (Esfahani and Kuhn, 2018; Deng et al., 2014; Liu et al., 2019; Shehadeh et al., 2020a; Wang et al., 2019). Mitigating the impact of such uncertainties at multiple levels is, however, challenging for two primary reasons. First, reliable data on random parameters is often not available (Cardoen et al., 2010; Esfahani and Kuhn, 2018; Gupta and Denton, 2008). Even if available, historical data only represents a sample taken from the true (yet ambiguous) distributions of random parameters, and the future realizations of random parameters may not be distributed as in the past. If we attempt to optimize surgery schedules using a data sample from a biased distribution, then the resulting scheduling decisions may have a disappointing out-of-sample performance when implemented under the true distribution (Esfahani and Kuhn, 2018; Smith and Winkler, 2006).

Second, considering uncertainty, sequencing of surgical and recovery activities, and the multiple conflicting objectives of all parties involved (including management, clinicians, and patients), leads to a class of *complex and computationally prohibitive combinatorial and multi-criteria stochastic optimization problems*. In fact, elective surgery scheduling with two-stages (OR and a recovery unit) is a special case of stochastic hybrid flow shop scheduling, which is NP-hard (Pinedo, 2016; Ruiz and Vázquez-Rodríguez, 2010; Varmazyar et al., 2020; Wang et al., 2015). The theoretical complexity and solution intractability of these problems have long prevented optimization-based surgery scheduling and downstream capacity planning models from being implemented in practice (Cardoen et al., 2010; Gupta and Denton, 2008; Shehadeh et al., 2019; Zhu et al., 2019).

With more hospitals adopting new and innovative healthcare information technologies (HIT), hospitals are acquiring the necessary information technology support to model uncertainty and coordinate capacity usage among different units. Unfortunately, despite the advances in stochastic optimization (SO) and HIT, we still do not have systematic frameworks that integrate these fields for developing computationally efficient and data-driven SO methods for elective surgery scheduling and downstream capacity planning. While some studies develop and evaluate scheduling heuristics considering downstream resources, often through the use of sophisticated simulation, they lack performance guarantees (i.e., sub-optimal) and make unrealistic distributional assumptions, which lead to poor operational performance (see, e.g., see Table A.1 in Appendix and Ewen and Mönch (2014), Lee and Yih (2014, 2012), Saremi et al. (2013) to name a few).

Motivated by our research collaboration with a large health system in Pennsylvania that has provided insights into some current challenges in elective surgery scheduling, in this paper, we review the state-of-the-art SO approaches to (optimal) stochastic elective surgery scheduling and capacity planning within the block-booking framework (i.e., OR schedules are divided into multiple blocks of defined lengths, and each block is reserved for a specific specialty). Specifically, we focus our review on stochastic programming (SP), robust optimization (RO), and distributionally robust optimization (DRO) approaches for two primary classes of elective surgery scheduling problems: (1) elective surgery scheduling with SICU and ward capacity constraints, and (2) elective surgery scheduling with PACU capacity constraints.

In the first class, we consider surgical suites that perform moderate to high invasive surgeries (e.g., open-heart surgery), after which patients often recover in the SICU, then possibly in an inpatient ward for several days. In the second class, we consider surgical suites that offer lower risk, and often minimally invasive, surgeries (e.g., gallbladder surgery), after which a patient recovers in a PACU bed for several hours before being discharged home (or to inpatient ward in some rare cases). Surgery durations and postoperative LOS are at a different time scale in each class. Resource requirements and constraints are also different in each class. Hence, the practical and theoretical challenges associated with surgery selection, sequencing, and scheduling are materially different for each class and thus worth separate analysis.

Emergency surgeries are often treated in dedicated units in most hospitals, so we do not consider them in our review (Cardoen et al., 2010; Guerriero and Guido, 2011; Jebali and Diabat, 2015; Min and Yih, 2010; Neyshabouri and Berg, 2017; Zhang et al., 2019). We further focus our analysis on recent SP, RO, and DRO approaches for the above two elective surgery classes published in the past decade (i.e., January/2010–November/2020). We do not include any of the extensive research on related deterministic optimization and simulation approaches to elective surgery scheduling or any work related to pre-operative preparation units and other upstream problems. Such analysis is beyond the scope of our focused analysis. Comprehensive surveys on deterministic and other approaches for OR and elective surgery scheduling include Cardoen et al. (2010), Samudra et al. (2016), and Zhu et al. (2019). Finally, we do not summarize any literature that focuses only on downstream or hospital capacity planning without considering surgery scheduling.

While our focused analysis of SO approaches to elective surgery scheduling with downstream capacity constraints does not provide a comprehensive view of the operations research challenges faced by OR managers, it does provide insights into the design of efficient and data-driven SO approaches for elective surgery scheduling and downstream capacity planning and their impact on the OR performance. Our goal is to identify the challenges and illustrate effective techniques for modeling and solving elective surgery scheduling problems under uncertainty and downstream capacity constraints. To date, no paper has analyzed existing SO approaches for advance and allocation scheduling of elective surgery under limited capacity of recovery units and uncertainty. More broadly, SO approaches to elective surgery scheduling can be used in other applications, within and outside healthcare operations, where multiple jobs share the same downstream resources.

The remainder of the paper is structured as follows. In Section 2, we discuss the impact of uncertainty on the elective surgery schedule, highlighting effective techniques to obtain granular data on random surgery duration and LOS in the downstream recovery units. In Section 3, we provide a gentle introduction to SP, RO, and DRO. In Section 4, we describe the main elective surgery scheduling problems OR managers face, provide and overview of existing SO approaches, and highlight the impact and challenges associated with these efforts. In addition, we provide recipes for future research opportunities. In Section 5, we present some future research opportunities and open questions. Finally, we draw conclusions in Section 6. Table 1 presents the main acronyms used throughout the paper.

2. Impact of uncertainty

Stochasticity is an intrinsic property of the advance and allocation surgery scheduling problems (Zhang et al., 2020). As pointed out by Shore (2020), Min and Yih (2010), and Shehadeh and Padman (2021), surgery duration and postoperative LOS in the downstream units are, by far, two of the primary (patient-related) sources of disruption to the daily schedule of surgery and patient flow in the downstream recovery units (other random factors observed in outpatient settings include patient no-show and patient arrival time, see Ahmadi-Javid et al. (2017) for a comprehensive survey). In what follows, we discuss the challenges associated with these random parameters as well as opportunities to model them.

2.1. Impact of uncertainty in surgery duration

Different types of surgeries require different surgery durations. Even surgeries of the same type have significant variability in their durations. Some studies link surgery duration to clinical factors (e.g., patient's medical history, surgery type, etc.), surgical team experience (Eijkemans et al., 2010; Molina-Pariente et al., 2015; Zheng et al., 2008), type of anesthesia (Strum et al., 2000), intraoperative complications, and surgery start time (Cassera et al., 2009; Peskun et al., 2012). Some recent statistical studies demonstrate that surgery duration depends on surgeon workload or surgery work-content such as the combination of surgery types, surgery sequence, priority constraints, degree of time-overlap between surgeries, etc (see, e.g., Shore (2020), Wang et al. (2018), and the references therein).

Table 1
Acronyms.

CVM	Composite variable modeling
DRO	Distributionally robust optimization
DR	Distributionally robust
EHR	Electronic health record
FJSS	Flexible job shop scheduling
HIT	Healthcare information technology
i.i.d	Independent and identically distributed
IT	Information technology
LP	Linear program/programming
LOS	length-of-stay
MDROM	Multi-stage DRO model
MILP	Mixed-integer linear program/programming
ML	Machine learning
OR	Operating room
PACU	Post-anesthesia care unit
PHU	Preoperative holding unit
RO	Robust optimization
RTLS	Real-time locating systems
SAA	Sample average approximation
SASS	Single-provider appointments sequencing and scheduling
SICU	Surgical intensive care unit
Select_Assign	Surgery selection and assignment problem
Seq_Sched	Sequencing and scheduling problem
SO	Stochastic optimization
SP	Stochastic program/programming
w.p.1	with probability one

Usually, practitioners use surgery duration as an input to make surgery allocation and scheduling decisions. Wang et al. (2018) find that allocation and scheduling decisions, in turn, influence surgery duration. Given the numerous factors impacting surgery duration, it is hard to predict in advance, particularly when the surgery is scheduled. Ignoring the uncertainty in surgery duration often contributes to unpredictable idle time and overtime for OR and surgery delays, among others. OR overtime is costly and may lead to surgery cancellation and thus compromised quality of care, and idle time implies poor utilization of OR capacity (Bartek et al., 2019; Fügener et al., 2014; May et al., 2011; Shehadeh et al., 2019). Several empirical studies report surgery delay (i.e., waiting time on the day of surgery) as an essential factor affecting patients' satisfaction and perceived quality of service and hence the reputation of the surgical suite (Kocas, 2015; Leiba et al., 2002; Osuna, 1985).

2.2. Impact of uncertainty in LOS

Postoperative LOS in the subsequent recovery resources such as SICU (days), inpatient wards (days), PACU (hours) is also random and hard to predict in advance (Bai et al., 2017; Strand et al., 2010; Zhang et al., 2019). LOS in SICU primarily depends on surgery type, anesthetic regimen, patient health status and intra- and post-operative factors and complications (Collins et al., 1999; De Hert et al., 2004; Toptas et al., 2018). Postoperative LOS in PACU varies depending upon surgery type, level of sedation, and response of individual patients after surgery.

Variability of patient LOS in the downstream recovery units leads to unpredictable availability of recovery beds and congestion. In the case of lack of SICU beds, OR manager may cancel surgeries, which incurs extra costs (~$1700–$2000 per case, Argo et al. (2009)) and impacts patient health, and/or transfers a patient from SICU to another hospital unit with a lower level of care to free a bed for a scheduled surgery. Several studies show that the rates of SICU readmission and risk of death of patients who leave SICU before they are considered fit for discharge are significantly higher than those who were discharged electively. Jonnalagadda et al. (2005) show that 15% of

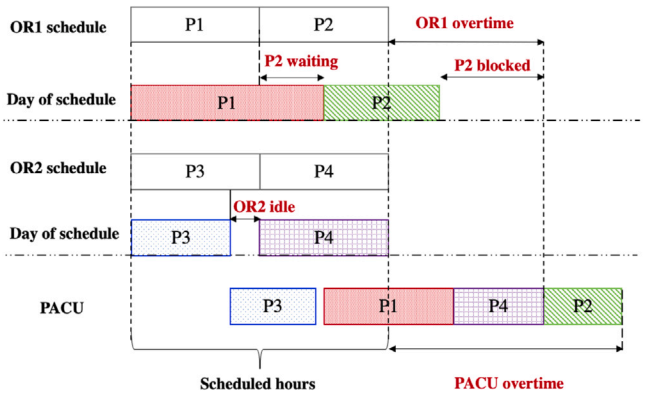


Fig. 2. Patient flows in a surgical suite with two ORs and one PACU bed. P1–P4 are scheduled surgeries.

surgery cancellations are due to the lack of recovery beds. Utzolino et al. (2010) state that the readmission rate to the SICU for patients with unplanned discharge from the SICU is ~3 times higher than those discharged electively (25.1% versus 8.3%). Goldfrad and Rowan (2000) and Utzolino et al. (2010) show that the mortality rate of patients who are readmitted to SICU is almost 6 times higher than those who are not readmitted.

PACU is also a well-known bottleneck for the overall operational performance of surgical suites. Insufficient PACU capacity may lead to OR blocking, in which case a patient is held in the OR after his/her surgery until a PACU bed becomes available. OR blocking is costly to all interested parties, since it may result in delays and cancellations of subsequent surgeries, reduced quality of care and overtime for both OR and PACU staff (Bai et al., 2017; Iser et al., 2008; Jonnalagadda et al., 2005; Lee and Yih, 2012, 2014). Fig. 2 illustrates a small example of patient flow and blocking in a surgical suite with two ORs and one PACU bed based on the partial schedule of each OR (Bai et al., 2017). If there is a surgery scheduled after P2 then OR1 overtime indicates the waiting time of this surgery.

PACU congestion is encountered in many, if not all hospitals (Bai et al., 2017). For example, Fairley et al. (2019) discusses how poor scheduling results in 20 hours of PACU blocking at a midsize academic pediatric hospital in California. These delays cost the hospital up to $44,000 in wasted OR time alone, which does not include overtime costs and the additional lost revenue from canceled surgeries due to OR blocking.

2.3. *Impact and challenges of multimodality*

There is extensive literature on surgery scheduling with random surgery durations (Cardoen et al., 2010; Batun, 2012; Burdett and Kozan, 2015; Denton et al., 2010; Berg et al., 2014; Gul et al., 2011, 2015; May et al., 2011; Shehadeh et al., 2019; Samudra et al., 2016). Most of these studies assume that surgery durations follow one (unimodal) fully known distribution, typically lognormal (Gul et al., 2011). Few studies consider the uncertainty of postoperative LOS in downstream recovery units, and these studies also assume that LOS follows one fully known distribution (Bai et al., 2018; Jebali and Diabat, 2015; Min and Yih, 2010; Zhang et al., 2019).

In reality, it is unlikely that decision-makers have enough data to infer the true distribution of these random parameters. Furthermore, given the diversity in patient characteristics and surgery contexts, the probability distributions of these random parameters may be multimodal and ambiguous. For example, based on data from Thorax Center Rotterdam of the Erasmus University Medical Center, Jebali and Diabat (2015) show that depending on the intraoperative complexity and surgery type, surgery durations may be short, medium, or long, and the distribution of short, medium, and long are different. Shehadeh et al. (2020a)'s study at a large medical center in Michigan suggests that colonoscopy durations are bimodal (i.e., may follow two distinct and ambiguous distributions). Similarly, depending on the intraoperative and postoperative complications, postoperative LOS could be different (e.g., short versus long), while conditioning on the scenario realization (e.g., simple versus complicated surgery), the expectation and distribution of LOS can also be different (Chen et al., 2020; Jebali and Diabat, 2015; Shehadeh et al., 2020a). When scheduling surgery, it is not possible to know whether or not a complication will occur. Such multi-modal ambiguity renders existing approaches sub-optimal or not applicable. To date, there is no stochastic optimization-based elective surgery scheduling model that has incorporated multi-modal distributional ambiguity of surgery duration and LOS in recovery units.

2.4. *Opportunities for modeling uncertainty*

There is a large body of literature, contributed by the information systems community, demonstrating how the use of information technology (IT) can improve performance within organizations (Dhillon, 2005; Mithas et al., 2012, 2016; Pang et al., 2014). An important IT in our study context is real-time locating systems (RTLS) that provide precise location-based information of people and items within a defined area in real or near-real-time (Boulos and Berry, 2012). RTLS consists of specialized fixed (radio-frequency identification, Bluetooth, Wifi, etc.) receivers or readers receiving wireless signals from small ID badges or tags attached to people or items to determine the location of tagged entities within a building or some other confined indoor or outdoor space. As such, RTLS enables granular data collection on random parameters, and therefore, have been employed in supply chains and retail industries (Angeles, 2005; Kim and Garrison, 2010), with very few applications in healthcare settings (Ebrahimzadeh et al., 2017; Kato-Lin and Padman, 2019; Lee and Chin, 2006; Lin and Padman, 2013; Oztekin et al., 2010).

In healthcare settings (including surgical suites), RTLS is mainly employed to track and identify objects and/or persons and has not been used to support, guide, or change operations and optimize surgery schedules. Data on random parameters are also documented in electronic health record systems (EHRs). However, standardized elective data repositories that harmonize and store data from disparate sources such as RTLS and EHR still do not exist for most major surgery types in most hospitals. Therefore, there is a need for more research on developing smart systems to meet this need (see Section 5.1). Such granular and integrated data can then be used to characterize the uncertainty of surgery durations and postoperative LOS in downstream units, their multimodality, and distributional ambiguity. Accordingly, researchers could build data-driven and distribution-free optimization-based approaches for surgery scheduling and downstream capacity planning that better mimic reality. This will not only lead to "smarter" surgery scheduling and capacity planning optimization approaches but will also elevate the value of IT in healthcare from just collecting, storing, retrieving, and sending information to guiding and changing operational activities. Researchers can start by collaborating with hospitals and companies that employ tracking systems in surgical suites.

Notably, Burdett and Kozan (2018) proposed a flexible job shop scheduling (FJSS) approach for scheduling healthcare activities in a hospital and used their results to motivate the need for investment in new IT systems that integrate the FJSS approach into appropriate hospital information and management system (HIMS). Such an integrated approach can inform the HIMS when every activity begins and ends and when every patient arrives and departs from the hospital's different units, which would enable the scheduling model to identify problems in advance and suggest better ways to operate.

3. A gentle introduction to stochastic optimization

Data uncertainty is ubiquitous in real-world optimization problems such as elective surgery scheduling. There are three main frameworks for optimization under uncertainty; stochastic programming (SP), robust optimization (RO), and distributionally robust optimization (DRO). SO models with *recourse*, particularly in a two-stage setting, have gained wide acceptance across SO application domains. Recourse models result when some (first-stage) decisions must be fixed before information relevant to uncertainty is available, while some (second-stage) decisions can be delayed until this information is available (Birge and Louveaux, 2011; Higle, 2005). The latter decisions are often called recourse decisions because they offer an opportunity to adjust to the received information about the uncertain data. For example, surgery assignments to surgical blocks are made before realizing uncertainty, then admission denial to the SICU or surgery cancellation, for example, are determined after observing LOS in SICU and SICU bed availability. In Section 4, we analyze existing SP, RO, and DRO models with recourse for elective surgery scheduling. Therefore, in this section, we briefly introduce and compare SP, RO, and DRO for completeness.

For illustrative purposes, we adopt the following notation of Rahimian and Mehrotra (2019). Let $x \in \mathcal{X} \subseteq \mathbb{R}^n$ represent the first-stage decisions, let $(\Xi, \mathcal{F})$ represents the underlying measurable space for a given space Ξ and σ-field $\mathcal{F}$ of Ξ, let $\xi : \Xi \mapsto \Omega \subseteq \mathbb{R}^d$ represents a vector of random parameters defined on a measurable space $(\Xi, \mathcal{F})$, $f(x, \xi)$: $\mathcal{X} \times \Xi \mapsto \mathbb{R}$ represents a random cost recourse function, and $g(x, \xi) : \mathcal{X} \in \Xi \mapsto \mathbb{R}^m$ represents a vector of random functions, i.e., $g(x, \cdot) := [g_1(x, \cdot), \dots, g_m(x, \cdot)]^\top$. Given this setup, Rahimian and Mehrotra (2019) define the general stochastic optimization (SO) problem as follows

$$\text{(SO)} \quad \inf_{x \in \mathcal{X}} \left\{ \mathcal{R}_{\mathbb{P}}\Big[f(x, \xi)\Big] \Big| \mathcal{R}_{\mathbb{P}}\Big[g(x, \xi)\Big] \leq 0 \right\} \tag{1}$$

where $\mathbb{P}$ is the known probability measure on $(\Xi, \mathcal{F})$, $\mathcal{R}_{\mathbb{P}} : \mathcal{Z} \mapsto \mathbb{R}$ is a componentwise real-valued functional under $\mathbb{P}$, and $\mathcal{Z}$ is a linear space of measurable functions on $(\Xi, \mathcal{F})$. As pointed out by Rahimian and Mehrotra (2019), *the functional* $\mathcal{R}_{\mathbb{P}}$ *accounts for quantifying the uncertainty in the outcomes of the decision, for a given fixed probability measure* $\mathbb{P}$.

Classical SP extends the linear optimization framework to minimize a risk measure (often the total expected cost) associated with the optimal *here-and-now* (first-stage) and *wait-and-see* (second-stage recourse) decisions under a *known probability distribution* $\mathbb{P}$ of random parameters. Mathematically, SP is a special case of the SO problem in (1) and has the following classical forms:

$$v = \inf_{x \in \mathcal{X}} \left\{ F := \mathbb{E}_{\mathbb{P}}\Big[f(x, \xi)\Big] \right\} \tag{2a}$$

$$\inf_{x \in \mathcal{X}} \left\{ f(x) \Big| \mathbb{E}_{\mathbb{P}}\Big[g(x, \xi)\Big] \leq 0 \right\} \tag{2b}$$

In (2a) and (2b), $\mathcal{R}_{\mathbb{P}}$ is the expected-value functional, i.e., $\mathcal{R}_{\mathbb{P}} := \mathbb{E}_{\mathbb{P}}[\cdot]$. While SP is a powerful modeling approach with nice convergence properties, it has two inherent shortcomings. First, SP suffers from the notorious curse of dimensionality and intractability. Indeed, the computation of expectation requires evaluating multi-dimensional integrals, which is, in general, intractable. Second, the assumption of full knowledge about the underlying probability distribution, which is usually estimated using limited sample data, might lead to disappointments when implementing the optimal SP decisions under the true distribution of uncertainty (or another sample drawn from the same population). This phenomenon is well-known as the optimizer's curse and is reminiscent of overfitting effects in statistical models (Esfahani and Kuhn, 2018; Smith and Winkler, 2006).

Given the difficulty in solving the SP models exactly, some studies resort to the Monte Carlo approximation approach. For illustrative purposes, we use the expectation model in (2a) to describe this approximation approach. In the Monte Carlo approach, we replace the distribution of ξ with a discrete empirical distribution based on N independent and identically distributed (i.i.d.) samples, then we solve the sample average approximation (SAA) formulation $v_N = \min \hat{F}_N := \sum_{n=1}^{N} N^{-1} f^n(x, \xi^n)$ instead of (2a). The sample average $\hat{F}_N$ is an unbiased estimator of the expected value $F := \mathbb{E}[f(x, \xi)]$ in (2a) (Shapiro, 2003; Mak et al., 1999). In addition, by the Law of Large Numbers and Shapiro (2003), we have $\hat{F}_N \to F$ with probability one (w.p.1) as $N \to \infty$ (Linderoth et al., 2006; Homem-de Mello and Bayraksan, 2014; Kleywegt et al., 2002). It follows that $v_N \to v$ w.p.1 as $N \to \infty$, i.e., the objective value of the SAA model converges to the optimal objective value of the SP model w.p.1 as the sample size $N \to \infty$.

If the SAA formulation has a large number of scenario-based constraints and variables or is a mixed-integer linear program (as in most elective surgery scheduling formulations), one would expect the computational effort and solution time to solve the SAA formulation to increase as the sample size increases. We refer the reader to (Linderoth et al., 2006; Homem-de Mello and Bayraksan, 2014; Kim et al., 2015a; Kleywegt et al., 2002; Shapiro et al., 2009; Shapiro and Homem-de Mello, 2000) and references therein for the technical details of SAA and detailed discussions on finding a sample size that provides a good trade-off between the computational effort required to solve the SAA formulation of the SP and the quality of approximation. It is worth mentioning that it is computationally prohibitive to determine the required sample size for a high-quality approximation of SP for most real-world stochastic optimization problems.

Classical RO approaches assume a complete ignorance about the probability distribution of uncertain parameters. Instead, RO assumes that uncertain parameters reside in a so-called "*uncertainty set*" $\mathcal{U} \subseteq \mathbb{R}^d$ of possible outcomes of ξ with some structure (e.g., ellipsoid or polyhedron, see, e.g., Bertsimas and Sim (2004), Ben-Tal et al. (2015), Soyster (1973) for a detailed discussion). In RO, optimization is based on the worst-case scenario occurring within the uncertainty set, which inevitably leads to over-conservatism and suboptimal decisions for other more-likely scenarios (Chen et al., 2020; Delage and Saif, 2018; Rahimian and Mehrotra, 2019; Thiele, 2010). Note that in SP, every single scenario from the distribution of uncertainty contributes to the optimal values. Thus, it is necessary to consider the complete scenario set in deriving an optimal solution to the SP model. In contrast, only the defined scenarios in the uncertainty set contribute to the RO model's optimal values. The following two RO models are special cases of the SO problem in (1)

$$\inf_{x \in \mathcal{X}} \sup_{\xi \in \mathcal{U}} f(x, \xi) \tag{3}$$

$$\inf_{x \in \mathcal{X}} \sup_{\xi \in \mathcal{U}} \left\{ f(x, \xi) \Big| \sup_{\xi \in \mathcal{U}} g(x, \xi) \leq 0 \right\} \tag{4}$$

DRO is another approach for modeling uncertainty. DRO offers a middle-ground approach instead of the black or white view of the SP and RO approaches about knowing the probability distribution of uncertain parameters. That is, in DRO, we assume that the distribution $\mathbb{P}_\xi$ of uncertain parameters belongs to a family of distributions referred to as the "*ambiguity set.*" Probability distributions residing in the ambiguity set $\mathcal{P}$ share specific parametric and statistical characteristics (e.g., mean value, variance, range, etc.) or are close-enough to a reference distribution. Mathematically, the ambiguous counterpart of the SO problem in (1) is as follows

$$\text{(DRO)} \quad \inf_{x \in \mathcal{X}} \sup_{\mathbb{P}_\xi \in \mathcal{P}} \left\{ \mathcal{R}_{\mathbb{P}_\xi}\Big[f(x, \xi)\Big] \Big| \sup_{\mathbb{P}_\xi \in \mathcal{P}} \mathcal{R}_{\mathbb{P}_\xi}\Big[g(x, \xi)\Big] \leq 0 \right\} \tag{5}$$

The DRO problem in (5) seeks to find decisions $x \in \mathcal{X}$ that minimizes the worst-case (maximum) of the functional $\mathcal{R}$ of the cost function $f(\cdot)$ among all probability distributions residing in the ambiguity set $\mathcal{P}$. Note that in DRO, the distribution $\mathbb{P}$ is a decision variable. The DRO counterpart of (2a) and (2b) are as follows

$$\inf_{x \in \mathcal{X}} \sup_{\mathbb{P}_\xi \in \mathcal{P}} \mathbb{E}_{\mathbb{P}_\xi}\Big[f(x, \xi)\Big] \tag{6a}$$

$$\inf_{x\in\mathcal{X}}\left\{ f(x)\middle| \sup_{\mathbb{P}_\xi\in\mathcal{P}} \mathbb{E}_{\mathbb{P}_\xi}\left[g(x,\xi)\right] \leq 0\right\} \tag{6b}$$

The ambiguity set must capture the true distribution with a high degree of certainty and be computationally manageable (i.e., allow for a tractable reformulation of the DRO model). There are many ways to construct the ambiguity set $\mathcal{P}$ and can be based on (1) empirical moments and their nearby regions (see, e.g., Delage and Ye (2010), Wagner (2008), Zhang et al. (2018)), and (2) statistical distances between a candidate distribution and a reference distribution, such as ϕ–divergence (Jiang and Guan, 2016), Wasserstein metric (Esfahani and Kuhn, 2018; Gao and Kleywegt, 2016), and norm-based distance (Jiang and Guan, 2018). We refer to Rahimian and Mehrotra (2019), Chen et al. (2020), Esfahani and Kuhn (2018) and the references therein for a detailed discussion. The tractability of DRO models primarily depends on the ambiguity set. Surprisingly, it turns out that DRO models with carefully designed ambiguity sets, in which the distribution of uncertain parameters is a decision variable, are often more tractable than their SP counterparts in many real-world applications (Delage and Saif, 2018; Rahimian and Mehrotra, 2019; Shehadeh and Padman, 2021; Wang et al., 2019; Wiesemann et al., 2014).

As pointed out by Rahimian and Mehrotra (2019), if $\mathcal{P}$ contains only the true distribution of ξ, DRO reduces to SP. If $\mathcal{P}$ consists of all probability distributions supported on $\mathcal{U}$, DRO reduces to RO. Therefore, $\mathcal{P}$ is a key ingredient in DRO that can put DRO between SP and RO. Consequently, DRO is less conservative than RO, which ignores all the distributional information about ξ, except for the support set $\mathcal{U}$.

4. Modeling elective surgery scheduling problems

Our review focuses on stochastic optimization (SO) approaches for elective surgery scheduling with downstream capacity constraints under the block scheduling framework. Under this scheduling framework, OR schedules are divided into multiple blocks of defined lengths (e.g., 4 h). Each block is dedicated to only one type of surgical specialty (e.g., urology, general, vascular, etc.). There can be multiple blocks of the same specialty during a cycle (e.g., a week) in the surgery schedule. Patients can be scheduled in any of the blocks dedicated to their surgery type during the planning horizon. Before operations begin, OR managers have to resolve two major questions:

1. *Selection and assignment.* How many surgeries to schedule in each block, considering the capacity of ORs and the shared downstream recovery resources? (advance scheduling)
2. *Sequencing and scheduling.* How to sequence the start times of surgeries in each block to maintain good operational performance in the ORs and subsequent recovery units? (allocation scheduling)

Typical performance metrics. Ideally, OR managers seek to schedule a subset of surgeries from the waiting list to maintain a good trade-off between the cost of performing and postponing surgeries, percentage of scheduled surgeries (i.e., access to surgery), OR idle and over time, recovery unit utilization, premature SICU transfers, OR blocking time, surgery cancellation and delay, OR and recovery units congestion, and clinical staff workload (see Cardoen et al. (2010), Bai et al. (2017), Min and Yih (2010), Neyshabouri and Berg (2017) for other metrics).

To date, there is no single stochastic optimization model that can answer these questions simultaneously. Collectively, answering these questions requires solving a large scale stochastic optimization problem with a large number of binary variables (for selection, assignment, sequencing, etc.), integer and continuous variables, numerous constraints, and many interrelated decisions. Therefore, researchers have instead employed a decomposition approach, breaking up the elective surgery scheduling problem into subproblems and then solving them sequentially. As such, the SO literature consists of approaches that answer the above questions separately. In the next subsections, we provide an overview of the existing SO approaches for each subproblem and highlight the impacts and challenges associated with these efforts. Additionally, we discuss and provide new directions for future research opportunities.

4.1. Surgery Selection and Assignment Problem (Select_Assign)

In this problem, studies focus on the decision process of selecting a subset of elective surgeries from a list of I waiting patients and assigning selected patients to B available surgery blocks. Each patient has a surgery type and can be assigned to any of the blocks dedicated to his/her surgery type during the planning horizon. Surgery durations are random and depend on surgery type. After surgery, a patient may need recovery in one or multiple of the following downstream resources (Fig. 1): (1) SICU for an uncertain number of consecutive days, (2) inpatient ward for an uncertain number of consecutive days, and (3) PACU for an uncertain number of consecutive hours or minutes. In some rare cases, patients may not require any critical care and are directly sent to a hospital ward or some patients may return to the SICU after being discharged (these rare pathways are represented by dashed lines in Fig. 1).

Associated with each patient, there is a cost $c_{i,b}$ of performing or delaying his/her surgery, which depends on his/her waiting time on the list and clinical priority (Zhang et al., 2019). The cost $c_{i,b'}$ of delaying surgery (i.e., postponing to the next planning horizon by assigning surgery to a dummy block b') is often assumed to be higher than performing surgery. Recovery units (SICU, PACU, and ward) have a fixed number of beds on each day $t \leq T$. Ward capacity is often much larger than the SICU and PACU capacities.

Next, we analyze the existing SO approaches for this particular elective surgery selection and assignment problem (Select_Assign). We first observe that, to the best of our knowledge, and according to the recent review of operating room planning and surgical case scheduling by Zhu et al. (2019), there are no SO models for Select_Assign with PACU capacity constraints (i.e., existing SO models assume fixed surgery selection and assignment decisions). Therefore, in what follows, we review and analyze the few existing two-stage SP, RO, and DRO approaches for Select_Assign with SICU and ward capacity constraints (see Table 2). To facilitate the discussion, we define the following notation. For all $i \in [I]$ and $b \in B \cup \{b'\}$, let the binary decision variable $x_{i,b}$ represent the assignment of surgery i to block b. We define feasible region $\mathcal{X}$ of variables x, which consist of constraints on surgery selection and assignment. For example, feasible region $\mathcal{X}$ in (7) is defined such that each surgery i is assigned to one of the blocks $B(i)$ dedicated to its type or the dummy block b' (i.e., postponed to the next planning period).

$$\mathcal{X} = \left\{ x : \begin{array}{l} \sum_{b\in B(i)\cup\{b'\}} x_{i,b} = 1, \ \forall i \in I \\ x_{i,b} \in \{0,1\}, \ \forall i \in I, b \in B \cup \{b'\} \end{array} \right\} \tag{7}$$

4.1.1. SP approaches for Select_Assign

SP Formulations. Existing two-stage SP models for elective Select_Assign with SICU/ward capacity constraints seek to find first-stage surgery assignment decision $x \in \mathcal{X}$ that minimizes the sum of the first-stage costs and the expected second-stage costs $Q(x,\xi)$ subject to uncertainty ξ with a known joint probability distribution $\mathbb{P}$, see, e.g., the SP in (8). The first-stage costs represent patient-related costs (which includes the cost of performing and postponing surgeries). The second-stage costs, $Q(x,\xi)$, often include one or more of the following: (1) OR overtime cost, (2) OR idle time cost, and (3) cost of insufficient capacity of downstream units. The random vector ξ represents a vector of realization of random parameters.

$$(\text{SP}) \quad \min_{x\in\mathcal{X}} \left\{ \sum_{i\in I} \sum_{b\in B\cup\{b'\}} c_{i,b}x_{i,b} + \mathbb{E}_{\mathbb{P}}[Q(x,\xi)] \right\} \tag{8}$$

Table 2
Stochastic optimization approaches for Select_Assign with limited SICU and ward capacity.

Article	Distributions of	Metrics	Model	Solution
	DT & LOS			Approach
Min and Yih (2010)[a]	assumed known	PT, OT	SP	SAA
Jebali and Diabat (2015)	assumed known	PT, OT, IT, SICU, Ward	SP	SAA
Jebali and Diabat (2017)	assumed known	PT, OT, SICU	SP-CC	SAA
Zhang et al. (2019)[b]	assumed known	PT, OT, SICU	SP+MDP	SAA
Zhang et al. (2020)[b]	assumed known	PT, OT	SP	CGBH
Neyshabouri and Berg (2017)	unknown[c]	PT, OT, SICU	RO	C&CG
Shehadeh and Padman (2021)	unknown*d*	PT, OT, IT, SICU	DRO	C&CG

Notation: DT is random duration, LOS is length-of-stay, PT is patient-related cost of performing and delaying surgery, OT is overtime, IT is idle time, SICU/Ward is cost of lack of SICU/Ward bed, SP is two-stage stochastic programming model, SP-CC is SP with chance-constrained, SAA is sample average approximation, RO is robust optimization, DRO is distributionally robust optimization, CGBH is column-generation-based heuristic, and C&CG is column-and-constraint generation.

[a] Model includes constraints that ensure the respect of SICU capacity.

[b] Consider the cost of keeping surgical blocks open.

Min and Yih (2010) proposed the first SP model for Select_Assign, which seeks to find surgery assignment decisions $x \in \mathcal{X}$ that minimize patient-related costs (performing and delaying surgery) and expected overtime cost (second-stage). Min and Yih (2010)'s SP incorporates the co-existing uncertainty of surgery durations and LOS and includes a hard constraint to ensure that the number of patients in the SICU will not exceed the number of SICU beds available each day.

Jebali and Diabat (2015) generalize the SP of Min and Yih (2010) by incorporating the uncertainty related to patient LOS in the general wards. In addition, Jebali and Diabat (2015)'s model incorporates the expected cost of OR idling and penalties for exceeding the regular SICU capacity into the objective function. As we show in Table 3, these generalizations lead to a larger model with a significantly larger number of variables and constraints. Jebali and Diabat (2017) propose a 2-stage chance-constrained SP that minimizes patient-related costs and expected operating room utilization costs (overtime and idle time) and penalty costs for exceeding SICU capacity.

Recently, Zhang et al. (2019) generalized Min and Yih (2010)'s approach by proposing a two-level optimization model that combines Markov decision process (MDP) and SP to minimize the total cost: cost incurred by maintaining open OR blocks, cost of performing and delaying surgeries, the expected OR overtime cost, and the cost of insufficient ICU capacity. At the first level, the MDP select the patients to schedule. At the second level, the SP assigns the selected patients to open OR blocks and compute the related costs. Other SPs for special cases of the general Select_Assign problem include Zhang et al. (2020) and references therein.

For illustrative purposes in Table 3, we present the respective ('approximate') sizes of three of the existing SPs for Select_Assign, in terms of the number of variables and constraints. Finally, we note that these SPs have complete recourse, i.e., $Q(x, \xi)$ is feasible for any feasible first-stage surgery assignments decisions $x \in \mathcal{X}$ (Birge and Louveaux, 2011).

SP Solution Approaches and Challenges. As we discussed in Section 3, there are two well-known difficulties in obtaining an (exact) optimal solution to the SP in (8). First, evaluating the value of $\mathbb{E}_{\mathbb{P}}[Q(x, \xi)]$ involves taking multi-dimensional integrals and solving a large number of similar integer programs. Second, both $\mathbb{E}_{\mathbb{P}}[Q(x, \xi)]$ and $Q(x, \xi)$ are often non-convex and discontinuous (Birge and Louveaux, 2011; Min and Yih, 2010). In view of these two difficulties, Min and Yih (2010), Jebali and Diabat (2015), and Zhang et al. (2019) resort to approximation solution approaches. In particular, these SP studies use the SAA approach to replace the continuous distributions of surgery durations and LOS with approximate discrete distributions by considering a sample of N randomly generated scenarios.

As detailed in Section 3, in SAA, the stochastic SP (8) is replaced by its deterministic MILP formulation as follows. First, a sample of N independent scenarios is generated (each scenario consists of a vector of realizations of surgery durations and LOS which are drawn independently from their distributions). Then, the sample average of the second-stage (i.e., $1/N \sum_{n=1}^{N} Q(x, n)$) is optimized using the generated sample. The sample size N is chosen based on a trade-off between the computational effort required to solve the resulting SAA-MILP and the quality of approximation of the expected value objective of the problem by its sample average (see Section 3).

Challenge 1: Large scale SAA-MILP formulations. As we show in Table 3, the size of the resulting MILP is a function of the number of surgeries, surgical blocks, and the sample size. In particular, as the sample size N grows, which is necessary for a high-quality approximation, the size of the MILP grows. It is well-known that an increase in MILP size suggests an increase in solution time for the linear programming (LP) relaxation of the MILP and, thus, the solution time via commercial solvers (Artigues et al., 2015; Keha et al., 2009; Klotz and Newman, 2013b). Therefore, it is computationally prohibitive to determine the required sample size for high-quality approximation of optimal schedules for large instances (in terms of I, T, and B) of Select_Assign. Min and Yih (2010), Jebali and Diabat (2015), Zhang et al. (2019), and more recently Shehadeh and Padman (2021) results confirm this hypothesis. Sophisticated and customized algorithms and decomposition methods (e.g., *L-shaped* and *cutting plane*-based algorithms (Birge and Louveaux, 2011)) may help in solving the SAA model. However, these methods often require complex parameter tuning, derivation of valid (lower or upper) bounding inequalities to improve and accelerate convergence, which requires access to support staff with optimization expertise that hospitals often do not have.

Challenge 2: Symmetry. Surgery selection and assignment problems are subject to a great deal of symmetry, i.e., given a feasible first-stage solution x, an equivalent solution can be generated by permuting the columns of $x \in \mathcal{X}$. For example, suppose that there are two blocks for cardiac surgery (blocks 4 and 5). It is possible to create a new schedule by moving all cardiac surgeries that were assigned to block 5 to block 4 and vice versa. This creates two equivalent solutions because the blocks are identical, and surgeries are of the same type (i.e., have common distributions of random parameters). In the same way that it allows multiple equivalent solutions, symmetry also allows multiple equivalent sub-problems in the branch-and-bound tree, potentially leading to larger trees and forcing a wasteful duplication of solution time and effort (Margot, 2010; Ostrowski et al., 2010). While there are many effective techniques to break the symmetry in various MILP problems, especially OR scheduling problems, unfortunately, there are no such techniques to Select_Assign that consider all the clinical and non-clinical aspects of this problem. Notably, Shehadeh and Padman (2021) proposed the first effective symmetry breaking inequalities, which break symmetries in the solution space of the first-stage surgery selection and assignment decisions. However, these inequalities do not

Table 3
Approximate sizes of three SP formulations of the selection and assignment problem (Select_Assig) with SICU and ward capacity constraints. Notation: I is number of surgery, B is number of surgical blocks, T is number of days, and N is number of scenarios of random parameters.

	Min and Yih (2010)[a]	Jebali and Diabat (2015)[b]	Zhang et al. (2019)
# Binary variables	$\mathcal{O}(I[B+T\ \ N])$	$\mathcal{O}(IT[B+2N])$	$\mathcal{O}(B[I+1]+ITN)$
# Continuous variables	$\mathcal{O}(NB)$	$\mathcal{O}(NT[3B+2])$	$\mathcal{O}(N[B+T])$
# First-stage constraints	$\mathcal{O}(2I+B)$	$\mathcal{O}(I+BT)$	$\mathcal{O}(I[2+B])$
# Second-stage constraints	$\mathcal{O}(N[B+BIT+T])$	$\mathcal{O}(N[3BT+2T+3TI+2I+5]$	$\mathcal{O}(N[B+BIT+T])$

[a] We set $\tau = T$ to simplify the analysis and comparison.

[b] We set $\tau = 0$ to simplify the analysis and comparison.

account for patient priority and urgency, provider preference, and other clinical and non-clinical aspects.

Challenge 3: Distributional Ambiguity. Even in a perfect world in which these SPs are easy to solve, the resulting schedule's quality may be questionable, given that the SP assumes that the distributions of random parameters are known with certainty. As we discussed in Section 2, in reality, it is challenging, if not impossible, to estimate the multi-modal distributions of surgery duration and LOS in downstream units for each type of surgery, especially with limited reliable data. Suppose we calibrate a SP for Select_Assign to a sample from a biased distribution. In this case, the resulting biased optimal surgery selection and assignment will have a disappointing performance when implemented under the true distribution or another sample from the biased distribution, which will negatively impact OR operational performance.

4.1.2. RO approaches for Select_Assign

As we discussed in Section 3, in RO, optimization is based on the worst-case scenario occurring within an uncertainty set of random parameters. Notably, Neyshabouri and Berg (2017) is the first and only RO approach for Select_Assign (see Table 2). Neyshabouri and Berg (2017) construct an uncertainty set based on the maximum positive deviations of surgery durations and LOS from their mean values. Then, they propose a two-stage RO model, where optimization is based on the maximum positive deviations from the mean values of random parameters within the uncertainty set. The first-stage of Neyshabouri and Berg (2017)'s model assigns surgeries to surgical blocks to minimize the cost of performing and delaying surgeries. In the second stage, the goal is to minimize the worst-case scenario for the overtime and denied SICU admission costs.

Neyshabouri and Berg (2017) use column-and-constraints generation (C&CG) to solve their RO model. As argued by Chen et al. (2020), Delage and Saif (2018), Rahimian and Mehrotra (2019), and Thiele (2010), by focusing the optimization on the worst-case scenario, classical RO models such as that of Neyshabouri and Berg (2017) may yield over-conservative and sub-optimal decisions for other more-likely scenarios and poor expected performance because it cannot capture the distributional information of uncertainty. Additionally, Neyshabouri and Berg (2017) show that it is challenging to solve some small instances of Select_Assign (e.g., 15 surgeries) within a reasonable time. Finally, Neyshabouri and Berg (2017) did not address the issue of symmetry.

4.1.3. DRO approaches for Select_Assign

As we discussed in Section 3, DRO has become an attractive approach for addressing optimization problems contaminated with uncertain data due to following three primary characteristics (Rahimian and Mehrotra, 2019; Wang et al., 2019). First, DRO alleviates the unrealistic assumption of the decision maker's complete knowledge of the probability distribution governing the uncertain surgery duration and LOS. Second, it is often more computationally tractable than their SP and RO counterparts (Delage and Saif, 2018; Rahimian and Mehrotra, 2019). Finally, DRO avoids the well-known over-conservatism and the poor expected performance of RO and allows for better utilization of the available data. The computational advantage of DRO further fosters its application in healthcare management, and we have seen many successful examples in healthcare scheduling (Deng et al., 2014; Jiang et al., 2017; Mak et al., 2014; Shehadeh et al., 2020a).

Despite the advantages of DRO, there are no data-driven DRO approaches for Select_Assign with downstream constraints that incorporate the multi-modality of surgery duration and LOS. Recently, Shehadeh and Padman (2021) proposed the first DRO counterpart of the SPs for Select_Assign presented in Table 3 with SICU capacity. Shehadeh and Padman (2021) assume that surgery duration and LOS are unimodal and construct a mean-support ambiguity set of all possible unimodal distributions of these random parameters. Accordingly, Shehadeh and Padman (2021) model seeks optimal surgery scheduling decisions to minimize the cost of performing and postponing surgeries and the worst-case expected costs associated with overtime and idle time of ORs and lack of SICU capacity (which causes premature discharges or transfers). To solve their model, Shehadeh and Padman (2021) proposed a computationally efficient C&CG method. By reformulating the OR block capacity (BC) recourse problem (which minimize OR idle time and overtime) as an equivalent linear program (instead of MILP as in Neyshabouri and Berg (2017)), Shehadeh and Padman (2021) eliminate the need for the addition of inequalities for the BC recourse problem into C&CG's master problem, which results in slower growth in the size of their master problem in each iteration and thus faster convergence as compared to Neyshabouri and Berg (2017) C&C

In addition, Shehadeh and Padman (2021) derive a new family of symmetry breaking inequalities, which break symmetries in the solution space of the first-stage surgery selection and assignment decisions. These inequities are independent of the method of modeling uncertainty and so are valid for any SP, RO, and deterministic formulation that employ the same first stage surgery selection and assignment decisions. With the intent of justifying a DRO approach, Shehadeh and Padman (2021) conduct extensive numerical experiments using the practice configuration presented by Min and Yih (2010) (a well-known benchmark instance of surgery scheduling with ICU capacity constraints). Shehadeh and Padman (2021) demonstrate that their DRO approach can solve large instances of the problem in a reasonable time and produce robust schedules that have superior operational performance under different distributions as compared to the challenging SP and RO approaches. For example, Shehadeh and Padman (2021) model can solve instances of 140 surgeries in less than 10 min, while the SP fails to solve such instances and terminate with large optimality gaps. The number of premature transfers from SICU of Shehadeh and Padman (2021) schedules are 40% less than that of SP schedules.

Shehadeh and Padman (2021)'s results and managerial implications also suggest that incorporating the cost of OR idle time leads to better OR time utilization and better access to surgery. Note that Neyshabouri and Berg (2017) as well as Min and Yih (2010), Zhang et al. (2019), and Zhang et al. (2020) did not consider the cost of OR idle time in their model. Ignoring the cost of idle time is a major limitation of these models because OR idle time is an essential metric of OR efficiency and utilization. Girotto et al. (2010) found that each hour of unused operative time in the OR costs $3,600 at the University of Rochester Medical Center. Childers and Maggard-Gibbons (2018) reports $37.45 as the mean cost for 1 min of OR time across California hospitals. Outpatient appointment and procedure scheduling literature also demonstrate the importance of including the cost of idle time schedule optimization (see, e.g., Berg et al. (2014), Jebali and Diabat (2015), Liu et al. (2019), Shehadeh et al. (2019)).

The mean-range ambiguity sets of Shehadeh and Padman (2021) do not incorporate the multi-modality of random parameters and are not data-driven. In addition, Shehadeh and Padman (2021) DRO model does not consider the ambiguity of LOS in the inpatient ward, the possibility of patients returning to the SICU after being discharged, and other quality of service performance metrics and risk attitudes of the OR stakeholders. Finally, Shehadeh and Padman (2021)'s symmetry breaking inequalities do not account for patient priority, provider preference, and other clinical and non-clinical aspects observed in the elective surgery practice.

4.1.4. *Opportunities: Recipes for data-driven distribution-free approaches*

Clearly, to be able to solve Select_Assign in practice, we need effective symmetry breaking techniques and data-driven distribution-free (DRO), and tractable models that incorporate the ambiguity and possible multimodality of random parameters. In addition, we need DRO models for Select_Assign with PACU constraints and models that consider the possibility of patients returning to the SICU or inpatient ward after being discharged from these units. As pointed out by Shehadeh et al. (2019), these data-driven DRO approaches should provide a good trade-off between "***tractability*** *(i.e., being able to solve problem instances of realistic sizes in an acceptable amount of time), and* ***implementability****, i.e., proposing approaches that can be easily translated into standard optimization software packages and decision support tools. Implementability in the above sense is necessary for an optimization-based decision support tool to gain wide adoption in healthcare, and other service-providing industries that do not have ongoing access to support staff with optimization expertise*."

Constructing data-driven, tractable, and multi-modal DRO approaches for Select_Assign is not impossible, given the recent advances in DRO (Rahimian and Mehrotra, 2019), numerous guidelines and recipes for formulating, reformulating, and solving difficult MILP problems in the literature (Brown and Dell, 2007; Conforti et al., 2014; Klotz and Newman, 2013b,a; Newman and Weiss, 2013), and the rich data that can be obtained from EHR, RTLS, and other data collection systems employed in surgical suites. Next, we provide a generic recipe for a DRO approach for Select_Assign, consisting of data-driven ambiguity sets, DRO models, and solution methods.

(1) Data-driven ambiguity sets. As we mentioned in Section 3, the ambiguity set $\mathcal{F}$ is a key ingredient of any DRO approach (Esfahani and Kuhn, 2018). Ideally, to solve Select_Assign, we need data-driven ambiguity sets $\mathcal{F}$ that (a) are rich enough to contain true data-generating and multi-modal distributions with high confidence, and at the same time, small enough to exclude pathological distributions which would incentivize overly conservative decisions (Esfahani and Kuhn, 2018), (b) are easy to parameterize from available data or based on clinical staff expertise (e.g., moment-based ambiguity sets), and (c) can facilitate a tractable reformulation of the DRO models as a structured mathematical program that can be solved efficiently with off-the-shelf optimization software or using efficient and implementable solution methods (to enable practical use of the models).

As we mentioned in Section 3, there are various ways to construct an ambiguity set, such as moment-based and statistical distance-based ambiguity set. Next, we describe an intuitive and novel approach to model the multi-modality of surgery duration and LOS (Chen et al., 2020). Assume that the probability distribution $\mathbb{P}$ of random parameter ξ is a mixture of S distinct distributions $\sum_{s=1}^{S} p_s\mathbb{P}_s$ with $\sum_{s=1}^{S} p_s = 1$, where each mixture component $\mathbb{P}_s$ is an ambiguous (unknown) distribution with support $\mathcal{U}_s$ and moments $\mathbb{E}_{\mathbb{P}_s}[\xi] = \mu_s$ and $\phi(\xi)$. As pointed out by Chen et al. (2020), the generalized moments characterized by convex function ϕ provide useful statistical characterizations of the uncertainty ξ, including (co)-variance and absolute deviation, among others. Accordingly, we consider the following ambiguity set

$$\mathcal{F} := \left\{ \mathbb{P} \in \mathcal{P}(\mathbb{R}^{I_\xi + I_v} \times [S]) : \begin{array}{ll} ((\tilde{\xi}, \tilde{v}), \tilde{s}) \in \mathbb{P} & \\ \mathbb{E}_{\mathbb{P}}[\tilde{\xi} | \tilde{s} = s] \in \mu_s & \forall s \in [S] \\ \mathbb{E}_{\mathbb{P}}[\tilde{v} | \tilde{s} = s] \leq \boldsymbol{\sigma}_s & \forall s \in [S] \\ \mathbb{P}[(\tilde{\xi}, \tilde{v}) \in Z_s | \tilde{s} = s] = 1 & \forall s \in [S] \\ \mathbb{P}[\tilde{s} = s] = p_s & \forall s \in [S] \end{array} \right\} \tag{9}$$

where $\mathcal{P}(\mathbb{R}^{I_\xi + I_v} \times [S])$ is the set of all distributions on $\mathbb{R}^{I_\xi + I_v} \times [S]$. Primary and auxiliary random variables $\tilde{\xi}$ and $\tilde{v}$ jointly reside in the support set $\mathcal{Z}_s := \{(\tilde{\xi}, \tilde{v}) | \xi \in \mathcal{U}_s, v \geq \phi(\xi)\}$, for different scenarios $s \in [S]$. Note that any distribution $\mathbb{P}$ in $\mathcal{F}$ can be written as $\sum_{s=1}^{S} p_s\mathbb{P}_s$.

Ambiguity set (9) is a special case of the so-called scenario-wise ambiguity set recently introduced by Chen et al. (2020). For different scenarios s, the random variable ξ could be different, while conditioning on the scenario realization, the expectation and distributions of ξ can also be different. For example, surgery durations could be bimodal, i.e., depending on the intraoperative complications, surgery durations may be short ($s = 1$) or long ($s = 2$), and the distributions of short, $\mathbb{P}_1$, and long, $\mathbb{P}_2$, surgeries are different. In this example, $\mathbb{P} = p\mathbb{P}_1 + (1 - p)\mathbb{P}_2$, with p=0 or $p = 1$ (i.e., p is a Bernoulli random variable). Unfortunately, when scheduling surgery, it is not known at that time whether a complication will occur or not during surgery, and so this ambiguity should be accounted for.

(2) DRO Select_Assign Models. Given data-driven ambiguity set $\mathcal{F}$, and depending on the risk attitude of OR manager, several DRO models for Select_Assign can be investigated. Specifically, as described in Jiang et al. (2017), for $x \in \mathcal{X}$, a DRO model may consider a risk measure ϱ of $Q(x, \xi)$, where

(i) a risk-neutral OR manager may choose $\varrho(Q(x, \xi)) = \mathbb{E}_{\mathbb{P}}[Q(x, \xi)]$.
(ii) a risk-averse OR manager may choose $\varrho(Q(x, \xi)) = \text{CVaR}_{1-\epsilon}(Q(x, \xi))$, i.e., the conditional value at risk (CVaR) of second-stage cost $Q(x, \xi)$ with $1 - \epsilon \in (0, 1)$ confidence.

Then, the DRO models impose a generic min–max DR objective in the form of (10a) and/or generic DR constraints in the form of (10b)

$$\min_{x \in \mathcal{X}} \Big(\sum_{i \in I} \sum_{b \in B \cup \{b'\}} c_{i,b} x_{i,b} + \sup_{\mathbb{P} \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}[Q(x, \xi)] \Big) \tag{10a}$$

$$\sup_{\mathbb{P} \in \mathcal{F}} \varrho(Q(x, \xi)) \leq \bar{\mathcal{M}} \tag{10b}$$

where $\bar{\mathcal{M}} \in \mathbb{R}$ represents a bounding threshold for the risk measure from above. Both DR objective (10a) and constraints (10b) protect the risk measure by hedging against all probability distributions in $\mathcal{F}$. The DR objective minimizes the first stage cost and the worst-case expectation of random second stage cost $Q(x, \xi)$ over a family of distributions that resides on $\mathcal{F}$. The CVaR constraints provide a safe guarantee on the performance metrics with high probabilities. For example, we can use (10b) to ensure that OR blocking is controlled under the threshold $\bar{\mathcal{M}}$ with the smallest possible probability being no less than $1 - \epsilon$, i.e., $\inf_{\mathbb{P} \in \mathcal{F}} \mathbb{P}\{\text{blocking} \leq \bar{\mathcal{M}}\} \geq 1 - \epsilon$, which will provide an appropriate upper bound guarantee on OR blocking. We refer the reader to Rahimian and Mehrotra (2019) for a detailed discussion of other forms of risk-averse DRO approaches.

One can use data-driven ambiguity set such as the one in (9) and build on the existing formulation we analyzed here to formulate and investigate DRO models for the Select_Assign problem. For example, Shehadeh et al. (2020a) proposed the first tractable and data-driven DRO-MILP upstream single-provider scheduling approach that considers the bimodality of colonoscopy duration as a function of the pre-procedure bowel preparation. Using data from a large academic medical center collected electronically, Shehadeh et al. (2020a) demonstrate that colonoscopy durations are bimodal, i.e., depending on the prep quality, they can follow two different probability distributions, one for those with adequate prep and the other for those with inadequate prep. Accordingly, Shehadeh et al. (2020a) define a distributionally robust outpatient colonoscopy scheduling problem that seeks optimal appointment sequence and schedule to minimize the worst-case weighted expected sum of patient waiting, provider idling, and provider overtime, where the worst-case is taken over an ambiguity set characterized through the known mean and support of the prep quality and durations. Shehadeh et al. (2020a) is a special case of ambiguity set in (9) with two distinct distributions with known mean and support. Shehadeh et al. (2020a) work (which generalizes that of Mak et al. (2014) and Jiang et al. (2017) by incorporating sequencing decisions and bimodality of service time) does not consider surgery selection decisions, does not incorporate the capacity of downstream units and uncertainty of postoperative LOS in these units, and is not applicable for scheduling in multiple-ORs with multiple downstream units.

(3) Solution methods. Given the two-stage characteristics of Select_Assign, it is natural to attempt to solve it using decomposition approaches such as cutting plane (Delage and Ye, 2010; Thiele et al., 2010; Ye, 1996) and column-and-constraints generation (Zeng and Zhao, 2013). These methods are implemented in a master-sub problem framework. The master problem is a relaxation of the DRO problem, so the tightness of the bounded provided by the master problem is a key to convergence efficiency. One can analyze the structural properties of the master and subproblem of the model and accordingly derive valid inequalities to strengthen the master problem and accelerate the convergence.

A broader and theoretically ambitious aim is to construct data-driven ambiguity sets that allow for deriving structured equivalent reformulations of DRO models that can be easily translated into standard optimization software packages, not requiring customized algorithmic development or tuning. Besides the exact algorithms, there is also an opportunity to design computationally efficient approximation algorithms to solve the DRO models (Goh and Sim, 2010). For example, one can solve a more restricted version of the problem called affinely adjustable robust counterpart (AARC). Tractable DRO heuristics like AARC with distributional ambiguity has not been investigated for this class of problems. Such heuristics are shown to be computationally efficient, and in some cases, also theoretically optimal. Thus, it has the potential to bring the best of both worlds: computational efficiency and solution quality (or precision). Researchers can explore potential technical and practical conditions of Select_Assign to design near-optimal heuristics and approximations. Comparing the efficacy of solving the DRO models directly, using decomposition methods, and using approximation methods will provide worthwhile theoretical and empirical contributions to computational operations research. Finally, novel and efficient symmetry breaking inequalities that consider patient priority, provider preference, and other clinical and non-clinical aspects of Select_Assign can improve models' solvability.

4.2. Surgery Sequencing and Scheduling Problem (Seq_Sched)

In practice, selected surgeries are initially assigned to a surgeon, date, and operating room several weeks or even months before their scheduled date (a result of solving the selection and assignment problem, for example). The actual scheduled start times on the day of surgery for these surgeries, however, are typically not set until a few days in advance. At this point, OR managers construct the final schedule (i.e., determine surgery sequence and the scheduled start times) and notify the patients when to report to the hospital (Ahmadi-Javid et al., 2017; Berg et al., 2014; Shehadeh et al., 2019; Zhu et al., 2019). In the next subsections, we discuss the art, challenges, and opportunities of formulating and solving Seq_Sched with SICU/ward capacity and PACU capacity.

4.2.1. SO approaches for Seq_Sched with SICU/ward capacity

For OR suites that perform different types of (heterogeneous) surgeries that require SICU recovery, surgery sequencing and scheduling in each block is crucial to maintain good operational efficiency and patient flow (e.g., minimize patients and surgical team waiting time). Given that each block is dedicated to a specific surgical team, the sequencing problem in each block resembles that of the single-provider stochastic appointments sequencing and scheduling (SASS) problem with random surgery duration. The standard time-based criteria to evaluate the schedule are provider overtime, idle time, and surgery waiting times (Ahmadi-Javid et al., 2017; Berg et al., 2014; Denton et al., 2007; Shehadeh et al., 2019).

SASS is a well-known complex stochastic combinatorial optimization problem, given the inherent implied stochastic sequencing problem that underlies assigning start times to each surgery. As such, SASS has received considerable attention in the operations research community (Berg et al., 2014; Denton et al., 2007, 2010; Gupta, 2007; Mancilla and Storer, 2012). Recently, Shehadeh et al. (2019) proposed a new SP for SASS and compared their model to other formulations in the literature both empirically and theoretically, demonstrating where significant improvements in performance can be gained with their model. Therefore, we refer the reader to Shehadeh et al. (2019) for a detailed analysis of these SPs (see also Table A.1 in Appendix).

There are also only a few DRO approaches for appointment scheduling, for which we refer the reader to the pioneering work by Jiang et al. (2017), Kong et al. (2013, 2015), Mak et al. (2014), and Zhang et al. (2017). All these papers report on optimized surgery start times, assuming a fixed sequence of surgeries. As we mentioned earlier, Shehadeh et al. (2020a) propose the first DRO model for SASS that optimizes both the sequencing and scheduling decisions. By reformulating this DRO model as a MILP, Shehadeh et al. (2020a) provide an implementable DRO-MILP tool to derive insights into outpatient colonoscopy scheduling, and SASS in general, with bimodal surgery durations.

4.2.2. SO approaches for Seq_Sched with PACU capacity

Practical challenges. The volume of patients sent from the OR to the PACU can vary greatly depending upon the types and durations of scheduled surgeries. Postoperative LOS in PACU varies depending upon surgery type, level of sedation, and the response of individual patients. With several surgeons performing surgeries at a high volume, PACU cannot release patients quickly enough since surgery durations are often much shorter than their consecutive recovery times. As we mentioned earlier, unavailable PACU beds may lead to OR blocking, in which case the patient is held in the operating room until space opens up in the PACU (see Fig. 2). OR blocking is costly since it contributes to delays and cancellation of subsequent surgeries, reduced quality of care, overtime for both OR and PACU staff, and, ultimately, a loss of revenue (Bai et al., 2017; Iser et al., 2008; Jonnalagadda et al., 2005).

Theoretical challenges. PACU capacity (i.e., number of beds) impacts surgery scheduling decisions (number of surgeries, sequencing, and start times), and surgery schedules will, in turn, influence the performance of the PACU (Bai et al., 2017). When PACU capacity is considered, ORs are linked with this shared downstream unit. Thus, Select_Assign and Seq_Sched in multiple-linked ORs may not be decomposable, more complicated than scheduling in multiple-independent ORs (as in Select_Assign and Seq_Sched with SICU/ward capacity), and continues to be an open challenge. The additional complexity stems from the need to consider the blocking issue and thus the precedence relationships between surgeries in the same and multiple blocks and track (in real-time) the random start and completion times of surgeries in OR as well as arrivals and departure times from PACU. Note that blocking is not clinically feasible for surgeries that require recovery in SICU because it is not possible to hold a patient in the OR for several days after surgery until a SICU bed becomes available.

Another complexity of Seq_Sched with PACU constraints stems from the need to optimize an "*adaptive*" PACU admission policy, i.e., a priority rule that determines the sequence of PACU admissions when two or more surgeries finish simultaneously. Finally, surgery selection and assignment problem and the sequencing problem (in multiple ORs and PACU) are subject to a great deal of symmetry.

Existing SO approaches. Seq_Sched with PACU is still an open challenge in the field of surgery scheduling and operations research. Ideally, OR managers want to design a schedule and PACU admission policy that minimizes (a) patient waiting time, (b) OR blocking time, (c) OR over and idle times, (d) PACU staff overtime, and (e) surgery cancellation. To the best of our knowledge, to date, there are only two published stochastic optimization approaches for the multiple-OR and PACU elective surgery Seq_Sched; the paper of Bai et al. (2017) and Bai et al. (2020) (see Table A.1 in Appendix for a summary of other approaches). Due to the complexity of determining the optimal sequence of surgeries, both Bai et al. (2017) and Bai et al. (2020) assume that the set of surgeries to operate on in each block is known. Accordingly, they focus on determining surgery start times that minimize the expectation of metrics (a)–(d) above.

Given the set of selected surgeries, Bai et al. (2017) propose a stochastic optimization model based on a Discrete Event Dynamic System to determine surgery start times that minimize the expected cost of patient waiting time, surgeon idle time, OR blocking time, OR overtime and PACU overtime. Bai et al. (2017) assume that the OR manager will employ the First-come-First served policy for PACU admission. In an extensive numerical study, Bai et al. (2017) demonstrate that their method identifies near-optimal solutions for small instances of the problem. However, this conclusion depends on the small Seq_Sched instances that they considered and thus not generalizable.

Bai et al. (2020) discretized the time horizon into a finite number of time intervals and formulated the elective Seq_Sched in multiple ORs with limited PACU capacity as a time-indexed mixed-integer SAA model. Given that the model was impossible to solve for small instances of the problem, Bai et al. (2020) propose to address this challenging stochastic combinatorial problem in two stages. In the first stage, Bai et al. (2020) propose a new surgery sequencing heuristic based on solving a surrogate optimization problem. The surrogate problem is a time-indexed SAA optimization model that considers duration uncertainty and the PACU capacity constraints, but with a different objective function ("surrogate objective") that is highly correlated with the original objective function. The surrogate problem aims to minimize the cost of patient waiting time, OR blocking time, OR overtime, and PACU overtime.

To obtain surgery sequence, Bai et al. (2020) solve the much easier surrogate model than the original SP model via Lagrangian relaxation. Then, given the surgery sequence from the first-stage, in the second stage, Bai et al. (2020) determine the associated scheduled start times via existing surgery scheduling methods, including those of Bai et al. (2017), Iser et al. (2008), and Marcon and Dexter (2007). Bai et al. (2020) numerical experiments show that their sequencing heuristic outperforms benchmark methods by 13% to 51%, or, equivalently, generates an average savings of \$760 to \$7420 per day in surgical suites with 4 to 10 ORs. However, Bai et al. (2020) also mention that their experiments are based on special instances of Seq_Sched with PACU constraints, and so it is not clear whether the method would produce near-optimal or sub-optimal solutions to other or larger instances of the problem.

Note that both Bai et al. (2017) and Bai et al. (2020) assume perfect information about the probability distribution of LOS in PACU. However, as we mentioned in Section 2, the distribution of surgery durations and LOS in PACU is often ambiguous and may be multi-modal, which render the fixed sequence SP approaches of Bai et al. (2017) and Bai et al. (2020) not applicable for handling ambiguity and multi-modality.

4.2.3. Opportunities

Clearly, stochastic Seq_Sched with PACU is not well-studied in the literature, and there are many open questions and challenges to tackle. First, we need data-driven models to characterize and analyze the distribution of the blocking problem. OR blocking may be impossible to eliminate entirely, especially when PACU capacity is tight. And so, by analyzing the distribution of blocking, we could identify an acceptable and realistic upper bound for blocking time. This will limit the scope of the optimization problem from eliminating blocking to controlling blocking under a realistic upper bound.

Second, there is a need to design optimal and implementable admission policy to the PACU. The First-Come-First-Served (FCFS) and Critical-Patient-First (CPF) are two of the well-known and easy-to-implement polices. However, no research has investigated the (sub-)optimality of these policies rigorously. Thus, we need data-driven, distribution-free models that can jointly optimize surgery scheduling decisions and PACU admission policy while providing a good trade-off between tractability and implementability. In theory, optimizing both the scheduling decisions and PACU admission policy requires formulating the problem as a multi-stage DRO model (MDROM) with risk-averse or risk-neutral objectives and/or constraints (or a multi-stage SP model if the distribution of uncertainty is known). The first stage of the MDROM pertains to deciding the number of patients to schedule, their assignments to surgical blocks, and the sequencing of their start times. In each stage (e.g., completion of one or multiple surgeries simultaneously, departure of a patient from PACU, etc.) after the first stage, the approach chooses the worst-case probability distribution from the ambiguity set under which the uncertain duration and LOS in PACU are observed and then optimizes a policy that grants admission to PACU, block a patient in OR until a bed is available, or cancel one of the remaining surgeries based on the information available up to (and including) the current stage. This process continues until the last stage is reached.

Given the difficulty in solving the existing two-stage SP models and the possible ambiguity of the number of MDROM stages, such MDROM approaches may be intractable. Next, we provide some additional techniques to formulate the MDROM and deal with some anticipated challenges:

- *Composite variable modeling (CVM)*. Bai et al. (2017) and Bai et al. (2020) models are computationally challenging to solve in part due to a large number of binary variables and complicated constraints. A composite variable is a binary variable that encompass multiple elemental decisions (Cohn, 2002). CVM is a well-known method to reduce the size, eliminate complicating constraints, and strengthen the LP relaxation of combinatorial optimization problems. For example, instead of using two sets of binary variables to represent the precedence and sequencing decisions or a TSP tour to represent surgery sequence, as in Bai et al. (2020) formulation, we can use one set of binary position assignment variables, which implicitly determine the precedence relationship between surgeries in each block. Keha et al. (2009) and Shehadeh et al. (2019) demonstrate that position assignment-based reformulations of precedence-based formulations for single machine (i.e., one provider or OR) stochastic sequencing problems without downstream PACU constraints improves LP relaxations and solvability of these problems.

- *Variable transformation.* As detailed in Section 4.1.3, Shehadeh and Padman (2021) used variable transformation to transform LOS in SICU (and thus start and completion time in SICU) to an arrival–departure process to–from SICU, which helped them to derive a tractable DR-MILP model. Recall that LOS in SICU is in the range of days, and LOS in PACU is in the range of hours. Thus, one can leverage similar variable transformation ideas as in Shehadeh and Padman (2021) (which is based on the work of Neyshabouri and Berg (2017)) to transform LOS in PACU to an arrival–departure process to–from PACU. Then, formulate a DRO or SP model for this arrival–departure process. This may eliminate the need for the challenging-to-solve large-scale, time-indexed formulation, which computes patients' actual start and completion times in OR and PACU (the arrival/departure times can be used to compute the time spent in these units).

- *Two-stage approximation*. The mathematically optimal admission policy to PACU resulting from solving the multi-stage model may not be implementable in practice. To design and test feasible and implementable distributionally robust PACU admission policies, one can start by evaluating the conditions under which some of the well-known, feasible, and easy-to-implement policies such as the First-Come-First-Serve (FCFS) and Critical-Patient-First (CPF) are (sub-)optimal. One approach is to implement each feasible policy in a two-stage DRO model, enforcing non-anticipativity and yielding upper bounds on the multi-stage DRO (minimization) model's optimal value. By relaxing the non-anticipativity of the multi-stage model, which may make it easier to solve, we can obtain a lower bound. By evaluating the gap between these bounds, we can obtain a better sense of which policy is near-optimal or provide a tighter upper bound on the mathematically optimal distributionally robust PACU admission policy. Recently, Shehadeh et al. (2020b) presented a similar two-stage approximation idea to derive the first near-optimal, easy-to-implement adaptive policy for the multi-stage resequencing and rescheduling problem of unpunctual arrivals in outpatient clinics.

Finally, from a theoretical perspective, it also worth investigating and comparing risk-averse and risk-neutral DRO (and other stochastic optimization) models for Seq_Sched with PACU capacity that incorporate multi-modal ambiguity sets of surgery durations and other random parameters that might be observed in outpatient procedure centers, such as no-shows and arrival times (Ahmadi-Javid et al., 2017). More broadly, Seq_Sched is an embedded sub-problem in an integrated model for elective surgery selection, assignment, and scheduling with PACU constraints. Thus, by designing efficient SO techniques to solve Seq_Sched, we can find efficient methods to solve an integrated model for elective surgery scheduling with downstream capacity constraints.

5. Future research

While great research efforts and a wide variety of studies have been reported over the past decades to improve elective surgery selection, assignment, sequencing, and scheduling with consideration of downstream capacity and uncertainty, much work is still needed to solve these problems in practice. A few critical research opportunities for the future are discussed below.

5.1. Elective surgery databases

The availability of granular data on patients, resources, and clinic staff and their movements during pre-operative, surgical and recovery activities is essential for developing data-driven surgery scheduling and capacity planning approaches that better mimic reality. As pointed out by Wilson and Doyle (2008), the safe, efficient, and coordinated passage of a patient through the surgical and downstream recovery begins long before the patient arrives at the hospital for surgery. The journey starts when a patient's health concern is recognized and a physician concurs that surgery is needed. From the time of seeking the first medical attention onward, multiple healthcare providers will collect hundreds of data elements and store them in various systems—electronic or paper, integrated or otherwise. Within this complex jungle of information repositories, some patient data will be redundant, some will be contradictory, some will be missing, some will pass from system to system, and some will reside only within a single database (Wilson and Doyle, 2008). Thus, when the data are integrated, pre-processed, and then stored in one database/system, it could be the key to creating new knowledge and evidence for practice innovation for OR stakeholders.

The first step toward developing integrated and standardized surgery databases is establishing a reliable and efficient data collection method with appropriate tools, technologies, and standards. There are numerous methods to collect healthcare data for research and hospital administrative purposes. For example, suppose we want to collect data on LOS in the hospital. As detailed in Sarkies et al. (2015), such data can be collected manually from ward-based resources such as nursing handover records, paper-based ward discharge/transfer records, paper-based inpatient medical records, direct observation by experienced personnel, etc. This is indeed a very time and effort-intensive collection method that is subject to various human errors. Retrospective data extraction from administrative reports, scanned medical records, and electronic patient management systems is another method to obtain data. While this approach has been extensively used in healthcare research as a gold standard approach, retrieving medical records and transforming them into research data is resource-intensive and requires exceptional knowledge of the medical context and research skills (Hogan and Wagner, 1997; Maresh et al., 1986; Wilton and Pennisi, 1994)

Modern health information technologies, tools, and devices offer alternatives to the aforementioned traditional data collection methods. For example, as mentioned earlier, RTLS provides precise location-based information of tagged entities (people, equipment, etc.) within a defined area in real-time. Thus, such health ITs enable granular data collection and often yield massive data on tagged entities that can be used to build data-driven optimization approaches. Unfortunately, a few health systems employ modern health ITs for recording and collecting health data in part due to the prohibitive cost of health IT, the potential need for additional trained staff to ensure both compliance and careful collection of data, insufficient research about its benefits and unproven return on investment, apprehension about change and philosophical opposition to IT, and data privacy issues. Therefore, more research is needed to show the (financial and health) benefits of employing modern health ITs in guiding, optimizing, and changing operational activities in surgical suits. Moreover, there is a need for new technologies that can efficiently pull out, harmonize, and store collected data from different resources in an accessible manner that does not need an advanced computer or IT skills. Government support and fund will encourage implementing new health IT in hospitals to collect data and conduct research.

Developing standardized surgical databases additionally requires more hospitals to adopt advanced health IT and collective efforts from hospitals (including managers and clinical staff) at the national level to ensure a standardized collection of data. Therefore, policies and strategies for developing standards-compliant surgical databases is an important research direction, and a pre-requisite to generalizable modeling and analytic studies. The data type may depend on the health system and surgical suite under study and the optimization objective/task. However, in most OR and surgery scheduling problems, we often need a combination of timestamps along different stages of the surgical and recovery processes (e.g., surgery start and end time, time in OR, admission and discharge times from post-operative recovery units, etc.), clinical data (e.g., medical history, surgery details, etc.), and real-time demand, supply, and capacity data (e.g., hourly ICU occupancy/bed capacity during the day, availability of clinical staff, availability of surgical supplies, etc.) to analyze and model key variables (e.g., surgery duration, LOS, clinical staff overtime in upstream and downstream units, patient waiting time on the day of surgery, idle time, number of blockings between pre-operative stages, number of transfers between post-operative stages, number of premature discharges from ICU, financial data associated with these tasks, and many others).

5.2. Exploiting the Power of Machine Learning (ML)

Modern ML methods are quite powerful at extracting information from large (and diverse) data sets and guiding decision-making through classification and prediction, among others. In particular, systems incorporating Artificial Intelligence (AI) and ML techniques are increasingly and successfully used to guide decision-making in the healthcare sector and answer clinically meaningful questions (Ghassemi et al., 2018). ML tools have also been employed to design data-driven uncertainty and ambiguity sets for SP, RO, and DRO (see Bertsimas et al. (2018), Rahimian and Mehrotra (2019), and references therein for a detailed discussion). For example, ML tools have been used for ranking and selection of the most important uncertain parameters to be included in the definition of the ambiguity set, obtaining tight estimations of the moment- and distance-based parameters of the ambiguity sets, etc (see, e.g, Bertsimas et al. (2018), Esfahani and Kuhn (2018), Rahimian and Mehrotra (2019) and references therein for a detailed discussion). Recently, Jia and Shen (2019) proposed a learning-enhanced Benders decomposition algorithm for solving two-stage stochastic programs with complete recourse based on finite samples of the uncertain parameters. This algorithm includes two phases: (1) sampling cuts and collecting information from training problems, and (2) solving testing problems with support vector machines (SVM) cut classifier. Jia and Shen (2019)'s results show that the SVM cut classifier works effectively for identifying valuable cuts and how their algorithm reduces the total solving time of all instances for different problems with various sizes and complexities.

Despite the plethora and potential power of ML tools and literature, to date, ML tools have not been employed to design data-driven ambiguity sets and distribution-free models for advanced and allocation surgery scheduling with downstream capacity constraints. Therefore, we see a huge opportunity to use historical clinical, pre-operative, and post-operative patient data in building ML-based prediction models for some statistical properties (e.g., mean, range, etc.) of LOS and surgery durations with tight confidence intervals (CI), for example. We can then use the predicted values and their CIs to construct and calibrate the ambiguity set and build the DRO models. Simply put, DRO is applied to deal with ambiguous probability distributions, and ML is used to construct and calibrate the ambiguity set of probability distributions of uncertain problem data. ML can also restrict the DRO model to a subset of important uncertain parameters and probability distribution, ensuring computational tractability (Guevara et al., 2020). The accuracy of ML-based predictions and thus the quality of ML-based ambiguity sets may be a major and yet theoretically interesting challenge that interested researchers may need to address.

A significant issue faced by researchers using ML to derive insights from EHR or other HIT is external validity or generalizability, i.e., the performance of learned models at sites other than the those that generate the data used for training the models (Callahan and Shah, 2017; Iezzoni, 1999). One popular opinion is that models not validated on data from one or multiple external sites are not useful and questionable. In other words, the community supporting this opinion claim that models lacking external validity have failed in their task. Such thinking is not necessarily true; for example, Callahan and Shah (2017) argue that a model with fewer variables is often considered more generalizable and more usable than one with more variables. Additionally, one of the benefits of ML is that a model can be trained with data from any local site and evaluated using data from that site itself (Callahan and Shah, 2017) . Therefore, if a model achieves satisfactory performance at one site with a given data set, it has achieved its purpose. So future research integrating ML and DRO for elective surgery scheduling should focus on sharing the model building steps and work flow to retrain a model at a new OR suite. Indeed, by learning an ML-DRO model using site-specific data and updating its parameters as the dataset changes over time will produce a model that is best suited for optimization at that site and dataset. The success of such an ML-DRO approach across range of OR suites will provide strong evidence that the model-building process is valid and can be replicated, i.e., provide a good starting point for other models (Callahan and Shah, 2017; Peck et al., 2013).

Another avenue for future research is to investigate the benefits of ML-DRO approaches that allow ML to dynamically update or tune the input parameters of the ambiguity set and optimization model. Within the context of OR and elective surgery scheduling, ML is mainly used to estimate input parameters for optimization. For example, Fairley et al. (2019) proposed an optimization and ML approach for sequencing operating room procedures to minimize PACU unavailability delays. First, Fairley et al. (2019) used ML (gradient tree boosting and classification and regression trees) to predict surgery duration and LOS in PACU. Second, they developed two sequential, deterministic integer programs that use the predicted inputs and other data to generate an optimal schedule. Finally, they used discrete event simulation to estimate the optimized schedule's performance in the presence of uncertainty in surgical and recovery durations.

5.3. The Multi-criteria and Cost Estimate Dilemma

In theory, real-world multi-criteria optimization problems such as resource-constrained elective surgery scheduling can be solved using multi-objective optimization (MOO) approaches such as interactive approach, hierarchical approach, or a weighted objective function (see, e,g., Ehrgott (2005), Hoogeveen (2005), and references therein). Most of the existing surgery scheduling models assume that decision-makers can articulate their objective. In reality, the decision maker's priorities and objectives are challenging to articulate and could vary over time. Besides, the decision-maker is rarely a unique individual, and often there is a group of people that make decisions.

Furthermore, the actual costs associated with some performance metrics are not always known, nor can be estimated. And some metrics are interrelated. For example, the cost of OR overtime can be approximated by the cost of keeping OR open, surgeon and OR staff overtime cost, among others. As mentioned earlier, OR overtime may lead to the cancellation of one or multiple surgeries, which incurs extra costs (e.g., ~$1700–$2000 per case as reported by Argo et al. (2009) and impacts patient health. Decision-makers cannot always predict or measure the impact of surgery cancellation on patient health and the associated cost. Additionally, the cost of 1 min of OR, SICU, PACU, and inpatient wards, the cost of care in these units, and the labor cost of OR and downstream unit staff vary widely and differ between private and government health facilities. The large number of payers (e.g., type of insurance) and payment options for surgical care (and other health care services) further complicate the issues. Finally, even if we have precise estimates of costs and definitions of metrics, scheduling metrics are often conflicting, and it is often challenging to arbitrate the conflict between them.

Given the above challenges, more work is needed to bridge the gap between MOO approaches, academic research on costs associated with OR and surgical practice, and the actual costs incurred in practice. Future research should also carefully design sensitivity analysis experiments that examine trade-offs between costs, utilization, and capacity.

5.4. Preoperative activities, unpunctuality, and no-show

Our analysis focused on SO approaches for downstream resource-constrained elective surgery selection, assignment, sequencing, and scheduling problems. In some surgical suites, especially those offering outpatient elective procedures, some upstream preoperative factors can impact surgery schedule. On the day of surgery, patients are typically sent first to a preoperative preparation or holding unit (PHU) before their surgery. Preoperative activities include, but not limited to, dressing for OR or gowning, completing forms, receiving medication, meeting with the anesthesiologist, initiation of IV infusions, etc.

Inpatients move directly to PHU from the inpatient wards. In contrast, outpatients first check-in, and then they are called back to PHU to start the preparation process for surgery.

PHU represents a bottleneck in surgical suites that perform short-duration elective surgeries (e.g., upper endoscopy) in high volumes and have tight PHU capacity (beds, nurses, etc.). Suppose a preparation bed is not available when the patient is supposed to start the preoperative activities, and it is not possible to prepare the patient in another hospital unit. In this case, patient preparation is delayed until a preparation bed becomes available. A delay in the start time of preparation activities may cause OR starvation if the surgical team and OR are available (i.e., the OR and surgical team remain idle until patient preparation is completed). Surgery sequence and scheduled times, surgery type, PHU capacity, and preparation time uncertainty are major factors contributing to PHU congestion.

Outpatients may arrive later than their scheduled time, delaying the start times of their preoperative activities and consequently the start of their surgeries (extremely late patients are often rescheduled). Patient lateness also leads to OR starvation and surgical team idling. No-show or patient absenteeism is another random factor that contributes to underutilization in outpatient surgical centers. While the uncertainty in patient arrival time, patient no-show, and preoperative preparation time, and preoperative unit capacity have been studied in the context of appointment scheduling (Ahmadi-Javid et al., 2017; Bai et al., 2020; Gul et al., 2011; Moosavi and Ebrahimnejad, 2018; Shehadeh et al., 2020b), none of the "few" existing stochastic optimization approaches considered the combined multimodality and distributional ambiguity of these preoperative factors with the downstream factors we analyzed in this paper. Therefore, there is a need for tractable, data-driven, and distribution-free holistic approaches for elective surgery scheduling considering preoperative preparation units, OR, and postoperative recovery units. Our general recipes for modeling uncertainty in this focused analysis of SO approaches for OR+downstream units offer useful directions for modeling and studying uncertainty associated with preoperative activities.

5.5. Integrated approaches

Developing tractable and implementable data-driven approaches that integrate surgery selection, assignment, sequencing, and scheduling considering OR and downstream capacity constraints could have tremendous practical significance (on OR and recovery units utilization and performance) and novel theoretical contributions. The primary challenge in developing an integrated model is to find the appropriate trade-off between (*i*) model scope and realism (how much does it approximate actual practice?), (*ii*) model solvability and tractability (how much time does it take to solve an instance of the integrated problem?), and (*iii*) model and solution implementability (can we translate them into easy-to-use decision support tools that do not require access to support staff with advanced optimization expertise?).

A number of papers in the operations research literature recognize the extent to which model formulation affects tractability and administer advice to this end. Some of the existing formulations often rely on big-M coefficients that take large values and undermine computational efficiency. Big-M is often used to relax some of the constraints or enforce certain conditions, among other uses (e.g., ensure that the waiting time of a patient is zero when the patient is not scheduled). Camm et al. (1990) provide specific guidance for practitioners interested in the ubiquitous big-M and provide many practical examples to demonstrate that almost always, there is a maximum theoretical and practical size for the big-M coefficients that can be derived using the structural properties of the problem. Such a smart choice of big-M value could improve the model solvability compared to using unnecessarily large values for this parameter, diminishing model tractability.

Brown and Dell (2007) provide various excellent examples of how to start formulating a model, focusing on commonly appearing integer linear programming structure, for example, (1) the use of big-M, (2) packing, covering, and partitioning constraints, (3) cardinality constraints, and (4) the use of the inclusive and exclusive "or" conditions. Trick (2005) discuss how some well-known reformulation techniques theoretically designed to improve the LP relaxations bound for integer programs do not necessarily apply in light of modern integer programming solvers, which already recognize standard tightening inequalities (Savelsbergh, 1994).

The models and studies we analyze here (and even the deterministic formulations that we did not analyze) represent a good start for developing a tractable and implementable approach, for example, using efficient generalization and reformulation techniques. Klotz and Newman (2013b) present suggestions for appropriate use of state-of-the-art optimizers and guidelines for the careful formulation, both of which can vastly improve performance. As pointed out by Newman and Weiss (2013), preprocessing an already-formulated model before passing it to the branch-and-bound (and other) algorithm can significantly improve an integer program's tractability. Some preprocessing techniques are embedded in solvers and could serve as a secondary screening mechanism on a well-formulated model.

5.6. Capacity planning during the periods of high demand

Decades of clinical and operations research-based studies have focused on developing surgery schedules that improve access to surgery and ensure the maximal utilization of critical resources such as intensive care units (ICU). ICUs are very expensive to operate and require skilled staff (Dobson et al., 2010; Kim et al., 2015b). Not surprisingly, even a fully integrated model may be sub-optimal and cause the ICU to fall short of meeting the existing and emerging demand due to unexpected events. For example, we recently saw a shortage in the available ICU beds during the novel Coronavirus (COVID-19) outbreak due to the associated surge in ICU demand. Thus, many surgical centers that were not prepared for such an unexpected disruption had to cancel elective surgeries to free ICU beds for COVID-19 patients. Therefore, it is important to develop policies for ICU bed and resource allocations and usage taking into account potential surge in demand via intelligent and robust admission and discharge decisions and optimization approaches.

During times of high demand, most hospitals are focused on delivering surgical and intensive care for critical and high-risk patients who most need them. However, when it comes to choosing among patients who can potentially benefit from such treatment, there do not appear to be easy answers, and hospital managers do a lot of manual shuffling and juggling of schedules to ensure that their patients have access to surgery and intensive care. Even in a perfect world where we can identify the surgical and ICU benefits at the individual patient level, and there is wide agreement on some objectives such as maximizing the expected number of survivors and patient outcomes, it is quite obvious that selecting high-risk patients for surgery and allocating ICU beds for them is not necessarily the right thing to do and might be unfair (Kim et al., 2015b). For example, if a patient may benefit significantly from surgery or intensive care, but is also likely to have long length-of-stay, then cost, benefits, utilization, and trade-offs amongst them have to be considered.

Simply put, making surgery selection and admission/discharge from ICU decisions during the period of high demand is a complex task that requires careful consideration of multiple conflicting objectives and not only the health conditions of current patients but also a collective assessment of operational requirements of all existing patients in the ICU as well as the expected demand for ICU in the near future. Therefore, as pointed out by Kim et al. (2015b) there is a tremendous need for *mathematical modeling and analysis to develop insights and policies that can be useful when making these complex decisions in practice, particularly* *under stressful conditions with high demand and limited resource capacity*. At the outset, it may seem that adding more ICU beds can improve patient access to surgery and thus increase OR utilization. However, many resources are needed, and these complicating constraints may make ICU expansion seem impossible. As such, there is also a need for models that can assist OR managers and hospital decision makers in evaluating the risks and benefits of ICU expansion.

Burdett and Kozan (2016) pointed out that there are few optimization-based approaches for studying hospital capacity. In fact, there is no single hospital capacity problem nor a standard recipe for optimizing hospital capacity under uncertainty. Recent notable and relevant mathematical programming-based hospital resource and case-mix capacity planning include Abdelaziz and Masmoudi (2012), Andersen et al. (2017), Burdett et al. (2017), Burdett and Kozan (2016), Ma and Demeulemeester (2013), Vanberkel et al. (2014), Yahia et al. (2016) and references therein. Burdett and Kozan (2016) proposed a holistic multi-criteria MILP formulation that includes several key units in hospitals such as the recovery wards, OR, intensive care units, and emergency departments. Using real-data from Princess Alexandra Hospital in Australia, they demonstrate how this deterministic approach can be used to study capacity issues, capacity expansion scenarios, utilization, and case-mix selection. However, Burdett and Kozan (2016) and the aforementioned studies did not consider the distributional ambiguity of random parameters in their models.

5.7. Decision support tools

In addition to theoretically modeling and efficiently solving the resource-constrained elective surgery selection, assignment, sequencing, and scheduling problems, there is a need to translate these approaches into usable decision support tools (i.e., a computer-based, user-driven, interactive systems) to assist OR managers to better plan for elective surgeries and manage their downstream recovery units. Besides the ability to easily and quickly generate surgery schedules under various parameter settings, these tools should allow OR managers to study and visualize patient flows (pathways) within each block, from OR to recovery units, and between recovery units. Moreover, OR managers should be able to evaluate the risks and rewards of extending the capacity of the downstream recovery units (Kahn, 2012). Recovery beds are costly and require skilled staff that may not be always available and are costly to hire. If the physical space and budget allow, adding more SICU beds, for example, may reduce SICU congestion and premature discharge but may lead to SICU underutilization. Therefore, OR managers need these data-driven decision support tools to study several capacity expansion scenarios and find the number of recovery beds that would provide a good trade-off between OR and SICU utilization. Shehadeh and Padman (2021) show how their DRO approach for elective surgery scheduling with limited SICU capacity can be used as a decision support tool to examine trade-offs between costs, utilization, and capacity.

Gartner and Padman (2017a) present a recipe for developing a unified framework and a digital workbench for the strategic, tactical, and operational hospital management plan driven by information technology and analytics. They propose a workbench (named E-HOSPITAL) that combines the three classical hierarchical decision-making levels in one integrated environment. At each level, several decision problems can be chosen. They present extensions of mathematical models from the literature and incorporated them into the digital platform. Multiple stakeholders can use the E-HOSPITAL in healthcare delivery settings and for pedagogical purposes on topics such as healthcare analytics, services management, and information systems.

6. Conclusions

Motivated by our research collaboration with a large health system in Pennsylvania that has provided insights into some current challenges in elective surgery scheduling, in this paper, we review the state-of-the-art SO approaches to (optimal) stochastic elective surgery scheduling with downstream capacity constraints within the block-booking framework. Specifically, we focus our review on recent SP, RO, and DRO approaches for the advance and allocation scheduling of two major classes of elective surgery scheduling problems: (1) elective surgery scheduling with intensive care unit and ward capacity constraints, and (2) elective surgery scheduling with post anesthesia unit capacity constraints. We focus our analysis on recent SP, RO, and DRO approaches for these elective surgery classes published in the past decade (2010–2020). Emergency surgeries are often treated in dedicated units in most hospitals, and so we do not consider them in our review.

We recognize that stochasticity of surgery duration and postoperative LOS is an intrinsic property of these two classes of elective surgery scheduling problems. We also discuss the impact of ignoring the ambiguity and possible multi-modality of the distribution of these random parameters. Surgery durations and postoperative length-of-stay (LOS) are at a different time scale in each class. Resource requirements and constraints are different in each scenario. Hence, the practical and theoretical challenges associated with surgery selection, sequencing, and scheduling are materially different for each class. Therefore, we describe the art of formulating and solving each class separately, provide an analysis of existing SO approaches and their challenges, and highlight areas of opportunity for developing tractable, implementable, and data-driven approaches. Our key findings are

- Compared to the other OR and healthcare scheduling and planning problems, there is a lack of data-driven and tractable SO approaches for stochastic elective surgery scheduling with downstream capacity constraints.
- To date, there is no single SO approach that simultaneously optimizes elective surgery selection and assignment (advance scheduling) and sequencing and scheduling (allocation scheduling) decisions. This is, in part, due to the fact that optimizing these decisions requires solving large scale stochastic, mixed-integer programming, and multi-criteria optimization problems. Therefore, researchers have instead employed a decomposition approach, breaking up the elective surgery scheduling problem into subproblems and then solving them sequentially.
- Within the limited SO literature that considers the stochasticity of the subsequent postoperative stay and the capacity of recovery units, most studies assume that surgery duration and postoperative LOS follow fully known unimodal distributions, typically lognormal. To date, there are no data-driven and distribution-free models for elective surgery scheduling and downstream capacity planning that model the ambiguity and possible multi-modality of the distribution of surgery durations and postoperative LOS.
- Elective surgery selection and assignment problems with downstream capacity constraints are subject to a great deal of symmetry. However, there are no computationally efficient symmetry breaking techniques that consider ‘all the clinical and non-clinical factors’ involved in delivering surgical care.
- Selection, assignment, sequencing, and scheduling of elective surgery scheduling with SICU and ward capacity constraints are more studied and relatively less complicated than elective surgery scheduling with PACU capacity constraints. There is only one DRO approach for a special case of the former problem, and only two SP approaches for special cases of the latter.

- The high theoretical complexity, intractability, and limited practical applicability has prevented stochastic optimization-based surgery scheduling models from being implemented in practice. We have motivated several recipes that could help develop tractable and distribution-free integrated models and algorithms such as CVM, variable transformation, reformulation, tighter formulation, valid and problem-based bounding inequalities, two-stage approximations, near-optimal approximations, and ambiguity sets that support structured formulations that can be solved directly and efficiently.

We discuss and provide new directions for future research opportunities by recognizing these challenges and gaps in the literature. Our main recommendation for future research includes recipes for (1) developing standardized and granular elective surgery databases using the recent advances in healthcare information technologies such as real-time locating systems, sensors and wearable devices, and electronic healthcare record systems, (2) designing data-driven ML- and optimization-based methods for modeling the uncertainty and multi-modality of random parameters, (3) developing data-driven and distribution-free, tractable, and implementable integrated elective surgery scheduling and downstream capacity planning approaches that can be easily embedded in decision support tools, and (4) developing methods for downstream capacity planning during periods of high demand.

Acknowledgment

We thank all researchers who have contributed significantly to the literature on such a class of complex combinatorial and multi-criteria stochastic resource-constrained scheduling problems. We are grateful to the anonymous reviewers for their insightful comments and suggestions that allowed us to improve the paper. Dr. Karmel S. Shehadeh dedicates her effort in this paper to every little dreamer in the whole world who has a dream so big and so exciting. Believe in your dreams and do whatever it takes to achieve them—the best is yet to come for you.

Appendix. Recent approaches to elective surgery scheduling under the block scheduling policy

See Table A.1.

Table A.1
Some of the recent approaches (published between 2010 and 2020) to elective surgery scheduling under the block scheduling policy were. Note: this list is not comprehensive and is provided for illustrative purposes only.

	DT	LOS	DWU	# OR	SAAP	Sequencing	Metrics	Model	Sol. Approach
Denton et al. (2007)	✓			S		✓	OT, IT, WT	SP (SMILP)	Heuristics
Mancilla and Storer (2012)	✓			S		✓	OT, IT, WT	SP (SMILP)	heuristics
Berg et al. (2014)	✓			S		✓	OT, IT, WT	SP (SMILP)	ED, B&B-PH, SH
Shehadeh et al. (2019)	✓			S		✓	OT, IT, WT	SP (SMILP)	CPLEX
Jiang et al. (2017)	✓			S			OT, IT, WT	DR-MILP	ED
Mak et al. (2014)	✓			S		✓	OT, IT, WT	DR MISOCP	CPLEX
Shehadeh et al. (2020a)	✓			S		✓	OT, IT, WT	DR-MILP	CPLEX
Liu et al. (2019)	✓	✓	Ward	S			OT, IT	MDP	DP
Min and Yih (2010)	✓	✓	SICU	M	✓		OT SP (SMILP)	SAA-CPLEX	
Jebali and Diabat (2015)	✓	✓	SICU, Ward	M	✓		OT, IT, ExSICU, ExWard	SP (SMILP)	SAA-CPLEX
Jebali and Diabat (2017)	✓	✓	SICU	M	✓		OT, IT, ExSICU	2-stage CCSP	SAA-CPLEX
Zhang et al. (2019)	✓	✓	SICU	M	✓		OT, ExSICU	SP (SMILP)	SAA-CPLEX
Zhang et al. (2020)	✓	✓	SICU	M	✓		OT	SP (SMILP)	SAA-CPLEX
Neyshabouri and Berg (2017)	✓	✓	SICU	M	✓		OT, ExSICU	RO	C&CG
Moosavi and Ebrahimnejad (2018)	✓	✓	SICU, Ward	M			OT, IT, WTB, ER, LT	RO	MIP-based LNS
Shehadeh and Padman (2021)	✓	✓	SICU	M	✓		OT, IT, ExSICU	DRO	tractable C&CG
Hsu et al. (2003)	Deter	Deter	PACU	M		✓	Makespan, OT	Deter–MILP	TBS-based heuristics
Pham and Klinkert (2008)	Deter	Deter	PACU	M		✓	Makespan	Deter–MILP	CPLEX
Lee and Yih (2014)	✓	✓	PACU	M			WTB, WTA, OR, IT	FJS-FL	GA-based heuristics
Bai et al. (2017)	✓	✓	PACU	M			WTB, OT, ORBT, OT PACU OT	SimulationOpt	SGD
Lee and Yih (2012)	✓	✓	PACU	M		✓	WTB, WTA, OT, IT	Simulation	GA–based heuristics
Bai et al. (2020)	✓	✓	PACU	M			WTB, OT, IT, ORBT PACU OT	LR and SG	LR and SG
Gul et al. (2011)	✓	✓	PACU	M		✓	WTB, WTA, OT	Simulation	GA–based heuristics
Saremi et al. (2013)	✓	✓	PACU	M		✓	WTB, CT, SC	Simulation	TBS–based heuristics
Ewen and Mönch (2014)	✓	✓	PACU	M		✓	WTB, OT, IT	Simulation	GA–based heuristics

Notation: DT is surgery duration, LOS is length-of-stay, # OR is number of ORs, Deter is deterministic, WTB is patient waiting time before surgery, WTA is patient waiting time after surgery for PACU bed, OT is overtime, IT is idle time, ER is earliness, LT is lateness, ExSICU is cost of exceeding SICU capacity, ExWard is cost of exceeding ward capacity, ORBT is OR blocking time, PACU OT is PACU overtime, CR is patient's completion time, SC number of surgery cancellations, SP is two-stage stochastic programming model, SMILP is stochastic mixed-integer linear program, DP is dynamic programming, MDP is Markov decision process, SAA is sample average approximation, MIP-based LNS is Mixed Integer Programming based Local Search Neighborhood, CCSP is chance-constrained SP, RO is robust optimization, DRO is distributionally robust optimization, FJS-FL is flexible job shop with fuzzy logic, CG-based heuristic is Column-generation-based heuristic, C&CG is column-and-constraint generation, GA is Genetic Algorithm, TBS is Tabu search, LR is Lagrangian relaxation, SG is Subgradient method, ED is exact decomposition methods, B&B-PH is B&B with progressive hedging, Heuris-based BDecomp is heuristic solution approach based on Benders' Decomp, SH is sequencing heuristics, SimulationOpt is simulation optimization, SGD is sample-gradient descent. [a] Some models include patient-related cost. [b] With surrogate objective for sequencing.

References

Abdelaziz, F.B., Masmoudi, M., 2012. A multiobjective stochastic program for hospital bed planning. J. Oper. Res. Soc. 63 (4), 530–538.

Ahmadi-Javid, A., Jalali, Z., Klassen, K.J., 2017. Outpatient appointment systems in healthcare: A review of optimization studies. European J. Oper. Res. 258 (1), 3–34.

Andersen, A.R., Nielsen, B.F., Reinhardt, L.B., 2017. Optimization of hospital ward resources with patient relocation using Markov chain modeling. European J. Oper. Res. 260 (3), 1152–1163.

Angeles, R., 2005. RFID Technologies: supply-chain applications and implementation issues. Inf. Syst. Manage. 22 (1), 51–65.

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

Artigues, C., Koné, O., Lopez, P., Mongeau, M., 2015. Mixed-integer linear programming formulations. In: Schwindt, C., Zimmermann, J. (Eds.), Handbook on Project Management and Scheduling, Vol. 1. Springer, pp. 17–41.

Bai, J., Fügener, A., Schoenfelder, J., Brunner, J.O., 2018. Operations research in intensive care unit management: a literature review. Health Care Manag. Sci. 21 (1), 1–24.

Bai, M., Storer, R.H., Tonkay, G.L., 2017. A sample gradient-based algorithm for a multiple-OR and PACU surgery scheduling problem. IISE Trans. 49 (4), 367–380.

Bai, M., Storer, R.H., Tonkay, G.L., 2020. Surgery sequencing coordination with recovery resource constraints. Available at SSRN 3653618.

Bartek, M.A., Saxena, R.C., Solomon, S., Fong, C.T., Behara, L.D., Venigandla, R., Velagapudi, K., Lang, J.D., Nair, B.G., 2019. Improving operating room efficiency: A machine learning approach to predict case-time duration. J. Am. Coll. Surg.

Batun, S., 2012. Scheduling multiple operating rooms under uncertainty. (Ph.D. thesis). University of Pittsburgh.

Ben-Tal, A., Den Hertog, D., Vial, J.-P., 2015. Deriving robust counterparts of nonlinear uncertain inequalities. Math. Program. 149 (1–2), 265–299.

Berg, B.P., Denton, B.T., Erdogan, S.A., Rohleder, T., Huschka, T., 2014. Optimal booking and scheduling in outpatient procedure centers. Comput. Oper. Res. 50, 24–37.

Bertsimas, D., Gupta, V., Kallus, N., 2018. Data-driven robust optimization. Math. Program. 167 (2), 235–292.

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

Birge, J.R., Louveaux, F., 2011. Introduction To Stochastic Programming. Springer Science & Business Media.

Boulos, M.N.K., Berry, G., 2012. Real-time locating systems (RTLS) in healthcare: a condensed primer. Int. J. Health Geogr. 11 (1), 25.

Bovim, T.R., Christiansen, M., Gullhav, A.N., Range, T.M., Hellemo, L., 2020. Stochastic master surgery scheduling. European J. Oper. Res.

Brilli, R.J., Spevetz, A., Branson, R.D., Campbell, G.M., Cohen, H., Dasta, J.F., Harvey, M.A., Kelley, M.A., Kelly, K.M., Rudis, M.I., et al., 2001. Critical care delivery in the intensive care unit: defining clinical roles and the best practice model. Crit. Care Med. 29 (10), 2007–2019.

Brown, G.G., Dell, R.F., 2007. Formulating integer linear programs: A rogues' gallery. Inf. Trans. Educ. 7 (2), 153–159.

Burdett, R.L., Kozan, E., 2015. Techniques to effectively buffer schedules in the face of uncertainties. Comput. Ind. Eng. 87, 16–29.

Burdett, R., Kozan, E., 2016. A multi-criteria approach for hospital capacity analysis. European J. Oper. Res. 255 (2), 505–521.

Burdett, R.L., Kozan, E., 2018. An integrated approach for scheduling health care activities in a hospital. European J. Oper. Res. 264 (2), 756–773.

Burdett, R.L., Kozan, E., Sinnott, M., Cook, D., Tian, Y.-C., 2017. A mixed integer linear programing approach to perform hospital capacity assessments. Expert Syst. Appl. 77, 170–188.

Callahan, A., Shah, N.H., 2017. Machine learning in healthcare. In: Key Advances in Clinical Informatics. Elsevier, pp. 279–291.

Camm, J.D., Raturi, A.S., Tsubakitani, S., 1990. Cutting big M down to size. Interfaces 20 (5), 61–66.

Cardoen, B., Demeulemeester, E., Beliën, J., 2010. Operating room planning and scheduling: A literature review. European J. Oper. Res. 201 (3), 921–932.

Cassera, M.A., Zheng, B., Martinec, D.V., Dunst, C.M., Swanström, L.L., 2009. Surgical time independently affected by surgical team size. Am. J. Surg. 198 (2), 216–222.

Chang, J.-H., Chen, K.-W., Chen, K.-B., Poon, K.-S., Liu, S.-K., 2014. Case review analysis of operating room decisions to cancel surgery. BMC Surg. 14 (1), 47.

Chen, Z., Sim, M., Xiong, P., 2020. Robust stochastic optimization made easy with rsome. Manage. Sci.

Childers, C.P., Maggard-Gibbons, M., 2018. Understanding costs of care in the operating room. JAMA Surg. 153 (4), e176233–e176233.

Cohn, A.E.M., 2002. Composite-variable modeling for large-scale problems in transportation and logistics. (Ph.D. thesis). Massachusetts Institute of Technology.

Collins, T.C., Daley, J., Henderson, W.H., Khuri, S.F., 1999. Risk factors for prolonged length of stay after major elective surgery. Ann. Surg. 230 (2), 251.

Conforti, M., Cornuéjols, G., Zambelli, G., et al., 2014. Integer Programming. Vol. 271, Springer.

De Hert, S.G., Van der Linden, P.J., Cromheecke, S., Meeus, R., Pieter, W., De Blier, I.G., Stockman, B.A., Rodrigus, I.E., 2004. Choice of primary anesthetic regimen can influence intensive care unit length of stay after coronary surgery with cardiopulmonary bypass. Anesthesiology: J. Am. Soc. Anesthesiol. 101 (1), 9–20.

Delage, E., Saif, A., 2018. The Value of Randomized Solutions in Mixed-Integer Distributionally Robust Optimization Problems. GERAD HEC Montréal.

Delage, E., Ye, Y., 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. Oper. Res. 58 (3), 595–612.

Deng, Y., Shen, S., 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., Huschka, T.R., 2010. Optimal allocation of surgery blocks to operating rooms under uncertainty. Oper. Res. 58 (4-part-1), 802–816.

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

Dhillon, G., 2005. Gaining benefits from IS/IT implementation: Interpretations from case studies. Int. J. Inf. Manage. 25 (6), 502–515.

Dobson, G., Lee, H.-H., Pinker, E., 2010. A model of ICU bumping. Oper. Res. 58 (6), 1564–1576.

Ebrahimzadeh, F., Nabovati, E., Hasibian, M.R., Eslami, S., 2017. Evaluation of the effects of radio-frequency identification technology on patient tracking in hospitals: A systematic review. J. Patient Saf.

Ehrgott, M., 2005. Multicriteria Optimization. Vol. 491, Springer Science & Business Media.

Eijkemans, M.J., Van Houdenhoven, M., Nguyen, T., Boersma, E., Steyerberg, E.W., Kazemier, G., 2010. Predicting the unpredictablea new prediction model for operating room times using individual characteristics and the surgeon's estimate. Anesthesiology: J. Am. Soc. Anesthesiol. 112 (1), 41–49.

Esfahani, P.M., Kuhn, D., 2018. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. Math. Program. 171 (1–2), 115–166.

Ewen, H., Mönch, L., 2014. A simulation-based framework to schedule surgeries in an eye hospital. IIE Trans. Healthc. Syst. Eng. 4 (4), 191–208.

Fairley, M., Scheinker, D., Brandeau, M.L., 2019. Improving the efficiency of the operating room environment with an optimization and machine learning model. Health Care Manag. Sci. 22 (4), 756–767.

Freeman, N.K., Melouk, S.H., Mittenthal, J., 2016. A scenario-based approach for operating theater scheduling under uncertainty. Manuf. Serv. Oper. Manag. 18 (2), 245–261.

Fügener, A., Hans, E.W., Kolisch, R., Kortbeek, N., Vanberkel, P.T., 2014. Master surgery scheduling with consideration of multiple downstream units. European J. Oper. Res. 239 (1), 227–236.

Gao, R., Kleywegt, A.J., 2016. Distributionally robust stochastic optimization with wasserstein distance. arXiv preprint arXiv:1604.02199.

Gartner, D., Padman, R., 2017a. E-HOSPITAL–a digital workbench for hospital operations and services planning using information technology and algebraic languages. Stud. Health Technol. Inform. 245, 84.

Gartner, D., Padman, R., 2017b. Mathematical programming and heuristics for patient scheduling in hospitals: A survey. In: Handbook of Research on Healthcare Administration and Management. IGI Global, pp. 627–645.

Gartner, D., Padman, R., 2019. Flexible hospital-wide elective patient scheduling. J. Oper. Res. Soc. 1–15.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A.L., Ranganath, R., 2018. Opportunities in machine learning for healthcare. arXiv preprint arXiv:1806.00388.

Girotto, J.A., Koltz, P.F., Drugas, G., 2010. Optimizing your operating room: or, why large, traditional hospitals don't work. Int. J. Surg. 8 (5), 359–367.

Goh, J., Sim, M., 2010. Distributionally robust optimization and its tractable approximations. Oper. Res. 58 (4-part-1), 902–917.

Goldfrad, C., Rowan, K., 2000. Consequences of discharges from intensive care at night. Lancet 355 (9210), 1138–1142.

Guerriero, F., Guido, R., 2011. Operational research in the management of the operating theatre: a survey. Health Care Manag. Sci. 14 (1), 89–114.

Guevara, E., Babonneau, F., Homem-de Mello, T., Moret, S., 2020. A machine learning and distributionally robust optimization framework for strategic energy planning under uncertainty. Appl. Energy 271, 115005.

Gul, S., Denton, B.T., Fowler, J.W., 2015. A progressive hedging approach for surgery planning under uncertainty. INFORMS J. Comput. 27 (4), 755–772.

Gul, S., Denton, B.T., Fowler, J.W., Huschka, T., 2011. Bi-criteria scheduling of surgical services for an outpatient procedure center. Prod. Oper. Manage. 20 (3), 406–417.

Gupta, D., 2007. Surgical suites' operations management. Prod. Oper. Manage. 16 (6), 689–700.

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

Halpern, N.A., Pastores, S.M., Thaler, H.T., Greenstein, R.J., 2007. Critical care medicine use and cost among Medicare beneficiaries 1995–2000: Major discrepancies between two United States federal Medicare databases. Crit. Care Med. 35 (3), 692–699.

Higle, J.L., 2005. Stochastic programming: Optimization when uncertainty matters. In: Emerging Theory, Methods, and Applications. Informs, pp. 30–53.

Hof, S., Fügener, A., Schoenfelder, J., Brunner, J.O., 2017. Case mix planning in hospitals: a review and future agenda. Health Care Manag. Sci. 20 (2), 207–220.

Hogan, W.R., Wagner, M.M., 1997. Accuracy of data in computer-based patient records. J. Am. Med. Infor. Assoc. 4 (5), 342–355.

Hoogeveen, H., 2005. Multicriteria scheduling. European J. Oper. Res. 167 (3), 592–623.

Hsu, V.N., De Matta, R., Lee, C.-Y., 2003. Scheduling patients in an ambulatory surgical center. Nav. Res. Logist. 50 (3), 218–238.

Iezzoni, L.I., 1999. Statistically derived predictive models: caveat emptor. J. Gen. Intern. Med. 14 (6), 388.

Iser, J.H., Denton, B.T., King, R.E., 2008. Heuristics for balancing operating room and post-anesthesia resources under uncertainty. In: 2008 Winter Simulation Conference. IEEE, pp. 1601–1608.

Jackson, R.L., 2002. The business of surgery. Health Manage. Technol. 23 (7), 20–22.

Jebali, A., Diabat, A., 2015. A stochastic model for operating room planning under capacity constraints. Int. J. Prod. Res. 53 (24), 7252–7270.

Jebali, A., Diabat, A., 2017. A chance-constrained operating room planning with elective and emergency cases under downstream capacity constraints. Comput. Ind. Eng. 114, 329–344.

Jia, H., Shen, S., 2019. Benders cut classification via support vector machines for solving two-stage stochastic programs. arXiv preprint arXiv:1906.05994.

Jiang, R., Guan, Y., 2016. Data-driven chance constrained stochastic program. Math. Program. 158 (1–2), 291–327.

Jiang, R., Guan, Y., 2018. Risk-averse two-stage stochastic program with distributional ambiguity. Oper. Res. 66 (5), 1390–1405.

Jiang, R., Shen, S., Zhang, Y., 2017. Integer programming approaches for appointment scheduling with random no-shows and service durations. Oper. Res. 65 (6), 1638–1656.

Jonnalagadda, R., Walrond, E., Hariharan, S., Walrond, M., Prasad, C., 2005. Evaluation of the reasons for cancellations and delays of surgical procedures in a developing country. Int. J. Clin. Pract. 59 (6), 716–720.

Kahn, J.M., 2012. The risks and rewards of expanding ICU capacity. Crit. Care 16 (5), 156.

Kato-Lin, Y.-C., Padman, R., 2019. Rfid technology-enabled Markov reward process for sequencing care coordination in ambulatory care: A case study. Int. J. Inf. Manage. 48, 12–21.

Keha, A.B., Khowala, K., Fowler, J.W., 2009. Mixed integer programming formulations for single machine scheduling problems. Comput. Ind. Eng. 56 (1), 357–367.

Kim, S.-H., Chan, C.W., Olivares, M., Escobar, G., 2015b. Icu admission control: An empirical study of capacity allocation and its implication for patient outcomes. Manage. Sci. 61 (1), 19–38.

Kim, S., Garrison, G., 2010. Understanding users' behaviors regarding supply chain technology: Determinants impacting the adoption and implementation of RFID technology in South Korea. Int. J. Inf. Manage. 30 (5), 388–398.

Kim, S., Pasupathy, R., Henderson, S.G., 2015a. A guide to sample average approximation. In: Handbook of Simulation Optimization. Springer, pp. 207–243.

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

Klotz, E., Newman, A.M., 2013a. Practical guidelines for solving difficult linear programs. Surv. Oper. Res. Manag. Sci. 18 (1–2), 1–17.

Klotz, E., Newman, A.M., 2013b. Practical guidelines for solving difficult mixed integer linear programs. Surv. Oper. Res. Manag. Sci. 18 (1–2), 18–32.

Kocas, C., 2015. An extension of osuna's model to observable queues. J. Math. Psych. 66, 53–58.

Kong, Q., Lee, C.-Y., Teo, C.-P., Zheng, Z., 2013. Scheduling arrivals to a stochastic service delivery system using copositive cones. Oper. Res. 61 (3), 711–726.

Kong, Q., Li, S., Liu, N., Teo, C.-P., Yan, Z., 2015. Appointment scheduling under schedule-dependent patient no-show behavior.

Lee, M.R., Chin, L.-P., 2006. Amalgamating RFID and wireless networks for clinical path management. In: 2006 IET International Conference on Wireless, Mobile and Multimedia Networks. IET, pp. 1–3.

Lee, S., Yih, Y., 2012. Surgery scheduling of multiple operating rooms under uncertainty and resource constraints of post-anesthesia care units. In: IIE Annual Conference. Proceedings. Institute of Industrial and Systems Engineers (IISE), p. 1.

Lee, S., Yih, Y., 2014. Reducing patient-flow delays in surgical suites through determining start-times of surgical cases. European J. Oper. Res. 238 (2), 620–629.

Leiba, A., Weiss, Y., Carroll, J.S., Benedek, P., Bar-dayan, Y., 2002. Waiting time is a major predictor of patient satisfaction in a primary military clinic. Mil. Med. 167 (10), 842–845.

Li, F., Gupta, D., Potthoff, S., 2016. Improving operating room schedules. Health Care Manag. Sci. 19 (3), 261–278.

Lin, Y.-C., Padman, R., 2013. Process visibility analysis in ambulatory care: A simulation study with RFID data.. Stud. Health Technol. Inform. 192, 768–772.

Linderoth, J., Shapiro, A., Wright, S., 2006. The empirical behavior of sampling methods for stochastic programming. Ann. Oper. Res. 142 (1), 215–241.

Liu, N., Truong, V.-A., Wang, X., Anderson, B.R., 2019. Integrated scheduling and capacity planning with considerations for patients' length-of-stays. Prod. Oper. Manage..

Ma, G., Demeulemeester, E., 2013. A multilevel integrative approach to hospital case mix and capacity planning. Comput. Oper. Res. 40 (9), 2198–2207.

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

Magerlein, J.M., Martin, J.B., 1978. Surgical demand scheduling: A review. Health Serv. Res. 13 (4), 418.

Mak, W.-K., Morton, D.P., Wood, R.K., 1999. Monte Carlo bounding techniques for determining solution quality in stochastic programs. Oper. Res. Lett. 24 (1–2), 47–56.

Mak, H.-Y., Rong, Y., Zhang, J., 2014. Appointment scheduling with limited distributional information. Manage. Sci. 61 (2), 316–334.

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

Marcon, E., Dexter, F., 2007. An observational study of surgeons' sequencing of cases and its impact on postanesthesia care unit and holding area staffing requirements at hospitals. Anesth. Analg. 105 (1), 119–126.

Maresh, M., Dawson, A., Beard, R., 1986. Assessment of an on-line computerized perinatal data collection and information system. BJOG: Inter. J. Obstet. Gynaecol. 93 (12), 1239–1245.

Margot, F., 2010. Symmetry in integer linear programming. In: 50 Years of Integer Programming 1958-2008. Springer, pp. 647–686.

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

Homem-de Mello, T., Bayraksan, G., 2014. Monte Carlo sampling-based methods for stochastic optimization. Surv. Oper. Res. Manag. Sci. 19 (1), 56–85.

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

Mithas, S., Krishnan, M.S., Fornell, C., 2016. Research note—Information technology, customer satisfaction, and profit: Theory and evidence. Inf. Syst. Res. 27 (1), 166–181.

Mithas, S., Tafti, A., Bardhan, I., Goh, J.M., 2012. Information technology and firm profitability: mechanisms and empirical evidence. MIS Q. 205–224.

Molina-Pariente, J.M., Fernandez-Viagas, V., Framinan, J.M., 2015. Integrated operating room planning and scheduling problem with assistant surgeon dependent surgery durations. Comput. Ind. Eng. 82, 8–20.

Moosavi, A., Ebrahimnejad, S., 2018. Scheduling of elective patients considering upstream and downstream units and emergency demand using robust optimization. Comput. Ind. Eng. 120, 216–233.

Newman, A.M., Weiss, M., 2013. A survey of linear and mixed-integer optimization tutorials. Inf. Trans. Educ. 14 (1), 26–38.

Neyshabouri, S., Berg, B.P., 2017. Two-stage robust optimization approach to elective surgery and downstream capacity planning. European J. Oper. Res. 260 (1), 21–40.

Ostrowski, J., Anjos, M.F., Vannelli, A., 2010. Symmetry in Scheduling Problems. Citeseer.

Osuna, E.E., 1985. The psychological cost of waiting. J. Math. Psych. 29 (1), 82–105.

Ouyang, H., Argon, N.T., Ziya, S., 2020. Allocation of intensive care unit beds in periods of high demand. Oper. Res. 68 (2), 591–608.

Oztekin, A., Pajouh, F.M., Delen, D., Swim, L.K., 2010. An RFID network design methodology for asset tracking in healthcare. Decis. Support Syst. 49 (1), 100–109.

Pang, M.-S., Tafti, A., Krishnan, M.S., 2014. Information technology and administrative efficiency in US state governments. MIS Q. 38 (4), 1079–1102.

Peck, J.S., Gaehde, S.A., Nightingale, D.J., Gelman, D.Y., Huckins, D.S., Lemons, M.F., Dickson, E.W., Benneyan, J.C., 2013. Generalizability of a simple approach for predicting hospital admission from an emergency department. Acad. Emerg. Med. 20 (11), 1156–1163.

Peskun, C., Walmsley, D., Waddell, J., Schemitsch, E., 2012. Effect of surgeon fatigue on hip and knee arthroplasty. Can. J. Surg. 55 (2), 81.

Pham, D.-N., Klinkert, A., 2008. Surgical case scheduling as a generalized job shop scheduling problem. European J. Oper. Res. 185 (3), 1011–1025.

Pinedo, M.L., 2016. Scheduling: Theory, Algorithms, and Systems. Springer.

Rahimian, H., Mehrotra, S., 2019. Distributionally robust optimization: A review. arXiv preprint arXiv:1908.05659.

Reis Miranda, D., Jegers, M., 2012. Monitoring costs in the ICU: a search for a pertinent methodology. Acta Anaesthesiol. Scand. 56 (9), 1104–1113.

Ruiz, R., Vázquez-Rodríguez, J.A., 2010. The hybrid flow shop scheduling problem. European J. Oper. Res. 205 (1), 1–18.

Samudra, M., Van Riet, C., Demeulemeester, E., Cardoen, B., Vansteenkiste, N., Rademakers, F.E., 2016. Scheduling operating rooms: achievements, challenges and pitfalls. J. Sched. 19 (5), 493–525.

Saremi, A., Jula, P., ElMekkawy, T., Wang, G.G., 2013. Appointment scheduling of outpatient surgical services in a multistage operating room department. Int. J. Prod. Econ. 141 (2), 646–658.

Sarkies, M.N., Bowles, K.-A., Skinner, E., Mitchell, D., Haas, R., Ho, M., Salter, K., May, K., Markham, D., O'Brien, L., et al., 2015. Data collection methods in health services research: hospital length of stay and discharge destination. Appl. Clin. Inform. 6 (1), 96.

Savelsbergh, M.W., 1994. Preprocessing and probing techniques for mixed integer programming problems. ORSA J. Comput. 6 (4), 445–454.

Shapiro, A., 2003. Monte Carlo sampling approach to stochastic programming. In: ESAIM: Proceedings. Vol. 13, EDP Sciences, pp. 65–73.

Shapiro, A., Dentcheva, D., Ruszczyński, A., 2009. Lectures on stochastic programming: modeling and theory. SIAM.

Shapiro, A., Homem-de Mello, T., 2000. On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs. SIAM J. Optim. 11 (1), 70–86.

Shehadeh, K.S., Cohn, A.E., Epelman, M.A., 2019. Analysis of models for the stochastic outpatient procedure scheduling problem. European J. Oper. Res. 279 (3), 721–731.

Shehadeh, K.S., Cohn, A.E., Jiang, R., 2020a. A distributionally robust optimization approach for outpatient colonoscopy scheduling. European J. Oper. Res. 283 (2), 549–561.

Shehadeh, K.S., Cohn, A.E., Jiang, R., 2020b. Using stochastic programming to solve an outpatient appointment scheduling problem with random service and arrival times. Nav. Res. Logist..

Shehadeh, K.S., Padman, R., 2021. A distributionally robust optimization approach for stochastic elective surgery scheduling with limited intensive care unit capacity. European J. Oper. Res. 290 (3), 901–913.

Shore, H., 2020. An explanatory bi-variate model for surgery-duration and its empirical validation. Commun. Stat.: Case Studies, Data Anal. Appl. 1–25.

Smith, J.E., Winkler, R.L., 2006. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. Manage. Sci. 52 (3), 311–322.

Soyster, A.L., 1973. Convex programming with set-inclusive constraints and applications to inexact linear programming. Oper. Res. 21 (5), 1154–1157.

Strand, K., Walther, S.M., Reinikainen, M., Ala-Kokko, T., Nolin, T., Martner, J., Mussalo, P., Sø reide, E., Flaatten, H.K., 2010. Variations in the length of stay of intensive care unit nonsurvivors in three scandinavian countries. Crit. Care 14 (5), R175.

Strum, D.P., Sampson, A.R., May, J.H., Vargas, L.G., 2000. Surgeon and type of anesthesia predict variability in surgical procedure times. Anesthesiology: J. Am. Soc. Anesthesiol. 92 (5), 1454–1466.

Thiele, A., 2010. A note on issues of over-conservatism in robust optimization with cost uncertainty. Optim. 59 (7), 1033–1040.

Thiele, A., Terry, T., Epelman, M., 2010. Robust linear optimization with recourse. University of michigan. IOE Technical Report TR09-01.

Toptas, M., Sengul Samanci, N., Akkoc, I., Yucetas, E., Cebeci, E., Sen, O., Can, M.M., Ozturk, S., 2018. Factors affecting the length of stay in the intensive care unit: our clinical experience. BioMed Res. Int. 2018.

Trick, M., 2005. Formulations and reformulations in integer programming. In: International Conference on Integration of Artificial Intelligence (AI) and Operations Research (OR) Techniques in Constraint Programming. Springer, pp. 366–379.

Utzolino, S., Kaffarnik, M., Keck, T., Berlet, M., Hopt, U.T., 2010. Unplanned discharges from a surgical intensive care unit: readmissions and mortality. J. Crit. Care 25 (3), 375–381.

Vanberkel, P.T., Boucherie, R.J., Hans, E.W., Hurink, J.L., 2014. Optimizing the strategic patient mix combining queueing theory and dynamic programming. Comput. Oper. Res. 43, 271–279.

Varmazyar, M., Akhavan-Tabatabaei, R., Salmasi, N., Modarres, M., 2020. Operating room scheduling problem under uncertainty: Application of continuous phase-type distributions. IISE Trans. 52 (2), 216–235.

Viapiano, J., Ward, D.S., 2000. Operating room utilization: the need for data. Inter. Anesthesiol. Clin. 38 (4), 127–140.

Wagner, M.R., 2008. Stochastic 0–1 linear programming under limited distributional information. Oper. Res. Lett. 36 (2), 150–156.

Wang, J., Cabrera, J., Tsui, K.-L., Guo, H., Bakker, M., Kostis, J.B., et al., 2018. Clinical and non-clinical effects on surgery duration: statistical modeling and analysis. arXiv preprint arXiv:1801.04110.

Wang, Y., Tang, J., Pan, Z., Yan, C., 2015. Particle swarm optimization-based planning and scheduling for a laminar-flow operating room with downstream resources. Soft Comput. 19 (10), 2913–2926.

Wang, Y., Zhang, Y., Tang, J., 2019. A distributionally robust optimization approach for surgery block allocation. European J. Oper. Res. 273 (2), 740–753.

Wiesemann, W., Kuhn, D., Sim, M., 2014. Distributionally robust convex optimization. Oper. Res. 62 (6), 1358–1376.

Wilson, M.L., Doyle, C., 2008. Integration of ORMS and AIMS. In: Anesthesia Informatics. Springer, pp. 345–360.

Wilton, R., Pennisi, A.J., 1994. Evaluating the accuracy of transcribed computer-stored immunization data. Pediatr. 94 (6), 902–906.

Yahia, Z., Eltawil, A.B., Harraz, N.A., 2016. The operating room case-mix problem under uncertainty and nurses capacity constraints. Health Care Manag. Sci. 19 (4), 383–394.

Ye, Y., 1996. Complexity analysis of the analytic center cutting plane method that uses multiple cuts. Math. Program. 78 (1), 85–104.

Zeng, B., Zhao, L., 2013. Solving two-stage robust optimization problems using a column-and-constraint generation method. Oper. Res. Lett. 41 (5), 457–461.

Zhang, J., Dridi, M., El Moudni, A., 2019. A two-level optimization model for elective surgery scheduling with downstream capacity constraints. European J. Oper. Res. 276 (2), 602–613.

Zhang, J., Dridi, M., El Moudni, A., 2020. Column-generation-based heuristic approaches to stochastic surgery scheduling with downstream capacity constraints. Int. J. Prod. Econ. 107764.

Zhang, Y., Jiang, R., Shen, S., 2018. Ambiguous chance-constrained binary programs under mean-covariance information. SIAM J. Optim. 28 (4), 2922–2944.

Zhang, Y., Shen, S., Erdogan, S.A., 2017. Distributionally robust appointment scheduling with moment-based ambiguity set. Oper. Res. Lett. 45 (2), 139–144.

Zheng, B., Denk, P., Martinec, D., Gatta, P., Whiteford, M., Swanström, L., 2008. Building an efficient surgical team using a bench model simulation: construct validity of the legacy inanimate system for endoscopic team training (LISETT). Surg. Endosc. 22 (4), 930–937.

Zhu, S., Fan, W., Yang, S., Pei, J., Pardalos, P.M., 2019. Operating room planning and surgical case scheduling: a review of literature. J. Comb. Optim. 37 (3), 757–805.