Coordinated and Priority-based Surgical Care: An Integrated Distributionally Robust Stochastic Optimization Approach

Esmaeil Keyvanshokooh[1]*, Pooyan Kazemian[2], Mohammad Fattahi[3], Mark P. Van Oyen[4]
[1]Department of Information & Operations Management, Mays Business School, Texas A&M University, College Station, TX 77845, USA, keyvan@mays.tamu.edu,
[2]Department of Operations, Weatherhead School of Management, Case Western Reserve University, Cleveland, OH 44106, USA, pooyan.kazemian@case.edu,
[3]Newcastle Business School, Northumbria University, United Kingdom, mohammad.fattahi@northumbria.ac.uk,
[4]Department of Industrial & Operations Engineering, University of Michigan, Ann Arbor, MI 48109, USA, vanoyen@umich.edu.

We study a Coordinated clinic and surgery Appointment Scheduling (CAS) problem for in-advance scheduling of surgical patients. Our models seek to provide timely access to care by coordinating clinic and surgery appointments to ensure that patients can see a surgeon in the clinic and (if needed) schedule their surgery within a maximum wait time target based on patient classes. There are different types of uncertainty including the number of appointment requests, whether a patient requires surgery, and surgery durations. We develop an Integrated Multi-stage Stochastic and Distributionally Robust Optimization (IMSDRO) approach to determine the optimal clinic and surgery dates for patients such that the access target constraints are satisfied, and the clinical and surgical overtimes are minimized. The IMSDRO approach synergizes multi-stage stochastic optimization with distributionally robust optimization to simultaneously incorporate multiple types of uncertainties by including stochastic scenarios for appointment request arrivals and ambiguity sets for surgery durations. Several new transformations are introduced to turn the nonlinear model derived from the IMSDRO approach to a tractable one, and a constraint generation algorithm is developed to solve it efficiently. We propose a data-driven Rolling Horizon Procedure (RHP) to facilitate implementation. We use case data to assess the performance of our policies. The results suggest that our policy can significantly improve surgical access delay times compared to the current practice. Our methodology is not limited to a particular setting and can be applied to other service industries where access delay matters.

Key words: healthcare coordination, access delay to care, data-driven optimization, multi-stage stochastic optimization, distributionally robust optimization.

History: Received: April 2019; Accepted: August 2021 by Edward Anderson, after 3 revisions.

1. Introduction

In a typical service system, it is inevitable that waiting times or delays will be experienced by customers due to the inherent uncertainty in both arrival processes and service times. In healthcare settings, a long waiting time to receive care is not only an annoyance, but it can also deteriorate health outcomes due to adverse events and increase healthcare costs because of the potential need

* Corresponding author.

for additional complicated procedures (Liu et al. 2017, Deglise-Hawkinson et al. 2018, and Oudhoff et al. 2007). *Timely access to care* is an essential feature of any high-quality and modern healthcare delivery system (Kaplan et al. 2015). We define "access delay" (or access to care) as the number of days between the day a patient's appointment request/referral is received by a medical center and their appointment day with a provider. Access delay can be mitigated by efficiently matching the available resource capacity to patient demand. This is, however, challenging given the inherent and various sources of uncertainty within any healthcare delivery system (Mays et al. 2009). Currently, the U.S. is experiencing an increase in demand for medical care due to an aging and growing population, which is outpacing the growth of healthcare providers (Markit 2017). This limited capacity along with sharply increasing demand leads to barriers to adequate access to care, and also highlights the importance of efficient utilization of resources, including providers and operating rooms. In this context, *coordination of patient care* throughout the course of treatment and across various clinic and surgery visits helps ensure that patients receive appropriate follow-up treatments without enduring long waiting times that can undermine their health condition.

This research is motivated by our collaborations with multiple healthcare institutions that desire to achieve timely access to surgery in their specialized surgical units. Patients with various acuity levels are referred to these surgical units either by their primary care physicians or by other hospital units. These patients first require a clinic consultation appointment with a surgeon, which then may need to be followed by a surgical procedure in the operating room. The decision of whether a patient requires a surgery is made during the patient's clinic consultation visit along with details of the surgery. In this paper, we develop a new optimization-based approach to coordinate clinic and surgery appointments for these surgical units such that all patients with various acuity levels can be offered a clinic consultation visit and a surgical time (if surgery is needed) within a pre-defined target time window that is clinically safe for them to wait using the minimum overtime possible. We call this the *Coordinated clinic and surgery Appointment Scheduling* (CAS) problem.

It is also worth noting that variability in *appointment request arrival numbers* and *surgery durations* can cause excessive patient waiting times and poor utilization of healthcare resources or high overtime. Unlike prior research that assumes the probability distribution of surgery duration is known (e.g., Denton and Gupta 2003, Denton et al. 2007, Erdogan and Denton 2013 and Diamant et al. 2018), our contribution is to consider how distributional robustness can be achieved using a model where only marginal information including mean, variance and range on surgery duration is used. Creating an accurate probability distribution for surgery duration, which can depend on the surgery type as well as the surgeon performing the surgery, requires a large amount of historical data. In many healthcare settings, however, a wide range of surgeries, limited numbers of cases of each type, and surgeons changing over time result in insufficient historical data to accurately

estimate surgery duration distributions for each combination of surgery-surgeon type. Further, it might be impossible to fit distributions tailored to the surgeon and the surgery type for some of the less common procedures. For example, as reported by Macario (2010), for approximately half of scheduled cases in the U.S. on any weekday, only five or fewer cases of the same surgery type (narrowly defined) and by the same surgeon have been performed. This motivates our interest in a *robust* scheduling policy that could perform relatively well against a class of surgery duration distributions satisfying only the above described moment (marginal) information in the CAS problem. This paper also incorporates the more traditional Poisson arrival process model. We assume there is usually enough historical data on appointment requests from which stochastic scenarios for the number of patient appointment requests can be made (Erdogan and Denton 2013).

Methodologically, we advance the literature by synergizing multi-stage stochastic optimization and distributionally robust optimization approaches such that the uncertainty in the number of patient appointment requests and surgery durations are modeled by a scenario tree and a moment-based ambiguity set, respectively. We call this new approach the *Integrated Multi-stage Stochastic and Distributionally Robust Optimization* (IMSDRO). The IMSDRO approach (i) specifies the optimal clinic date, (ii) determines the optimal surgery date with the same surgeon who performed the clinic visit (given surgery is needed), and (iii) minimizes and balances the clinic and surgery overtimes of surgeons. The IMSDRO approach *guarantees* that the pre-defined priority-based clinical and surgical access delay targets are met for all patients. Clinical and surgical overtimes are used, as needed, to achieve these predefined priority-based access delay targets.

In light of all the above discussions, we address the following two questions in this paper. (i) How can one develop an optimization-based model that seeks to establish timely access to specialized surgery by coordinating clinic and surgery appointments such that a patient is guaranteed to see a surgeon in the clinic and (if needed) receive surgery within a maximum wait time target that is clinically safe for them to wait? (ii) How can one synergize multi-stage stochastic optimization with distributionally robust optimization to concurrently deal with different types of uncertainty to generate a model that is efficiently solvable and implementable in everyday clinical practice?

1.1. Related Literature

Our work is related to multiple research areas, namely, appointment scheduling, healthcare coordination, distributionally robust optimization, and stochastic programming.

Appointment scheduling and healthcare coordination. There is a growing literature on appointment scheduling in healthcare. Recent papers include Liu et al. (2017), Lemay et al. (2017), Wang et al. (2018), Morrice et al. (2018), Liu et al. (2019a), Liu et al. (2019b), Jung et al. (2019), He et al. (2019), Yu et al. (2020), Mandelbaum et al. (2020), Bandi and Gupta (2020), Zacharias and Yunes (2020), Grant et al. (2021), Zhou et al. (2021), and Keyvanshokooh et al. (2021).

Gupta and Wang (2008) defined two access delay types. *Direct access delay* is the time between the patient's arrival to the clinic on the day of their appointment and the time the doctor sees them; *indirect access delay* is the time between the patient's appointment referral and the time of their scheduled appointment. Most works have concentrated on direct waiting times and far fewer considered indirect waiting times. Patrick et al. (2008) proposed a Markov decision process (MDP) to develop policies that minimize the number of patients that do not get a single appointment by a clinically determined maximum wait time target. Gupta and Wang (2008) studied an MDP under patients' preference for a clinic to decide how to manage access to its slots when patients can choose between a single same-day or future appointment. Liu et al. (2010) presented an MDP under no-show and cancellation to allocate each patient a single appointment date within a specific horizon. Saure et al. (2012) extended the work of Patrick et al. (2008) to require a sequence of appointment visits for each patient while reducing access delays. They assume multiple identical therapy machines and are thus able to model total capacity by aggregating individual capacities of machines, unlike our paper. Turkcan et al. (2012) considered a deterministic number of chemotherapy patients with multiple visits over time and formulate an optimization model to minimize access delay from their earliest start dates. Gocgun and Puterman (2014) studied a similar model to that of Patrick et al. (2008) and considered different patient types that require different levels of access to a single appointment. Diamant et al. (2018) developed an MDP under no-shows where patients undergo a series of assessments before being eligible for a surgery.

Our paper belongs to the stream of research on indirect access time. There are a number of key differences between the above papers and ours. First, we consider multiple non-identical surgeons as scarce resources as opposed to Saure et al. (2012), which modeled either one single resource or multiple identical resources. Second, it is mostly assumed that each patient needs either one (e.g., Patrick et al. 2008, Gupta and Wang 2008, Liu et al. 2010, and Gocgun and Puterman 2014) or multiple visits (e.g., Saure et al. 2012, Turkcan et al. 2012, and Diamant et al. 2018), and these are all assumed to be known at the time of receiving the request. But, in our problem, each patient requires a clinic visit, which may or may not be followed by a surgery, and the surgery need is realized at the clinic visit. Third, we do not consider no-shows and cancellations because they rarely occur in the settings of highly specialized clinics. Fourth, an important goal of our study is to achieve timely access to care using *access delay targets* and model uncertainty in surgery duration, which are not considered in Diamant et al. (2018) and Saure et al. (2012).

We address *healthcare coordination* in the sense of setting appointments for pairs of sequential visits that together achieve timely access to care. We found only two articles in this regard. Wang et al. (2018) proposed a coordinated pre-operative scheduling approach to evaluate patients' conditions prior to surgery. They model a two-station stochastic network, where each clinic may be

staffed by multiple parallel providers and patients see the first available one. They give a myopic scheduling policy due to their complex setting. Kazemian et al. (2017) used a simulation approach to evaluate their heuristic policies for coordinating clinic and surgery visits. Our work develops optimization models rather than heuristics for determining the clinic and surgery visit decisions, and models uncertainty in both surgery duration and arrival process. This provides a general approach to a broader range of systems because heuristics do not readily extend to new settings.

Stochastic optimization and distributionally robust optimization. As an alternative to MDP approaches and simulation to address uncertainty, two-stage stochastic optimization is usually employed to formulate appointment scheduling problems that incorporate uncertainty (see e.g., Denton and Gupta 2003, Mancilla and Storer 2012, and Parvin et al. 2018). However, the uncertainty in stochastic parameters such as demand is often realized *progressively*, and the decision at each stage should be a function of the observed feedback outcomes up to that stage. Multi-stage stochastic programming (MSSP) is a more suitable approach for modeling such a setting (see e.g., Erdogan and Denton 2013), which is the case in our paper.

Surgery durations across different patient classes and surgeons are not usually homogeneous; thus, it is challenging to characterize their exact probability distributions. To overcome this issue, distributionally robust optimization (DRO) approaches optimize the worst-case performance over an ambiguity set, which represents a class of probability distributions with specified moment information. Kong et al. (2013) formulated a DRO model in which an ambiguity set is used to include all distributions of service times with common mean and covariance and derived a semidefinite program. Mak et al. (2014) considered a similar problem except that service durations are independently distributed, and reformulated their DRO model as a conic program. However, their formulation requires the assumption that service durations could take on negative values. Jiang et al. (2017) considered a single-server DRO scheduling problem given a fixed sequence of appointments with ambiguous no-shows and service durations and derive mixed-integer nonlinear models.

There are key differences between the above papers and ours. First, the focus of their DRO models is not on real-world settings; however, we develop a DRO formulation for an appointment scheduling problem with realistic features. Second, we develop a new approach which integrates a DRO model with an MSSP model to incorporate different types of uncertainty as well. Third, we leverage a set of transformations to turn our nonlinear model into a tractable one, which can be efficiently solved by a new constraint generation algorithm.

The decisions made by most decision making under uncertainty approaches are often not implementable in practice. Rolling horizon type algorithms are usually developed to deal with this issue. For example, the rollout method for approximate dynamic programming (Bertsekas 2005, Bertsekas and Castanon 1999 and Bertsekas et al. 1997), Monte Carlo search tree method for reinforcement

learning (Browne et al. 2012, Gelly et al. 2012, and Munos et al. 2014) and rolling-horizon policies for MDPs (Hernández-Lerma and Lasserre 1990 and Alden and Smith 1992) are three main applications of this idea. However, we extend the idea of rolling horizon into our IMSDRO approach as a way of adapting to the effect of uncertainty in the novel case of MSSP integrated with DRO.

1.2. Main Contributions and Focus

Below, we summarize the major contributions of this paper to the existing literature.

(1) Integrated multi-stage stochastic and distributionally robust optimization. We believe that methodologically this paper is the first to develop an integrated multi-stage stochastic and distributionally robust optimization approach to simultaneously model two different types of uncertainties, namely the uncertainty in arrival process and the uncertainty in service time. While arrivals are often approximated by a Poisson process in operations models, in many services such as healthcare, the service time can depend greatly on what type of service is provided and by whom (Gupta and Denton 2008). In the context of a specialized surgical unit, several types of surgeries may be offered by a number of surgeons, making it challenging to elicit a complete probability distribution of surgery duration for all surgery type-surgeon combinations. Hence, a DRO approach that only relies on limited distributional information (e.g., mean, standard deviation, and range) combined with an MSSP model to model the arrival process is extremely valuable. In this paper, we first develop an MSSP model that defines the decisions to be made at each stage as a function of the observed outcomes up to that stage and models the uncertainty around appointment request arrivals by a Poisson process from which we can take enough random samples to make a *scenario tree*. We synergize this MSSP model with a DRO approach, which makes no assumption on the exact probability distribution of surgery duration. Instead, it describes a *moment-based ambiguity set*, which captures a class of distributions with specified moment information. The exact formulation derived by the IMSDRO approach is not tractable. We leverage a set of transformations to turn this nonlinear model into an approximate one that contains an embedded mixed-integer linear program in its constraints. We develop a new constraint generation algorithm that generates effective scenario cuts through this embedded optimization problem, to efficiently solve the model. Our IMSDRO methodology is flexible and can be applied to other service operations in which different types of uncertainty are to be modeled simultaneously. Our transformations can also be used for many other DRO models to turn them into tractable ones.

(2) Data-driven rolling horizon procedure. Since the decisions obtained by the IMSDRO approach are scenario dependent, they are not readily implementable in practice. We propose a *data-driven rolling horizon procedure* (RHP), which provides a framework to (i) make the decisions of the IMSDRO approach *implementable* in real practices, and (ii) empirically evaluate the performance of the scheduling policies obtained by the IMSDRO approach. The main advantage of the

RHP is that it allows practitioners to make use of the latest information that is revealed as time unfolds and adjust their decisions by dynamically utilizing the realization of uncertain parameters. This RHP resolves the critical limitation of traditional stochastic optimization policies, which are only valid for a limited number of scenarios. While the rolling horizon idea is, in general, similar to that of a rollout policy for constrained dynamic programs, a Monte Carlo search tree in reinforcement learning, and rolling-horizon MDPs, implementation of a data-driven rolling horizon procedure in the context of IMSDRO is novel.

(3) Healthcare coordination for timely access delay. This paper presents a class of scheduling policies that aim to coordinate clinic consultation and surgical appointments in a specialized surgical setting to accommodate patients of different acuity/priority levels within a predefined priority-based time window. The need to consider *care coordination* has been raised by May et al. (2011) and emphasized by the recent survey of Ahmadi-Javid et al. (2017). To the best of our knowledge, this paper is the first work to date that uses optimization approaches to study and model the impact of clinic and surgery appointment coordination to accommodate priority-based access delay targets. In §6, we demonstrate that coordinating clinic and surgery appointments in a specialized surgical unit using our new IMSDRO methodology can significantly improve surgical access delay for patients with acute conditions. We show that there is a trade-off between meeting access delay targets and incurring overtime. This allows decision makers to define a set of access delay targets that results in acceptable surgeon overtime. Although our work is motivated by healthcare, our models, methodology, and insights can also be extended to the general appointment-based service systems. We discuss several important practical implications and insights from our work in §7.

2. Problem Statement

In this section, we present the description and specifications of the CAS problem. This new problem is motivated by a real-world healthcare scheduling application in our collaborating hospitals.

Surgeons offer both clinic consultation and surgical procedures. There are a number of surgeons who work in two clinic and surgery teams. The set of surgeons is denoted by $\mathcal{K}$ where each surgeon is presented by $k \in \mathcal{K}$. Each period is typically a day. We use the terms day and period interchangeably. On any given day, one team sees patients in the clinic while the other team performs surgery in the operating rooms (ORs). Each surgeon switches between clinic and surgery teams on the following day and maintains his/her own clinical and surgical calendars. This is called an *every-other-day operating calendar* for surgeons. The system allows both clinical and surgical overtimes along with the regular clinical and surgical capacities for surgeons. Each surgeon $k \in \mathcal{K}$ has a regular clinical capacity of U_m^k on clinic day m, and a regular surgical capacity of V_n^k on surgery day n. These details can be easily modified to accommodate other healthcare settings.

Patients from different classes/types. There are different classes of patients whose requests are received by the surgical clinic. The set of *patient classes* is denoted by Γ where each patient class is presented by a tuple $\gamma = (\phi, \nu) \in \Gamma$, where ϕ is the *referral type* (i.e., local or remote) and ν is the *indications of disease* (e.g., colon cancer, rectal prolapse, diverticulitis, etc). Patients are referred either by other hospital units or by their primary care physicians to the surgical clinic to consult with a surgeon and evaluate the need for a surgery. When the surgical clinic receives an appointment request, the patient's electronic health records reveal the indication of disease as well as whether the patient is locally or remotely referred. The referral type and the indication of diseases together determine a patient's class. We use the terms request and referral interchangeably.

Each appointment request first requires a clinic consultation appointment with a surgeon, which then may need to be followed by a surgery. The decision of whether a patient requires a surgery is made during the patient's clinic visit. If a surgery is required, we assume that it has to be performed by the same surgeon who visited the patient at the clinic visit. This feature captures the *continuity of care* between patient-surgeon and is often preferred by patients since the patient has already established some trust and a relationship with the surgeon.

Various types of uncertainty. In light of the availability of historical data and the inherent uncertainty of the system, there are three types of uncertainty. The first is the total number of appointment requests received from each patient class in each period, which is realized at the end of the period. The second is whether a given patient requires a surgery or not, which is revealed at the clinic visit. If a surgery is needed, the third type of uncertainty concerns the surgery duration.

We know the probability distribution for the number of appointment requests made by each class in each period from which a set $\mathcal{S}$ of *stochastic scenarios* (indeed a scenario tree) is generated to model the existing uncertainty in appointment request arrivals. We represent this by $\mathcal{D}^s_{\gamma,t}$, which is the set of class $\gamma \in \Gamma$ patients whose request is received on any day t under scenario $s \in \mathcal{S}$ (see §3.1). Nonetheless, we have limited distributional information on the distribution of surgery duration $d_{\gamma,k}$ for each class $\gamma \in \Gamma$ and surgeon $k \in \mathcal{K}$ pair. There is usually a wide range of patient classes served by several different surgeons, which leads to having only a limited number of cases/examples for each patient class-surgeon combination. This makes it hard to fit distributions tailored to each surgeon and patient class pair, because individual surgeons may perform many surgery types with small annual volumes (see discussion in §1). A *moment-based ambiguity set* is employed to incorporate all such distributions with a common mean, standard deviation, and support (see §3.2).

Moreover, the patient class determines the probability r_γ that the patient will need a surgery. Indeed, whether a class γ patient needs a surgery follows a Bernoulli distribution with success probability r_γ. The surgery probability helps approximate the required surgery workload in the future. Accordingly, our approach in §3 considers the expected number of required surgeries (with

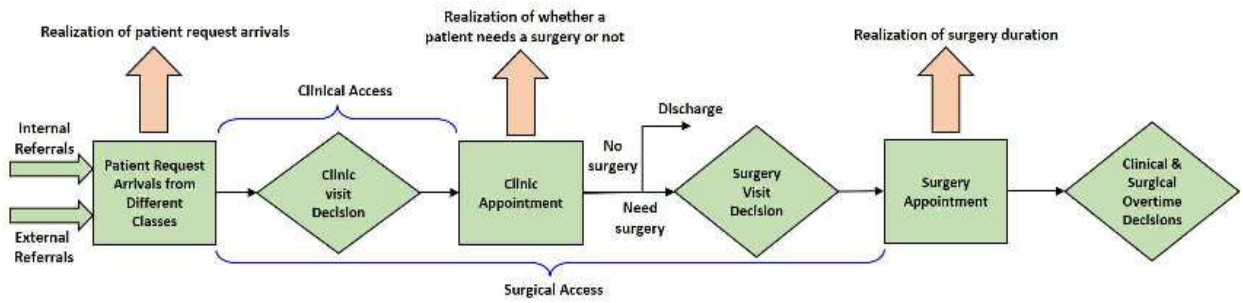


Figure 1 **The illustration of sequence of events, timing of different uncertainty realizations and proactive clinic and surgery scheduling decisions made for each patient request in the surgical clinic.**

respect to the future clinical visits) when making new clinical and surgical appointment decisions in each period.

We adopt a multi-stage decision-making setting as well because the uncertainty in parameters is progressively realized in each period. We then develop an integrated optimization-based approach, denoted above as the IMSDRO approach, synergizing an MSSP model with a DRO approach to simultaneously model all types of uncertainty. The goal is to find an optimal clinic and (if needed) surgery visit date for each patient with minimum overtimes for surgeons such that class-specific access delay targets are met for patients. We denote the clinical and surgical overtimes of surgeon $k \in \mathcal{K}$ on clinic day m and surgery day n under scenario $s \in \mathcal{S}$ by $q^k_{m,s}$ and $o^k_{n,s}$, respectively.

Clinical and surgical decisions. Figure 1 illustrates the sequence of events and decisions, surgical and clinical access delays and timing of uncertainty realizations. At the end of each period, the uncertainty about the number of appointment referrals received on that period is realized and the clinic appointments for those patients are scheduled at the end of that period. The clinic visit day is promised at the end of the arrival day and is denoted by $x^{k,m}_{p,\gamma,t,s}$, which is whether a class γ patient $p \in \mathcal{D}^s_{\gamma,t}$ whose request was received on any day t under scenario s, has clinic visit on day m with surgeon k. The next decision, the day of surgery, is made on the clinic appointment day. After completing the clinic visit, it becomes known whether the patient requires a surgery. If the patient needs a surgery, we schedule a surgery appointment, which must be with the same surgeon with whom she/he had the clinic visit. The surgery decision is denoted by $y^{k,n}_{\gamma,t,m,s}$, which is the number of class γ patients whose requests were received on any day t under scenario s and had clinic visit on m, and we choose surgery day n with surgeon k. After the realization of surgery need and duration, we calculate the clinical and the surgical overtimes $q^k_{m,s}$ and $o^k_{n,s}$, respectively.

Timely access to care. To ensure that patients are granted timely access to care, we place hard constraints on the allowable time intervals during which a patient may have clinic and surgery visits safely. For each patient class γ, we define a parameter called WTC_γ or "*minimum wait time target for the clinic visit of a patient class γ patient.*" For example, in our case study (see §6),

the value of this parameter only depends on ϕ as it was appropriate to assume the wait time to clinic is determined based on whether the patient is referred locally or remotely to the hospital. In particular, for the local referral, WTC_γ is zero because the patient is physically at or around the surgical clinic. But, for the remote referral, we allow a minimum of WTC_γ days (5 days in the case study) from when a patient referral is received until her/his clinic visit so as to give the patient time to make travel arrangement to the surgical clinic. We also define another important parameter called WTS_γ or "*maximum wait time target to surgery visit.*" This can be thought of as the maximum wait time that the patient's surgery, if needed, can be safely postponed from the time of patient referral. Our methodology ensures that all patients are offered at least one surgery visit within their WTS_γ. We define a parameter CSG_γ or "*minimum gap between clinic and surgery visits of a class γ patient.*" This corresponds to the minimum required number of days between the patient's clinic and surgery visits. While this can be zero, some surgeries require a period of preparation prior to the surgery. WTC_γ, WTS_γ and CSG_γ are set by the surgical clinic in our case study, but can be easily modified in other settings.

According to the above-defined parameters, if we receive a referral on any period t from patient class γ, we define (i) the earliest time $EC_{\gamma,t} = t + WTC_\gamma$, and the latest time $LC_{\gamma,t} = t + WTS_\gamma - CSG_\gamma$ for setting the clinic appointment, and (ii) the earliest time $ES_{\gamma,t} = t + WTC_\gamma + CSG_\gamma$ and the latest time $LS_{\gamma,t} = t + WTS_\gamma$ for choosing the surgery appointment (if needed) for this patient. In our approach, both clinic and surgery appointments are scheduled within these *clinical and surgical target time windows* that depend on the appointment request period. This requirement on each patient's flow pathway adds significant complexity to the CAS problem. Figure 2 depicts the allowable target time windows for having the clinic and surgery (if needed) appointments for a typical patient from class γ whose appointment request is received on any period t.

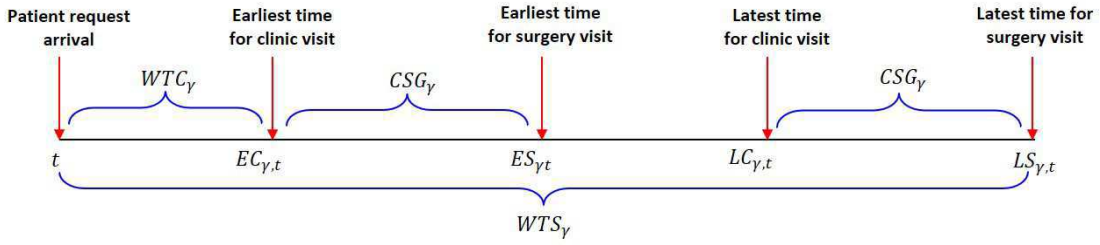


Figure 2 The illustration of minimum Wait Time for Clinic visit (WTC), minimum Clinic to Surgery visits Gap (CSG), maximum Wait Time to Surgery visit (WTS) for a patient whose request is received on any period t.

3. Integrated Multi-stage Stochastic and Distributionally Robust Optimization Methodology

Analytics overview. In this section, the IMSDRO methodology for the CAS problem is presented. In §3.1, we first assume that the surgery duration is deterministic and develop a Multi-stage Stochastic Mixed-Integer Program (MS-MIP) model in which a scenario tree is employed to model the uncertainty in the number of appointment requests on each period. In §3.2, we then extend this MS-MIP model to account for uncertainty in the surgery duration by developing a DRO approach that uses an ambiguity set constructed based on the empirical mean, standard deviation, and support of the surgery duration. Given that the resulting formulation is not tractable, we deploy a set of approximations based on the structural properties and a scenario cut-generating model, which results in an approximate tractable reformulation (IMSDRO-APRX).

3.1. Multi-stage Stochastic Mixed-Integer Model

We define three horizons (see Figure 3): (i) current scheduling horizon $\mathcal{L}$, (ii) current arrival horizon $\mathcal{T}$, and (iii) past arrival horizon $\mathcal{U}$. By using this modeling approach, we account for initial steady-state clinical and surgical workloads. The *current scheduling horizon* $\mathcal{L}$ is the set of periods from current period t_0 until period t_e, over which we decide the clinic and surgery appointment dates. There are two types of *patient arrival horizons*: (i) *"current" arrival horizon* $\mathcal{T}$ is the set of periods from current period t_0 until period t_b for new patient request arrivals, which is the first portion of the current scheduling horizon; (ii) *"past" arrival horizon* $\mathcal{U}$ is the set of periods from period 1 until period t_0-1 for past patient request arrivals over the previous scheduling horizon. The reason for defining the set $\mathcal{U}$ is twofold. First, the clinic visit of a patient whose request has been already received in $\mathcal{U}$ may happen on any period (day) in $\mathcal{L}$, so we may still need to make a surgery visit decision. Second, the surgery visit of the patients whose request is received in $\mathcal{U}$ may happen on any period in $\mathcal{L}$ and they consume surgery capacity during the horizon $\mathcal{L}$. Due to the access wait time targets to surgery, $|\mathcal{L}| = |\mathcal{T}| + \max_\gamma\{WTS_\gamma\}$ is the required length of current scheduling horizon. Note that t_0 is the first period of the current scheduling and arrival horizons, $t_b = t_0 + |\mathcal{T}| - 1$ is determined by the surgical unit and corresponds to the last period for new patient request arrivals, and $t_e = t_0 + |\mathcal{T}| + \max_\gamma\{WTS_\gamma\}$ corresponds to the end of current scheduling horizon.

A multi-stage stochastic model allows us to have several decision layers, where random outcomes are *progressively* realized, and the clinical and surgical decisions should be adapted to this process. In general, a T-stage stochastic model includes a sequence of stochastic parameters $\xi_1, \xi_2, \cdots, \xi_{T-1}$ with a discrete support (note that $T = |\mathcal{T}|$ in our model). A *scenario* is a realization of these stochastic parameters, and a *scenario tree* represents the progressive observation of random parameters. To model stochasticity in the number of appointment request arrivals as a scenario tree, a

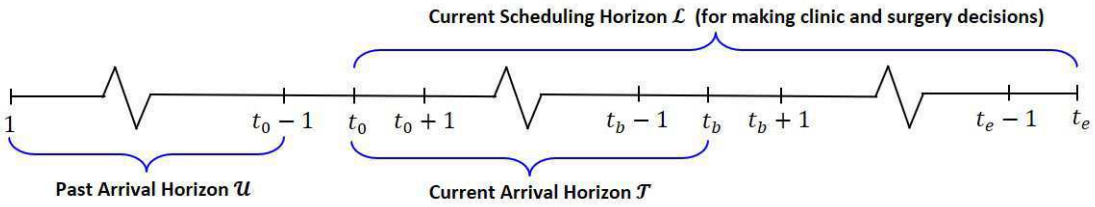


Figure 3 The illustration of arrival horizon $\mathcal{U}$ for patient request arrivals in the previous scheduling horizon, arrival horizon $\mathcal{T}$ for patient request arrivals in the current scheduling horizon, and current scheduling horizon $\mathcal{L}$.

set of scenarios $\mathcal{S}$ with a countable size $S=|\mathcal{S}|$ is defined. The corresponding scenarios' probabilities are $\pi_1,\pi_2,\cdots,\pi_S$, and a realization of stochastic parameters for scenario $s\in\mathcal{S}$ is presented by $(\xi^s_{t_0},\xi^s_{t_0+1},\cdots,\xi^s_{t_b})$ where $\xi^s_t=(\mathcal{D}^s_{\gamma,t}:\gamma\in\Gamma)$ is a realization for the number of requests on period $t\in\mathcal{T}$ over different classes under scenario $s\in\mathcal{S}$, and $\mathcal{D}^s_{\gamma,t}$ is the stochastic set of class γ patients whose request/referral is received in period $t\in\mathcal{T}$ under scenario $s\in\mathcal{S}$. Note that $\xi^s_{t_0}$ is the same (deterministic) for all scenarios $s\in\mathcal{S}$ because it is the number of appointment requests in the current period t_0 of the arrival horizon $\mathcal{T}$. We also define $\widetilde{\mathcal{D}}_{\gamma,t}$ as the deterministic set of class $\gamma\in\Gamma$ patients whose request was already received on day $t\in\mathcal{U}$. Formally, we have the following assumption for the number of appointment requests.

ASSUMPTION 1 **(Stochasticity Assumption)**. *There is full distributional information for the number of patient appointment requests in every period over the current arrival horizon $\mathcal{T}$. Such uncertainty is modeled by a stochastic process ξ with a realization of stochastic parameters presented by $(\xi^s_{t_0},\xi^s_{t_0+1},\cdots,\xi^s_{t_b})$ with a probability π_s under scenario $s\in\mathcal{S}$, where $\xi^s_t=(\mathcal{D}^s_{\gamma,t}:\gamma\in\Gamma)$ is a realization for the number of patient appointment requests in period $t\in\mathcal{T}$ under scenario $s\in\mathcal{S}$.*

In an MSSP, a policy should be *non-anticipative*, meaning that the decisions made at each stage must not be dependent on the future realization of stochastic parameters. There are two common ways for formulating an MSSP (Dupačová 1995). In the first, an MSSP is formulated as a sequence of nested two-stage stochastic programs in which non-anticipativity is implicitly imposed. In the second (used in this paper), a set of *non-anticipativity constraints* (NAC) is explicitly modeled.

Figure 4 (left-hand side) shows an example of a scenario tree with four stages and four scenarios for the CAS problem with two classes. In each scenario node, there is a realization $(|\mathcal{D}^s_{1,t}|,|\mathcal{D}^s_{2,t}|)$ where $|\mathcal{D}^s_{1,t}|$ and $|\mathcal{D}^s_{2,t}|$ are the number of class 1 and 2 appointment requests that are received on period $t\in\mathcal{T}$, respectively. For example, $\mathcal{D}^2_{1,3}=\{1,2,3\}$ and $\mathcal{D}^2_{2,3}=\{1,2\}$ are for the node at stage three and scenario two. Figure 4 (right-hand side) is an alternative presentation of the scenario tree, which is called *scenario fan*, where the individual scenarios observed in the particular stages are disaggregated over all periods to form four scenarios. However, this scenario fan is *not permissible*. If we solve the CAS problem for each of the scenarios, the solution found might not be feasible for

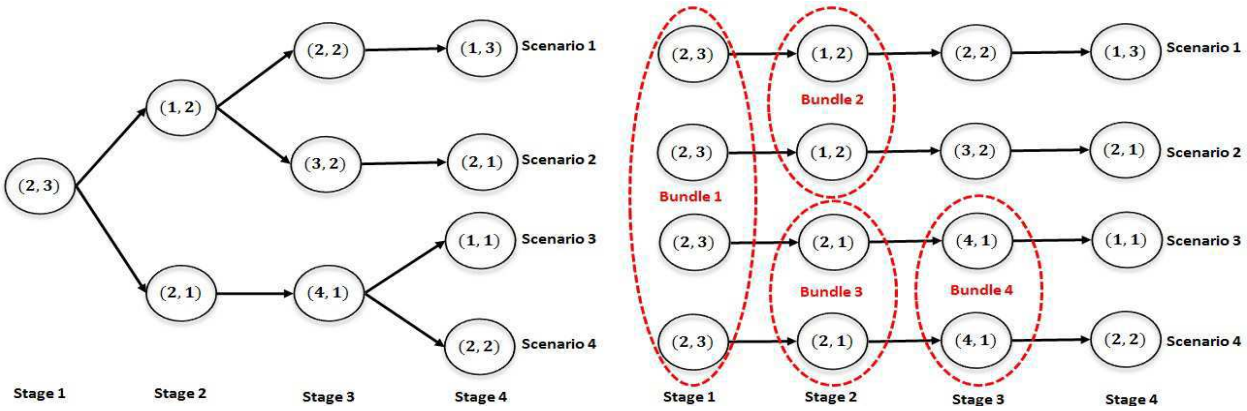


Figure 4 (LHS): An illustration of a scenario tree for the number of appointment request arrivals of 2 patient classes in a 4-stage MSSP with 4 scenarios where in each node (i, j) shows the number of appointment request arrivals of patient classes 1 and 2 at each stage t and scenario s, and (RHS): the corresponding scenario fan with four scenario bundles required for this 4-stage MSSP. The dashed ovals covering the nodes present NACs.

the overall problem because they imply decisions that anticipate future uncertain events. Thus, we need to enforce NACs to have permissible decisions. The dashed ovals covering the nodes represent NACs. For example, since all four scenarios have the same realizations at stage 1, they share the same scenario bundle, and so a NAC is imposed to guarantee that the same surgical and clinical decisions are made at all nodes in this scenario bundle. This is the same for scenarios 1 and 2 on period $t=1$, scenarios 3 and 4 on period $t=2$, and scenarios 3 and 4 on period $t=3$.

The other notations are given in Table 1. Tilda ($\sim$) is used to distinguish *decisions* x and y from *parameters* $\tilde{x}$, $\tilde{y}$, and $\tilde{z}$. The decision $\widehat{y}$ is also similar to decision y except it applies only to arrivals on current period t_0, which is deterministic (hence no dependence on scenario). We use bold notations whenever some indices of parameters/variables are removed.

The proposed multi-stage-stage stochastic model for the CAS problem is presented as follows, noting that the surgery duration $d_{\gamma,k}$ is assumed to be deterministic in this formulation.

Objective function. The objective (1) of the MS-MIP model is to minimize the expected total clinical and surgical overtimes of all surgeons over the scheduling horizon and scenarios.

$$\min \quad \sum_{s\in\mathcal{S}}\sum_{k\in\mathcal{K}}\pi_s\Big(\sum_{m\in\mathcal{L}} q^k_{m,s}+\sum_{n\in\mathcal{L}} o^k_{n,s}\Big). \tag{1}$$

Constraints for access delay to clinic appointments. Constraints (2)-(3) below guarantee that the clinic visit decision or $x^{k,m}_{p,\gamma,t,s}$ for each class γ patient whose request is received on any day t under scenario s should be an available date between the earliest clinic time $EC_{\gamma,t}=t+WTC_\gamma$ and the latest clinic time $LC_{\gamma,t}=t+WTS_\gamma-CSG_\gamma$ (see Figure 2 for the feasible clinic range).

$$x^{k,m}_{p,\gamma,t,s}=0, \quad \forall\gamma\in\Gamma, t\in\mathcal{T}, s\in\mathcal{S}, p\in\mathcal{D}^s_{\gamma,t}, k\in\mathcal{K},$$

Indices	
t, m, n	: Day indices (t is used for the day that an appointment request is received, and m and n are used for a clinic day and a surgery day, respectively.)
γ	: Patient class index, $\gamma=(\phi,\nu)\in\Gamma$ (ϕ is the referral type and ν is the disease indications).
k	: Surgeon index, $k\in\mathcal{K}$.
p	: Patient index, $p\in\mathcal{D}^s_{\gamma,t}$.
s	: Scenario index, $s\in\mathcal{S}$.
Deterministic and Stochastic Parameters	
U^k_m	: Total clinical capacity of surgeon $k\in\mathcal{K}$ on clinic day $m\in\mathcal{L}$.
V^k_n	: Total surgical capacity of surgeon $k\in\mathcal{K}$ on surgery day $n\in\mathcal{L}$.
c_γ	: Clinic duration of a patient class $\gamma\in\Gamma$.
$d_{\gamma,k}$	: Surgery duration of a patient class $\gamma\in\Gamma$ performed by surgeon $k\in\mathcal{K}$.
$\mathcal{D}^s_{\gamma,t}$	: Set of class $\gamma\in\Gamma$ patients whose request is received on day $t\in\mathcal{T}$ under scenario $s\in\mathcal{S}$.
$\widetilde{\mathcal{D}}_{\gamma,t}$	: Set of class $\gamma\in\Gamma$ patients whose request is already received on day $t\in\mathcal{U}$.
π_s	: Probability of occurrence of scenario $s\in\mathcal{S}$.
r_γ	: Surgery probability of a class $\gamma\in\Gamma$ patient.
$\widetilde{x}^{k,m}_{p,\gamma,t}$	: Binary parameter equal to 1 if a class $\gamma\in\Gamma$ patient p whose request was received on day $t\in\mathcal{U}$ has clinic visit on day $m\in\mathcal{T}\backslash\{t_0\}$ with surgeon $k\in\mathcal{K}$, and zero otherwise.
$\widetilde{z}^k_{\gamma,t}$	: The number of class $\gamma\in\Gamma$ patients whose request is received on day $t\in\mathcal{U}\cup\{t_0\}$, and has clinic visit on day t_0 with surgeon $k\in\mathcal{K}$, and also needs surgery.
$\widetilde{y}^{k,n}_{\gamma,t,m}$	: The number of class $\gamma\in\Gamma$ patients whose request is received on day $t\in\mathcal{U}$, and has clinic visit on day $m\in\mathcal{U}$, and surgery visit on day $n\in\mathcal{T}$ with surgeon $k\in\mathcal{K}$.
Stage Decision Variables	
$x^{k,m}_{p,\gamma,t,s}$	: Binary variable equal to 1 if a class $\gamma\in\Gamma$ patient p whose request is received on day $t\in\mathcal{T}$ under scenario $s\in\mathcal{S}$ has clinic visit on day $m\in\mathcal{L}$ with surgeon $k\in\mathcal{K}$, and 0 otherwise.
$y^{k,n}_{\gamma,t,m,s}$	: The number of class $\gamma\in\Gamma$ patients whose requests are received on day $t\in\mathcal{U}\cup\mathcal{T}$ under $s\in\mathcal{S}$, and have clinic visit on $m\in\mathcal{T}\backslash\{t_0\}$, and surgery visit on $n\in\mathcal{L}$ with surgeon $k\in\mathcal{K}$.
$\widehat{y}^{k,n}_{\gamma,t,t_0}$	: The number of class $\gamma\in\Gamma$ patients whose requests are received on day $t\in\mathcal{U}\cup\{t_0\}$, and have clinic visit on day t_0, and surgery visit on $n\in\mathcal{L}$ with surgeon $k\in\mathcal{K}$.
$q^k_{m,s}$	: Clinical overtime of surgeon $k\in\mathcal{K}$ on the clinic day $m\in\mathcal{L}$ under $s\in\mathcal{S}$.
$o^k_{n,s}$	: Surgical overtime of surgeon $k\in\mathcal{K}$ on the surgery day $n\in\mathcal{L}$ under $s\in\mathcal{S}$.

Table 1 **The description of indices, parameters and decisions of the MS-MIP model for the CAS problem.**

$$m\in\big[t_0,t+WTC_\gamma-1\big]\cup\big[t+WTS_\gamma-CSG_\gamma+1,t_e\big]. \tag{2}$$

$$\sum_{m=t+WTC_\gamma}^{t+WTS_\gamma-CSG_\gamma}\sum_{k\in\mathcal{K}}x^{k,m}_{p,\gamma,t,s}=1,\ \ \forall\gamma\in\Gamma,t\in\mathcal{T},s\in\mathcal{S},p\in\mathcal{D}^s_{\gamma,t}. \tag{3}$$

Constraints for access delay to surgery appointments. Constraints (4)-(6) state that the surgery of each patient is performed by the same surgeon who performed the clinic visit, and the surgery visit for each patient should be within a clinically safe range of days (see Figure 2 for the feasible surgery range). More explicitly, the class γ patients whose requests are received on either day $t\in\mathcal{U}$ or day $t\in\mathcal{T}$, and have clinic visit on day $m\in\mathcal{T}$ could have their surgery visit on any day between $m+CSG_\gamma$ and $t+WTS_\gamma$. We denote the surgery visit decisions by $y^{k,n}_{\gamma,t,m,s}$ and $\widehat{y}^{k,n}_{\gamma,t,t_0}$ for patients whose clinic visit is any day $m\in\mathcal{T}\backslash\{t_0\}$ and the current day t_0, respectively.

$$r_\gamma\Big(\sum_{p\in\widetilde{\mathcal{D}}_{\gamma,t}}\widetilde{x}^{k,m}_{p,\gamma,t}\Big)\ \le\sum_{n=m+CSG_\gamma}^{t+WTS_\gamma}y^{k,n}_{\gamma,t,m,s},\ \ \forall\gamma\in\Gamma,t\in\mathcal{U},m\in\mathcal{T}\backslash\{t_0\},k\in\mathcal{K},s\in\mathcal{S}. \tag{4}$$

$$r_\gamma\Big(\sum_{p\in\mathcal{D}^s_{\gamma,t}} x^{k,m}_{p,\gamma,t,s}\Big) \le \sum_{n=m+CSG_\gamma}^{t+WTS_\gamma} y^{k,n}_{\gamma,t,m,s},\ \forall\gamma\in\Gamma, t\in\mathcal{T}, m\in\mathcal{T}\backslash\{t_0\}, k\in\mathcal{K}, s\in\mathcal{S}. \tag{5}$$

$$\widetilde{z}^k_{\gamma,t} \le \sum_{n=t_0+CSG_\gamma}^{t+WTS_\gamma} \widehat{y}^{k,n}_{\gamma,t,t_0},\ \forall\gamma\in\Gamma, t\in\mathcal{U}\cup\{t_0\}, k\in\mathcal{K}. \tag{6}$$

Constraints (4) are for the class γ patients $\widetilde{\mathcal{D}}_{\gamma,t}$ whose request is received in the *previous* arrival horizon $\mathcal{U}$ (so they are already in the system) and their clinic visits are denoted by parameter $\widetilde{x}^{k,m}_{p,\gamma,t}$, and their surgery is being made on one day in the current horizon $\mathcal{T}$. However, constraints (5) are for the class γ patients $\mathcal{D}^s_{\gamma,t}$ whose request is received in the *current* arrival horizon $\mathcal{T}$ under scenario s and their clinic visits are denoted by $x^{k,m}_{p,\gamma,t,s}$. In both constraints (4) and (5), the clinic appointment of patients may happen on any day over horizon $\mathcal{T}\backslash\{t_0\}$, so their surgery need is specified by a surgery probability r_γ as their clinic visit has not happened yet. Recall that whether a class γ patient needs a surgery follows a Bernoulli distribution with success probability r_γ. Constraints (6) are for the patients denoted by parameter $\widetilde{z}^k_{\gamma,t}$ whose request is received on any day in horizon $\mathcal{U}\cup\{t_0\}$ (so they are already in the system), but unlike constraints (4)-(5), their clinic visit is on the current day t_0, and hence their surgery need is realized.

Clinical and surgical capacity constraints. Constraints (7)-(8) restrict the amount of clinical and surgical workloads (both regular capacity and overtime) for each surgeon $k\in\mathcal{K}$ on each day $n\in\mathcal{L}$, respectively, over the scheduling horizon. Note that $\tilde{x}^{k,m}_{p,\gamma,t}$ and $\tilde{y}^{k,n}_{\gamma,t,m}$ correspond to the decisions made in the previous periods; thus, they become parameters in the current time period.

$$\sum_{\gamma\in\Gamma} c_\gamma\Big(\sum_{t\in\mathcal{U}}\sum_{p\in\widetilde{\mathcal{D}}_{\gamma,t}} \widetilde{x}^{k,m}_{p,\gamma,t} + \sum_{t\in\mathcal{T}}\sum_{p\in\mathcal{D}^s_{\gamma,t}} x^{k,m}_{p,\gamma,t,s}\Big) \le U^k_m + q^k_{m,s},\ \forall m\in\mathcal{L}, k\in\mathcal{K}, s\in\mathcal{S}. \tag{7}$$

$$\sum_{\gamma\in\Gamma} d_{\gamma,k}\Big(\sum_{t\in\mathcal{U}}\sum_{m\in\mathcal{U}} \widetilde{y}^{k,n}_{\gamma,t,m} + \sum_{t\in\mathcal{T}}\sum_{m\in\mathcal{T}\backslash\{t_0\}} y^{k,n}_{\gamma,t,m,s} + \sum_{t\in\mathcal{U}\cup\{t_0\}} \widehat{y}^{k,n}_{\gamma,t,t_0}\Big) \le V^k_n + o^k_{n,s}, \forall n\in\mathcal{L}, k\in\mathcal{K}, s\in\mathcal{S}. \tag{8}$$

Non-anticipativity constraints. In any given stage over the scheduling horizon, the decision maker cannot foresee the future outcomes of the total number of appointment requests; therefore, the clinic and surgery decisions must satisfy NACs. This indicates that these decisions in a given stage t are identical for each pair (s, s') of scenarios with a common ancestor node in that stage (see Figure 4). If two scenarios s and s' share the same history of random parameters ξ^s and $\xi^{s'}$ up to stage t, then the decisions made at stage t are the same among all scenarios placed in the same scenario bundle. Constraints (9)-(10) are the corresponding NAC for the CAS problem.

$$x^{k,m}_{p,\gamma,t,s} = x^{k,m}_{p,\gamma,t,s'}, \forall k\in\mathcal{K}, \gamma\in\Gamma, m\in\mathcal{L}, t\in\mathcal{T}, p\in\mathcal{D}^s_{\gamma,t}, s, s'\in\mathcal{S}, (\xi^s_{t_0+1},\cdots,\xi^s_t) = (\xi^{s'}_{t_0+1},\cdots,\xi^{s'}_t). \tag{9}$$

$$y^{k,n}_{\gamma,t,m,s} = y^{k,n}_{\gamma,t,m,s'}, \forall k\in\mathcal{K}, \gamma\in\Gamma, m\in\mathcal{T}\backslash\{t_0\}, t\in\mathcal{T}\cup\mathcal{U}, n\in\mathcal{L}, s, s'\in\mathcal{S}, (\xi^s_{t_0+1},\cdots,\xi^s_t) = (\xi^{s'}_{t_0+1},\cdots,\xi^{s'}_t). \tag{10}$$

Note that we do not require defining the NACs for the other variables $q^k_{m,s}$ and $o^k_{n,s}$. The reason is because these auxiliary decisions are calculated directly from decisions $x^{k,m}_{p,\gamma,t,s}$ and $y^{k,n}_{\gamma,t,m,s}$ by constraints (7)-(8), and thereby preserving the non-anticipativity for them automatically.

Other constraints. Constraints (11)-(14) define the binary and non-negativity restrictions on the clinic and surgery appointment decisions, and clinical and surgical overtimes, respectively.

$$x^{k,m}_{p,\gamma,t,s} \in \{0,1\},\ \forall k \in \mathcal{K}, m \in \mathcal{L}, \gamma \in \Gamma, t \in \mathcal{T}, p \in \mathcal{D}^s_{\gamma,t}, s \in \mathcal{S}. \tag{11}$$

$$y^{k,n}_{\gamma,t,m,s} \geq 0,\ \ \forall k \in \mathcal{K}, m \in \mathcal{T}\backslash\{t_0\}, \gamma \in \Gamma, t \in \mathcal{U} \cup \mathcal{T}, n \in \mathcal{L}, s \in \mathcal{S}. \tag{12}$$

$$\widehat{y}^{k,n}_{\gamma,t,t_0} \geq 0,\ \ \forall k \in \mathcal{K}, n \in \mathcal{L}, \gamma \in \Gamma, t \in \mathcal{U} \cup \{t_0\}. \tag{13}$$

$$q^k_{m,s}, o^k_{n,s} \geq 0,\ \forall k \in \mathcal{K}, m, n \in \mathcal{L}, s \in \mathcal{S}. \tag{14}$$

Remark (Patient-centered Care). The above MS-MIP model has been developed to be *patient-centered* by putting *hard constraints* on access delay targets, thereby guaranteeing full service (i.e., clinic and surgery appointments) within a predefined priority-based safe interval. It is, however, inevitable to employ overtime on some days to achieve this goal, and the best scheduling policy is thus the one that meets such service level with the minimum possible clinical and surgical overtime. Generally, there is a trade-off between access delay targets and surgeon overtime. The tighter the access targets, the higher the overtime. In Appendix C, we develop an alternative bi-objective model, that strikes a balance between meeting access delay targets and incurring overtime. It allows decision makers to set penalties on violating access targets and incurring surgeon overtime.

3.2. Integrated Multi-stage Stochastic and Distributionally Robust Model

In this section, we extend the MS-MIP model (1)-(14), by incorporating ambiguous distributional information for surgery duration of each patient class and surgeon pair. Surgery duration is usually highly variable (see discussions in §1); however, there is often little uncertainty in clinic duration (e.g., in our partner hospitals, the clinic visits are scheduled in 15-minute time slots). The uncertainty in surgery duration is modeled by using an ambiguity set that is constructed based on the empirical mean, standard deviation, and support of the surgery duration. More precisely, besides Assumption 1, another important assumption in the IMSDRO is as follows.

Assumption 2 **(Ambiguity Assumption)**. *There is ambiguous distributional information about the surgery duration of each patient class and surgeon pair. This limited distributional information includes two stochastic moments (i.e., mean and standard deviation), and the support. A moment-based ambiguity set is used to model such uncertainty in the surgery duration.*

From the Assumption 2, the surgery duration vector $\boldsymbol{d} = (d_{\gamma,k} : \gamma \in \Gamma, k \in \mathcal{K})$ for different classes and surgeons has an unknown probability distribution P with a *polyhedral support set* Θ as follows:

$$\Theta = \Big\{\boldsymbol{d} \in \mathbb{R}_+^{|\Gamma|\times|\mathcal{K}|} : d^{LB}_{\gamma,k} \leq d_{\gamma,k} \leq d^{UB}_{\gamma,k},\ \forall \gamma \in \Gamma, k \in \mathcal{K}\Big\}, \tag{15}$$

where $\boldsymbol{d}^{LB}$ and $\boldsymbol{d}^{UB} \in \mathbb{R}_+^{|\Gamma|\times|\mathcal{K}|}$ denote the lower and upper bound vectors for the surgery duration $\boldsymbol{d}$, respectively. Such lower and upper bounds can be computed from available historical data.

DEFINITION 1 (**Marginal Moment-based Ambiguity Set**). *Given a set of $|L|$ observations of surgery duration $\boldsymbol{d}$, denoted by $\{\boldsymbol{d}^l\}_{l\in L}$ where $\boldsymbol{d}^l \in \mathbb{R}_+^{|\Gamma|\times|\mathcal{K}|}$, a moment-based ambiguity set $\Phi(\boldsymbol{\mu},\boldsymbol{\sigma},\Theta)$ is defined for the probability distribution P using the marginal mean vector $\boldsymbol{\mu} \in \mathbb{R}_+^{|\Gamma|\times|\mathcal{K}|}$ and standard deviation vector $\boldsymbol{\sigma} \in \mathbb{R}^{|\Gamma|\times|\mathcal{K}|}$ of these realizations of surgery durations as follows:*

$$\Phi(\boldsymbol{\mu},\boldsymbol{\sigma},\Theta) = \Big\{P : \int_\Theta dP(d) \;=\; 1, \tag{16a}$$

$$\int_\Theta d_{\gamma,k}\; dP(d) \;=\; \mu_{\gamma,k},\;\; \forall \gamma\in\Gamma, k\in\mathcal{K} \tag{16b}$$

$$\int_\Theta d^2_{\gamma,k}\; dP(d) \;=\; \mu^2_{\gamma,k} + \sigma^2_{\gamma,k},\;\; \forall \gamma\in\Gamma, k\in\mathcal{K}\Big\}. \tag{16c}$$

$\Phi(\boldsymbol{\mu},\boldsymbol{\sigma},\Theta)$ is the set of all plausible surgery distributions that satisfy (16a)-(16c). Constraint (16a) ensures that this moment-based ambiguity set contains only plausible probability distributions over the polyhedral support set Θ. Constraints (16b)-(16c) limit such probability distributions to have marginal first and second distributional moments being equal to those of the observed surgery durations. This set satisfies all candidate distributions whose marginal means and standard deviations match $\mu_{\gamma,k}$ and $\sigma_{\gamma,k}$, respectively, for each pair of patient class $\gamma\in\Gamma$ and surgeon $k\in\mathcal{K}$.

We are now ready to develop the IMSDRO model for the CAS problem, which is derived based on both Assumptions 1 and 2. The integrated model combines the MS-MIP model (1)-(14) with a DRO approach such that we can handle different types of uncertainty in one optimization model. We formulate this integrated model as the following *min-max* problem:

$$Z^{IMSDRO} \;=\; \min_{\boldsymbol{x},\boldsymbol{y},\widehat{\boldsymbol{y}},\boldsymbol{q}} \sum_{s\in\mathcal{S}} \pi_s \Bigg\{ \sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{L}} q^k_{m,s} + \max_{P\in\Phi(\boldsymbol{\mu},\boldsymbol{\sigma},\Theta)} \mathbb{E}_P\Big[f_s(\boldsymbol{y_s},\widehat{\boldsymbol{y}},\boldsymbol{d})\Big]\Bigg\} \tag{17a}$$

$$\text{s.t. } (\boldsymbol{x_s},\, \boldsymbol{y_s},\widehat{\boldsymbol{y}},\, \boldsymbol{q_s}) \in \mathcal{R}_s,\;\; \forall s\in\mathcal{S} \tag{17b}$$

where $\mathbb{E}_P$ is the expectation taken over the probability distribution P, and the feasible region $\mathcal{R}_s$ is defined by constraints (2)-(7) and (9)-(14) for each individual scenario $s\in\mathcal{S}$. Given the surgery appointment decisions $\boldsymbol{y_s}$ and $\widehat{\boldsymbol{y}}$, and a realization of random variable $\boldsymbol{d}$, $f_s(\boldsymbol{y_s},\widehat{\boldsymbol{y}},\boldsymbol{d})$ is defined by $\sum_{n\in\mathcal{L}}\sum_{k\in\mathcal{K}} \max\Big\{0, \sum_{\gamma\in\Gamma} d_{\gamma,k}\Big(\sum_{t\in\mathcal{U}}\sum_{m\in\mathcal{T}\setminus\{t_0\}} \widetilde{y}^{k,n}_{\gamma,t,m} + \sum_{t\in\mathcal{T}}\sum_{m\in\mathcal{T}\setminus\{t_0\}} y^{k,n}_{\gamma,t,m,s} + \sum_{t\in\mathcal{U}\cup\{t_0\}} \widehat{y}^{k,n}_{\gamma,t,t_0}\Big) - V^k_n\Big\}$.

Intuitively, $f_s(\boldsymbol{y_s},\widehat{\boldsymbol{y}},\boldsymbol{d})$ is the cumulative surgical overtimes of all surgeons over the scheduling horizon. The objective function of the IMSDRO model (17a)-(17b) then implies that we are making the clinic and surgery appointment decisions, and clinical and surgical overtimes decisions so as to minimize the expected clinical overtimes plus the *worst-case* expected surgical overtimes of all surgeons over the set of plausible surgery duration distributions $P\in\Phi(\boldsymbol{\mu},\boldsymbol{\sigma},\Theta)$. The distributionally robust part seeks the worst-case distribution P of $\boldsymbol{d}$ for which $\mathbb{E}_P\big[f_s(\boldsymbol{y_s},\widehat{\boldsymbol{y}},\boldsymbol{d})\big]$ is maximized.

Our next step is to reformulate the min-max IMSDRO model (17a)-(17b) into a *tractable* reformulation using the moment-based ambiguity set $\Phi(\boldsymbol{\mu}, \boldsymbol{\sigma}, \Theta)$. We first analyze the inner maximization problem in the model (17a)-(17b). For any fixed surgical decisions $\boldsymbol{y_s}$ and $\widehat{\boldsymbol{y}}$, and the uncertain realization vector $\boldsymbol{d}$, we consider the following *moment problem*:

$$\max_{P \in \Phi(\boldsymbol{\mu}, \boldsymbol{\sigma}, \Theta)} \mathbb{E}_P \Big[f_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{d}) \Big], \tag{18}$$

where the probability measure P is decision variable. We next expand this problem, which helps convert the min-max IMSDRO model (17a)-(17b) into an equivalent single-level minimization one.

Remark. The min-max IMSDRO model (17a)-(17b) with the moment problem (18) has a special structure, which is useful for many operational problems in which $f_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{d})$ has the form of cumulative maximization values. So, our methodologies can be used for a broad range of settings.

PROPOSITION 1 **(Reformulation of the min-max IMSDRO Model)**. *Under the moment-based ambiguity set $\Phi(\boldsymbol{\mu}, \boldsymbol{\sigma}, \Theta)$ for the probability distribution P of the surgery duration characterized by the constraints (16a)-(16c), the min-max IMSDRO model (17a)-(17b) can be reformulated as the following equivalent minimization problem,*

$$Z^{IMSDRO} = \min_{\boldsymbol{x}, \boldsymbol{y}, \widehat{\boldsymbol{y}}, \boldsymbol{q}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\beta}} \sum_{s \in \mathcal{S}} \pi_s \left\{ \sum_{k \in \mathcal{K}} \Big(\sum_{m \in \mathcal{L}} q^k_{m,s} + \sum_{\gamma \in \Gamma} \big(\mu_{\gamma,k}\, \alpha^k_{\gamma,s} + (\mu^2_{\gamma,k} + \sigma^2_{\gamma,k})\, \beta^k_{\gamma,s} \big) \Big) + \delta_s \right\} \tag{19a}$$

$$s.t.\ \delta_s \ \geq \ \max_{\boldsymbol{d} \in \Theta} \Big\{ f_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{d}) - \sum_{\gamma \in \Gamma} \sum_{k \in \mathcal{K}} d_{\gamma,k}\, \alpha^k_{\gamma,s} - \sum_{\gamma \in \Gamma} \sum_{k \in \mathcal{K}} d^2_{\gamma,k}\, \beta^k_{\gamma,s} \Big\}, \ \forall s \in \mathcal{S} \tag{19b}$$

$$(\boldsymbol{x_s}, \boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{q_s}) \in \mathcal{R}_s, \ \forall s \in \mathcal{S} \tag{19c}$$

$$\delta_s \in \mathbb{R}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s} \in \mathbb{R}^{|\Gamma| \times |\mathcal{K}|}, \ \forall s \in \mathcal{S}, \tag{19d}$$

where $\delta_s \in \mathbb{R}$, and $\boldsymbol{\alpha_s}, \boldsymbol{\beta_s} \in \mathbb{R}^{|\Gamma| \times |\mathcal{K}|}$ are dual variables for constraints (16a)-(16c), respectively.

Structural properties. The proof of Proposition 1 is provided in Appendix A.1. The reformulation (19a)-(19d) of the IMSDRO model in Proposition 1 is still nonlinear due to the maximization expression on the right-hand side of (19b). To obtain a tractable reformulation, we first attain a characterization of the overtime function $f_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{d})$ by converting it into an equivalent minimization linear program (LP) with the help of the surgical overtime definition (see LP (**??**)-(**??**) in the proof of Proposition 2 in Appendix A.2). We formulate its dual to merge it with the maximization over $\boldsymbol{d} \in \Theta$ in constraint (19b), and then reformulate the resulting problem based on the special structural properties, including (i) the surgery duration $\boldsymbol{d}$ has a *polyhedron-shaped support* Θ, and (ii) the dual variables for the $f_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{d})$ problem is *bounded* below and above by zero and one.

PROPOSITION 2 **(Reformulation of surgical overtime function)**. *For any fixed and feasible value of $\boldsymbol{y_s}$, $\widehat{\boldsymbol{y}}$, $\boldsymbol{\alpha_s}$, $\boldsymbol{\beta_s}$, $\boldsymbol{x_s}$, and $\boldsymbol{q_s}$ vectors and δ_s under scenario $s \in \mathcal{S}$ in the minimization*

problem (19a)-(19d), the value of the maximization problem on the right hand side of constraint (19b), i.e., $\Psi_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s}) = \max_{\boldsymbol{d}\in\Theta} \left\{ f_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{d}) - \sum_{\gamma\in\Gamma}\sum_{k\in\mathcal{K}} d_{\gamma,k}\alpha^k_{\gamma,s} - \sum_{\gamma\in\Gamma}\sum_{k\in\mathcal{K}} d^2_{\gamma,k}\beta^k_{\gamma,s} \right\}$*, is equivalent to the following problem under each scenario* $s \in \mathcal{S}$*:*

$$\Psi_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s}) = \max_{\lambda_s\in\Lambda_s} \sum_{\gamma\in\Gamma}\sum_{k\in\mathcal{K}} \left\{ \max_{d^{LB}_{\gamma,k} \le d_{\gamma,k} \le d^{UB}_{\gamma,k}} \left(\sum_{n\in\mathcal{L}} \left\{ \sum_{t\in\mathcal{U}} \sum_{m\in\mathcal{T}\setminus\{t_0\}} \widetilde{y}^{k,n}_{\gamma,t,m} + \sum_{t\in\mathcal{T}} \sum_{m\in\mathcal{T}\setminus\{t_0\}} y^{k,n}_{\gamma,t,m,s} + \sum_{t\in\mathcal{U}\cup\{t_0\}} \widehat{y}^{k,n}_{\gamma,t,t_0} \right\} \cdot \lambda^k_{n,s}\, d_{\gamma,k} - \alpha^k_{\gamma,s}\, d_{\gamma,k} - \beta^k_{\gamma,s}\, d^2_{\gamma,k} \right) - \sum_{n\in\mathcal{L}} V^k_n \lambda^k_{n,s} \right\}, \tag{20}$$

where feasible region Λ_s *is a polyhedron given by* $\Lambda_s = \{\boldsymbol{\lambda_s} \in \mathbb{R}^{|\mathcal{K}|\times|\mathcal{L}|} : 0 \le \lambda^k_{n,s} \le 1, \forall k \in \mathcal{K}, n \in \mathcal{L}\}$ *for each scenario* $s \in \mathcal{S}$*, and* $\lambda^k_{n,s}$ *is the dual variable associated with surgical overtime constraints.*

The proof of Proposition 2 is provided in Appendix A.2.

Discrete approximations. We next analyze the inner-maximization problem (21) embedded in $\Psi_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s})$ for each pair of class $\gamma \in \Gamma$ and surgeon $k \in \mathcal{K}$ and compute its optimal solution based on the structure of the polyhedron-shaped support Θ defined by the set (15):

$$\max_{d^{LB}_{\gamma,k} \le d_{\gamma,k} \le d^{UB}_{\gamma,k}} \left(\sum_{n\in\mathcal{L}} \left\{ \sum_{t\in\mathcal{U}} \sum_{m\in\mathcal{T}\setminus\{t_0\}} \widetilde{y}^{k,n}_{\gamma,t,m} + \sum_{t\in\mathcal{T}} \sum_{m\in\mathcal{T}\setminus\{t_0\}} y^{k,n}_{\gamma,t,m,s} + \sum_{t\in\mathcal{U}\cup\{t_0\}} \widehat{y}^{k,n}_{\gamma,t,t_0} \right\} \cdot \lambda^k_{n,s}\, d_{\gamma,k} - \alpha^k_{\gamma,s}\, d_{\gamma,k} - \beta^k_{\gamma,s}\, d^2_{\gamma,k} \right). \tag{21}$$

The inner-maximization problem (21) is a concave quadratic program. However, finding a closed-form solution for this problem over $d_{\gamma,k}$ is not trivial because the optimal value of $d_{\gamma,k}$ depends on all the coefficients in (21), which are themselves variables in problem (20). Even if the closed-form optimal solution for $d_{\gamma,k}$ is incorporated into (20), it becomes nonlinear because we obtain a quadratic expression in $\lambda^k_{n,s}$. To overcome this issue, we approximate (21) using a *piece-wise linear function* with equal length pieces. This is a common technique in optimization (Yang and Goh 1997). We define a set of $H+1$ segment points $\Upsilon_{\gamma,k} = \{\tilde{d}_{\gamma,k}(i)\}_{i=0}^{H}$ for the surgery duration of each class $\gamma \in \Gamma$ and surgeon $k \in \mathcal{K}$ pair, where $\tilde{d}_{\gamma,k}(i) = (1 - \frac{i}{H})\, d^{LB}_{\gamma,k} + (\frac{i}{H})\, d^{UB}_{\gamma,k}$, $i \in \{0, \cdots, H\}$ is the i^{th} *segment point* in the set $\Upsilon_{\gamma,k}$ for each class $\gamma \in \Gamma$ and surgeon $k \in \mathcal{K}$ pair.

The inner-maximization problem (21) then reduces to the following approximation problem of finding the maximum over $H+1$ different quantities for each (γ, k) pair under each scenario s:

$$\max_{i=0,\cdots,H} \left\{ \left(\sum_{n\in\mathcal{L}} \left\{ \sum_{t\in\mathcal{U}} \sum_{m\in\mathcal{T}\setminus\{t_0\}} \widetilde{y}^{k,n}_{\gamma,t,m}\lambda^k_{n,s} + \sum_{t\in\mathcal{T}} \sum_{m\in\mathcal{T}\setminus\{t_0\}} y^{k,n}_{\gamma,t,m,s}\lambda^k_{n,s} + \sum_{t\in\mathcal{U}\cup\{t_0\}} \widehat{y}^{k,n}_{\gamma,t,t_0}\lambda^k_{n,s} \right\} \right) \tilde{d}_{\gamma,k}(i) - \alpha^k_{\gamma,s}\, \tilde{d}_{\gamma,k}(i) - \beta^k_{\gamma,s}\, \tilde{d}_{\gamma,k}(i)^2 \right\}. \tag{22}$$

Choosing a large number of segment points for each (γ, k) pair models the support of the surgery duration distribution more precisely, thereby increasing the precision of the estimation made by the approximation problem (22) for (21); however, this comes at the cost of more computational time. We analyze how different choices of segment points affect the solution quality and computational time in Appendix D.

If we insert the approximation problem (22) into the optimization problem (20) derived in Proposition 2 under each scenario $s \in \mathcal{S}$, it yields an approximation called $\widetilde{\Psi}_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s})$ for the problem (20). In Theorem 1, we find an equivalent *mixed-integer linear program* for the approximation problem $\widetilde{\Psi}_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s})$ by leveraging McCormick-type constraints (McCormick 1976).

THEOREM 1 **(Scenario cut-generating problem)**. *Under each scenario $s \in \mathcal{S}$, the optimization problem (20) is approximated by the following mixed-integer linear program (MILP):*

$$\widetilde{\Psi}_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s}) = \max_{\boldsymbol{\tau}, \boldsymbol{\eta}, \boldsymbol{\lambda}} \ \chi_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s}; \boldsymbol{\tau_s}, \boldsymbol{\eta_s}, \boldsymbol{\lambda_s}) \tag{23a}$$

$$s.t. \ \sum_{i=0}^{H} \eta^k_{\gamma,i} \ = \ 1, \ \forall \gamma \in \Gamma, k \in \mathcal{K} \tag{23b}$$

$$\tau^k_{n,s,\gamma,i} - \lambda^k_{n,s} - \eta^k_{\gamma,i} \ \geq \ -1, \ \forall \gamma \in \Gamma, k \in \mathcal{K}, n \in \mathcal{L}, i = 0, \cdots, H \tag{23c}$$

$$\tau^k_{n,s,\gamma,i} - \lambda^k_{n,s} \ \leq \ 0, \ \ \forall \gamma \in \Gamma, k \in \mathcal{K}, n \in \mathcal{L}, i = 0, \cdots, H \tag{23d}$$

$$\tau^k_{n,s,\gamma,i} - \eta^k_{\gamma,i} \ \leq \ 0, \ \forall \gamma \in \Gamma, k \in \mathcal{K}, n \in \mathcal{L}, i = 0, \cdots, H \tag{23e}$$

$$\tau^k_{n,s,\gamma,i} \geq 0, \ \eta^k_{\gamma,i} \in \{0,1\}, \ 0 \leq \lambda^k_{n,s} \leq 1, \ \forall \gamma \in \Gamma, k \in \mathcal{K}, n \in \mathcal{L}, i = 0, \cdots, H \tag{23f}$$

where the objective function $\chi_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s}; \boldsymbol{\tau_s}, \boldsymbol{\eta_s}, \boldsymbol{\lambda_s})$ is defined for each scenario $s \in \mathcal{S}$ as follows:

$$\sum_{\gamma \in \Gamma} \sum_{k \in \mathcal{K}} \sum_{i=0}^{H} \Bigg\{ \sum_{n \in \mathcal{L}} \Big(\sum_{t \in \mathcal{U}} \sum_{m \in \mathcal{T} \setminus \{t_0\}} \widetilde{y}^{k,n}_{\gamma,t,m} + \sum_{t \in \mathcal{T}} \sum_{m \in \mathcal{T} \setminus \{t_0\}} y^{k,n}_{\gamma,t,m,s} + \sum_{t \in \mathcal{U} \cup \{t_0\}} \widehat{y}^{k,n}_{\gamma,t,t_0} \Big) \tilde{d}_{\gamma,k}(i) \, \tau^k_{n,s,\gamma,i} \tag{24}$$
$$- \alpha^k_{\gamma,s} \, \tilde{d}_{\gamma,k}(i) \, \eta^k_{\gamma,i} - \beta^k_{\gamma,s} \, \tilde{d}_{\gamma,k}(i)^2 \, \eta^k_{\gamma,i} \Bigg\} - \sum_{k \in \mathcal{K}} \sum_{n \in \mathcal{L}} V^k_n \, \lambda^k_{n,s}.$$

The proof of Theorem 1 is provided in Appendix A.3. The important implication of Theorem 1 is that it prevents having an embedded MILP model on the right-hand side of constraints (19b) by recognizing which scenario cuts must be added to replace the nonlinear constraints (19b). Using the results of Theorem 1, we can approximate the IMSDRO model (19a)-(19d) as follows:

$$\widetilde{Z}^{IMSDRO} = \min_{\boldsymbol{x}, \boldsymbol{y}, \widehat{\boldsymbol{y}}, \boldsymbol{q}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\beta}} \sum_{s \in \mathcal{S}} \pi_s \Bigg\{ \sum_{k \in \mathcal{K}} \Big(\sum_{m \in \mathcal{L}} q^k_{m,s} + \sum_{\gamma \in \Gamma} \Big(\mu_{\gamma,k} \, \alpha^k_{\gamma,s} + (\mu^2_{\gamma,k} + \sigma^2_{\gamma,k}) \, \beta^k_{\gamma,s} \Big) \Big) + \delta_s \Bigg\} \tag{25a}$$

$$\text{s.t.} \ \delta_s \ \geq \ \widetilde{\Psi}_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s}), \ \forall s \in \mathcal{S} \tag{25b}$$

$$(\boldsymbol{x_s}, \, \boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \, \boldsymbol{q_s}) \ \in \ \mathcal{R}_s, \ \forall s \in \mathcal{S} \tag{25c}$$

$$\delta_s \ \in \ \mathbb{R}, \ \boldsymbol{\alpha_s}, \ \boldsymbol{\beta_s} \ \in \ \mathbb{R}^{|\Gamma| \times |\mathcal{K}|}, \ \forall s \in \mathcal{S}. \tag{25d}$$

Remark. The minimization problem (25a)-(25d) is an approximation of the IMSDRO model (19a)-(19d). We shall call it the IMSDRO-APRX model. Although this model has a linear objective function with continuous and binary variables, due to the right-hand side of constraints (25b), which includes an embedded optimization problem (23a)-(23f), this model is not an MILP that is solvable by off-the-shelf MILP solvers (such as Gurobi and Cplex). In §4, we develop a constraint generation algorithm, which is based on iteratively generating constraints (25b) for each individual scenario $s \in \mathcal{S}$, as needed, to efficiently solve the IMSDRO-APRX model.

4. Constraint Generation Algorithm

We develop a new constraint generation algorithm for solving the IMSDRO-APRX model, which exploits the structure of the embedded MILP $\widetilde{\Psi}_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s})$ to generate effective scenario cuts. The main idea is explained as follows. The algorithm starts by solving the IMSDRO-APRX model without having any of the constraints (25b). At each iteration, it solves a *relaxed master problem* (RMP) to obtain a solution $(\boldsymbol{x_s}, \boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{q_s}, \delta_s, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s})$. Given this solution, it then solves what we call the *scenario cut-generating problem* $\widetilde{\Psi}_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s})$ or (23a)-(23f). If $\widehat{\boldsymbol{y}}, \boldsymbol{y_s}, \boldsymbol{\alpha_s}$, and $\boldsymbol{\beta_s}$ do not satisfy $\delta_s \geq \widetilde{\Psi}_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s})$, the scenario cut-generating problem returns scenario cuts in the form of (26b) back to RMP and the algorithm proceeds to next iteration. If $(\boldsymbol{x_s}, \boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{q_s}, \delta_s, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s})$ is optimal, the algorithm then terminates. The RMP at the J^{th} iteration is formulated as follows:

$$\widetilde{Z}^{RMP} = \min_{\boldsymbol{x},\boldsymbol{y},\widehat{\boldsymbol{y}},\boldsymbol{q},\boldsymbol{\delta},\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{s\in\mathcal{S}} \pi_s \left\{ \sum_{k\in\mathcal{K}} \left(\sum_{m\in\mathcal{L}} q^k_{m,s} + \sum_{\gamma\in\Gamma} \left(\mu_{\gamma,k}\, \alpha^k_{\gamma,s} + (\mu^2_{\gamma,k} + \sigma^2_{\gamma,k})\, \beta^k_{\gamma,s} \right) \right) + \delta_s \right\} \tag{26a}$$

$$\text{s.t. } \delta_s \geq \chi_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s}; \boldsymbol{\tau_s^{(j)}}, \boldsymbol{\eta_s^{(j)}}, \boldsymbol{\lambda_s^{(j)}}),\ \forall s \in \mathcal{S}, j = 1, \cdots, J-1 \tag{26b}$$

$$(\boldsymbol{x_s}, \boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{q_s}) \in \mathcal{R}_s,\ \forall s \in \mathcal{S} \tag{26c}$$

$$\delta_s \in \mathbb{R},\ \boldsymbol{\alpha_s}, \boldsymbol{\beta_s} \in \mathbb{R}^{|\Gamma|\times|\mathcal{K}|},\ \forall s \in \mathcal{S}, \tag{26d}$$

where the superscript j denotes all iterations up to the current iteration J, and the solution of the scenario cut-generating problem (23a)-(23f) at j-th iteration is denoted by $(\boldsymbol{\tau_s^{(j)}}, \boldsymbol{\eta_s^{(j)}}, \boldsymbol{\lambda_s^{(j)}})$ for each scenario s. The linear function $\chi_s(\boldsymbol{y_s}, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha_s}, \boldsymbol{\beta_s}; \boldsymbol{\tau_s^{(j)}}, \boldsymbol{\eta_s^{(j)}}, \boldsymbol{\lambda_s^{(j)}})$ in constraints (26b) is represented by the expression (24) with $(\boldsymbol{\tau_s}, \boldsymbol{\eta_s}, \boldsymbol{\lambda_s}) = (\boldsymbol{\tau_s^{(j)}}, \boldsymbol{\eta_s^{(j)}}, \boldsymbol{\lambda_s^{(j)}})$ for each scenario s. These scenario cuts are iteratively derived by passing the current solution $(\boldsymbol{x_s^{(J)}}, \boldsymbol{y_s^{(J)}}, \widehat{\boldsymbol{y}}^{(J)}, \boldsymbol{q_s^{(J)}}, \delta_s^{(J)}, \boldsymbol{\alpha_s^{(J)}}, \boldsymbol{\beta_s^{(J)}})$ of the RMP to the scenario cut-generating problem, and checking whether it satisfies (25b). If not, we add the corresponding scenario cut to the RMP for each scenario s. The details of the constraint generation algorithm are presented in Algorithm 1.

It is worth mentioning that the total number of scenario cuts that are being passed back to the RMP at each iteration is not necessarily equal to the total number of scenarios. Indeed, for only violated scenarios, Algorithm 1 passes back their corresponding scenario cuts to the RMP. In Appendix D, we compare multi-cut and single-cut versions of Algorithm 1 for the CAS problem.

Algorithm 1 Constraint Generation Algorithm for Solving the IMSDRO-APRX Model

Initialize iteration number $J = 0$, a positive tolerance ϵ, and also set the scenario parameters $Terminate(s) \leftarrow true$ for each scenario $s \in \mathcal{S}$.
Step I: **Solve the Relaxed Master Problem (RMP)**.
while ($\exists$ at least one $Terminate(s) \leftarrow true$ for a scenario $s \in \mathcal{S}$) **do**
    Set $J \leftarrow J + 1$, and $Terminate(s) \leftarrow false$ for each scenario $s \in \mathcal{S}$.
    Solve the RMP to get optimal solution $(\boldsymbol{x}_s^{(J)}, \boldsymbol{y}_s^{(J)}, \widehat{\boldsymbol{y}}^{(J)}, \boldsymbol{q}_s^{(J)}, \delta_s^{(J)}, \boldsymbol{\alpha}_s^{(J)}, \boldsymbol{\beta}_s^{(J)})$ for $\forall s \in \mathcal{S}$.
    Step II: **Cut-Generating Subroutine**.
    for each scenario $s \in \mathcal{S}$ **do**
        Solve the scenario cut-generating problem $\widetilde{\Psi}_s(\boldsymbol{y}_s^{(J)}, \widehat{\boldsymbol{y}}^{(J)}, \boldsymbol{\alpha}_s^{(J)}, \boldsymbol{\beta}_s^{(J)})$.
        Obtain the optimal solution $(\boldsymbol{\tau}_s^{(J)}, \boldsymbol{\eta}_s^{(J)}, \boldsymbol{\lambda}_s^{(J)})$.
        Step III: **Add Scenario Cuts to the RMP**.
        if $\delta_s^{(J)} < (1-\epsilon)\, \chi_s(\boldsymbol{y}_s^{(J)}, \widehat{\boldsymbol{y}}^{(J)}, \boldsymbol{\alpha}_s^{(J)}, \boldsymbol{\beta}_s^{(J)}; \boldsymbol{\tau}_s^{(J)}, \boldsymbol{\eta}_s^{(J)}, \boldsymbol{\lambda}_s^{(J)})$ **then**
            Add a scenario cut $\delta_s \geq \chi_s(\boldsymbol{y}_s, \widehat{\boldsymbol{y}}, \boldsymbol{\alpha}_s, \boldsymbol{\beta}_s; \boldsymbol{\tau}_s, \boldsymbol{\eta}_s, \boldsymbol{\lambda}_s)$ to the RMP for scenario $s \in \mathcal{S}$.
            Set $Terminate(s) \leftarrow true$ for scenario $s \in \mathcal{S}$.
Step IV: **Return the optimal policy**.
Return $(\boldsymbol{x}_s^{(J)}, \boldsymbol{y}_s^{(J)}, \widehat{\boldsymbol{y}}^{(J)}, \boldsymbol{q}_s^{(J)}, \delta_s^{(J)}, \boldsymbol{\alpha}_s^{(J)}, \boldsymbol{\beta}_s^{(J)})$ for each $s \in \mathcal{S}$ as the optimal policy.

THEOREM 2. *The constraint generation Algorithm 1 converges to an optimal scheduling policy for the IMSDRO-APRX model in a finite number of iterations.*

The proof of Theorem 2 is provided in Appendix A.4.

5. Data-Driven Rolling Horizon Procedure

Generally, the policy obtained from solving the IMSDRO-APRX model is not readily implementable for the real-world CAS problems because it is scenario dependent and does not allow for information gained over time to be used. Indeed, the critical limitation of scenario-based stochastic models is that their optimal policy is only valid for a limited set of scenarios. To resolve this issue, we develop a *data-driven* RHP to (i) make the scheduling policy *implementable* in practice, and (ii) evaluate the scheduling policy empirically on a rolling basis. It allows practitioners to make use of the latest data that is revealed as time progresses, and adjust their decisions in a rolling horizon fashion. By using this RHP, we dynamically observe the realization of the uncertain parameters in one period and update the scenario tree for the following periods.

In our data-driven RHP, we consider a set of data as the realization of uncertainties, which include the realized number of appointment requests, the realized surgery needs, and the realized surgery durations over an arrival horizon of $T = |\mathcal{T}|$ periods (days). Our approach for generating

this data set is explained in Appendix B. One member of this data set is then randomly drawn as a *sample path*, which provides a trajectory of the realization of the stochastic parameters over the planning horizon. The decision-making process in our data-driven RHP has three main steps: (i) the current period's stochastic quantities are realized, (ii) the scenario tree is updated, and (iii) the IMSDRO-APRX model is solved. Note that, in each period (day), only the stochastic quantities from the beginning of the planning horizon until the current period are observed, but the future uncertainty remains unknown to the model (thus, the RHP will not use any future information).

For implementation of a scheduling policy in the current period $t_0 \in \mathcal{T}$, given a scenario tree for the number of appointment requests over periods $t_0+1,\cdots,t_0+T-1$ and an ambiguity set for the surgery duration, we solve the T-stage IMSDRO-APRX model with an arrival horizon of T periods in which the uncertain parameters for period t_0 are known based on the realized path $\boldsymbol{\omega}$. We then implement the obtained optimal policy *only* for the current period t_0 and update the number of patients who need clinic and surgery appointments over $t_0+1,\cdots,t_0+T-1$ periods, as well as the remaining clinical and surgical capacities. We repeat this procedure, and "roll the patient arrival horizon forward one day" by adding a new period to the calendar at every step, so that at the following period t_0+1, the arrival horizon includes period t_0+2 to period t_0+T. Note that the length of the arrival horizon is always T periods (see Figure 5). By drawing enough realized sample paths, we can estimate the average clinical and surgical overtimes of surgeons over all sample paths. The data-driven RHP provides a framework that makes the decisions made by the IMSDRO approach implementable in practice.

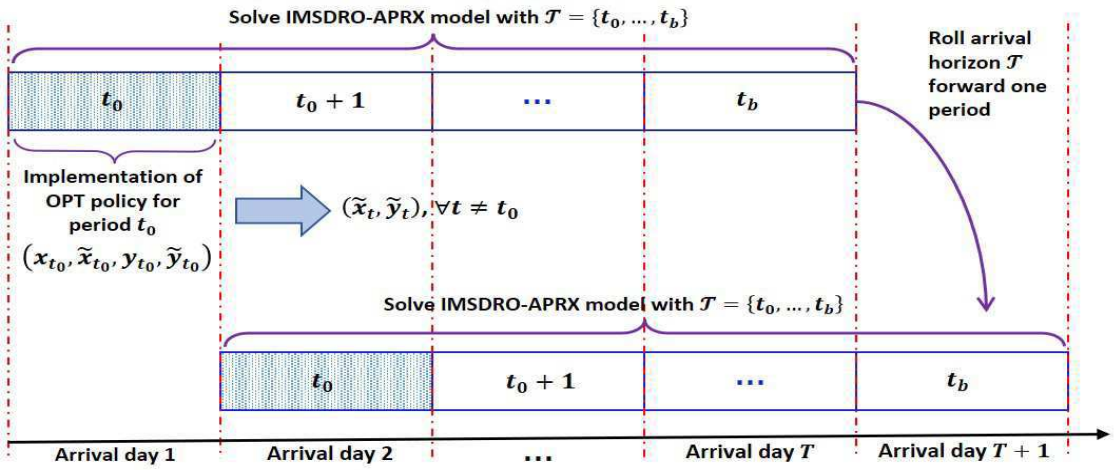


Figure 5 **The illustration of the data-driven rolling horizon procedure for solving the IMSDR-APRX model with an arrival horizon of $\mathcal{T} = \{t_0, \cdots, t_b\}$ on every stage (day) for the CRS problem.**

The details of the data-driven RHP are presented in Algorithm 2. Here, IMSDRO-APRX$(i, \boldsymbol{\omega})$ represents the problem in which the first period of the arrival horizon is day i, and its data is based

Algorithm 2 Data-driven Rolling Horizon Procedure

1: **Step I: Initialization**. Consider a sample path $\boldsymbol{\omega}$, which includes (i) the realized number of appointment requests ($\mathcal{D}_{\gamma,i}(\boldsymbol{\omega})$), (ii) the realizations of surgery durations ($d_{\gamma,k}(\boldsymbol{\omega})$), and (iii) the realized surgery needs ($\widetilde{z}^k_{\gamma,i}(\boldsymbol{\omega})$) for periods $i \in \mathcal{T} = \{t_0, \cdots, t_b\}$.

2: **Step II: Solve the IMSDRO-APRX model for each period** i.

3: **for** each arrival period $i \in \mathcal{T} = \{t_0, \cdots, t_b\}$ **do**

4: Consider a scenario tree with $|\mathcal{T}| - 1$ time periods.

5: Solve IMSDRO-APRX$(i, \boldsymbol{\omega})$ beginning with period i and parameters $\mathcal{D}_{\gamma,i}(\boldsymbol{\omega})$ and $\widetilde{z}^k_{\gamma,i}(\boldsymbol{\omega})$.

6: Given the following *implementable* decisions in period i for sample path $\boldsymbol{\omega}$,

$$x^{k,m}_{p,\gamma,i}(\boldsymbol{\omega}), \text{ where } k \in \mathcal{K}, \gamma \in \Gamma, m \in \mathcal{L}, p \in \mathcal{D}_{\gamma,i}(\boldsymbol{\omega}), \text{and } i \text{ is an arrival day},$$

$$\widetilde{x}^{k,i}_{p,\gamma,t}, \text{ where } k \in \mathcal{K}, \gamma \in \Gamma, t \in \mathcal{U}, p \in \widetilde{\mathcal{D}}_{\gamma,t}, \text{and } i \text{ is a clinic day},$$

$$\widehat{y}^{k,n}_{\gamma,t,i}(\boldsymbol{\omega}), \text{ where } k \in \mathcal{K}, \gamma \in \Gamma, t \in \mathcal{U} \cup \{i\}, \text{and } n \in \mathcal{L}, \text{and } i \text{ is a clinic day}$$

$$\widetilde{y}^{k,i}_{\gamma,t,m}, \text{ where } k \in \mathcal{K}, \gamma \in \Gamma, t \in \mathcal{U}, m \in \mathcal{U}, \text{and } i \text{ is a surgery day},$$

calculate clinic and surgery overtimes ($q^k_i(\boldsymbol{\omega})$, $o^k_i(\boldsymbol{\omega})$) in period i for each $k \in \mathcal{K}$ as follows:

$$q^k_i(\boldsymbol{\omega}) = \left(\sum_{\gamma \in \Gamma} \Big(\sum_{t \in \mathcal{U}} \sum_{p \in \widetilde{\mathcal{D}}_{\gamma,t}} c_\gamma \cdot \widetilde{x}^{k,i}_{p,\gamma,t} + \sum_{t \in \mathcal{T}} \sum_{p \in \mathcal{D}^s_{\gamma,t}} c_\gamma \cdot x^{k,i}_{p,\gamma,i}(\boldsymbol{\omega}) \Big) - U^k_i \right)^+$$

$$o^k_i(\boldsymbol{\omega}) = \left(\sum_{\gamma \in \Gamma} \Big(\sum_{t \in \mathcal{U}} \sum_{m \in \mathcal{T} \setminus \{i\}} d_{\gamma,k}(\boldsymbol{\omega}) \cdot \widetilde{y}^{k,i}_{\gamma,t,m} + \sum_{t \in \mathcal{U} \cup \{i\}} d_{\gamma,k}(\boldsymbol{\omega}) \cdot \widehat{y}^{k,i}_{\gamma,t,i}(\boldsymbol{\omega}) \Big) - V^k_i \right)^+.$$

7: Update the horizons $\mathcal{U}$, $\mathcal{T}$, and $\mathcal{L}$, and the following parameters for the next period $i+1$:

$$\widetilde{x}^{k,m-1}_{p,\gamma,t-1} \leftarrow \widetilde{x}^{k,m}_{p,\gamma,t}, \text{ for } t < i,\ m > i.$$

$$\widetilde{x}^{k,m-1}_{p,\gamma,t-1} \leftarrow x^{k,m}_{p,\gamma,t}(\boldsymbol{\omega}), \text{ for } t = i,\ m > i.$$

$$\widetilde{y}^{k,n-1}_{\gamma,t-1,m-1} \leftarrow \widehat{y}^{k,n}_{\gamma,t,i}(\boldsymbol{\omega}), \text{ for } t < i,\ m = i,\ n > i.$$

$$\widetilde{y}^{k,n-1}_{\gamma,t-1,m-1} \leftarrow \widetilde{y}^{k,n}_{\gamma,t,m}, \text{ for } t < i,\ m \le i,\ n > i.$$

8: **Step III: Calculate the objective function for the sample path** $\boldsymbol{\omega}$.

$$Q(\boldsymbol{\omega}) = \sum_{i=t_0}^{t_b} Q_i(\boldsymbol{\omega}), \text{ where } Q_i(\boldsymbol{\omega}) = \sum_{k \in \mathcal{K}} \big(q^k_i(\boldsymbol{\omega}) + o^k_i(\boldsymbol{\omega}) \big).$$

on the sample path $\boldsymbol{\omega}$. This sample path provides the related data for the new arrival horizon $\mathcal{T}$. Note that the next period, $i+1$, becomes the first period after rolling forward one period, and Algorithm 1 is used to solve IMSDRO-APRX$(i, \boldsymbol{\omega})$ again. After the T-stage IMSDRO-APRX$(i, \boldsymbol{\omega})$ model is solved, an i-th stage decision is implemented and new information is obtained, we roll forward (i.e., shift the time window) to solve another T-stage IMSDRO-APRX model with the uncertainty determined by the implemented i-th stage decision and by an observation of sample

path $\boldsymbol{\omega}$ only for i-th stage. Note that the scenario tree that is considered in step 4 can be updated at each iteration i using the scenario generation and reduction algorithm illustrated in Appendix B, to capture any possible seasonality or trend in the data.

6. Case Study: Empirical Results and Managerial Insights

We populate our models and algorithms based on appointment scheduling data from a highly specialized surgical clinic of a partner hospital. We provide numerical results to evaluate the performance of our approach compared to current practice of the surgical clinic and obtain managerial insights.

6.1. Experimental Setup

Appointment requests are received throughout the day either from other units within the same or nearby hospitals or from remote healthcare facilities. Each patient request asks for a clinic consultation appointment with one of the surgeons. Some patients may require a surgical procedure; this is determined during the patient's clinic consultation appointment. Appointment requests include information on (i) the referral type, i.e., either local or remote referral, and (ii) the indication of disease, which can be one of the 12 possible medical conditions.

The surgical clinic is currently scheduling appointment requests with the surgeon who has the earliest clinic consultation availability. However, this policy has resulted in long wait time to surgery, which is particularly troubling for patients with acute conditions. It has also been observed that some surgeons end their surgical day early on some days and very late on other days. Our models are designed to guarantee that all patients will be offered a clinic consultation and a surgical appointment (if needed) within a time window that is safe for them to wait.

We consider five priority classes; class 1 includes the most acute and urgent conditions, whereas class 5 is assigned to patients who only need a clinic appointment for follow-up/consult or those who do not need a surgery in the near future. Each class corresponds to a maximum wait time to surgery WTS_γ, except class 5 that does not require a surgery. Clinic to surgery gap CSG_γ is another parameter that depends on patient class. The minimum wait time for clinic visits (WTC_γ), however, depends only on the referral type. For local referrals (i.e., the patient is physically at the hospital or in the same region), WTC_γ is zero, whereas for remote referrals, we assume WTC_γ is five business days from the day the request is received to give the patient at least one week to make travel arrangements. Table 2 shows these values in days for different patient priority classes.

The probability of requiring a surgery depends on the patient class. The surgical clinic under study performs about 400 surgeries per month. This corresponds to a rate of about 60 appointment requests per business day. Each clinic appointment takes 15 minutes in length; that is, clinic days are divided into 15-minute time slots and each clinic appointment takes one slot. The surgical clinic

Class	WTC	CSG	WTS
1	0 or 5	0	5
2	0 or 5	1	7
3	0 or 5	2	10
4	0 or 5	3	18
5	0 or 5	NA	NA

Table 2 The values of Wait Time to Clinic (WTC), Clinic to Surgery Gap (CSG), and Wait Time to Surgery (WTS) in terms of number of days.

has eight surgeons (i.e., $|\mathcal{K}| = 8$). These 8 surgeons are divided into two teams taking alternating turns between the clinic and the operating room (OR) from one day to another (i.e., on a given day, four surgeons are seeing patients in the clinic and four surgeons are performing surgeries in the OR). Hence, each surgeon separately maintains both a clinical and a surgical calendar. The specialized surgical unit in our case study have access to dedicated ORs as well as to a number of swing ORs that they can use, if needed. Thus, operating room capacity can be flexible, if needed.

Surgery duration $d_{\gamma,k}$ depends on both patient class and the specific surgeon who performs the operation. Recall that patient class γ is a tuple of two elements: referral type and indication of disease. We assume surgery duration is independent of referral type but depends on the indication of disease. Therefore, given 8 surgeons and 12 disease indications, we consider 96 indication-surgeon pairs to define surgery duration. For each indication-surgeon combination, we employ the empirical surgery mean, standard deviation and support of past surgeries to construct the ambiguity sets. Appendix B elaborates on our approach for both generating the ambiguity set for the surgery duration and the scenario tree for the number of patient referrals.

Our analyses aim to address the following questions in §6.2, §6.3, and §6.4, respectively. (i) How effective is our stochastic-robust policy in terms of clinical and surgical access times and surgeons' overtime compared to different benchmark policies? (ii) What is the trade-off between meeting the clinical and surgical access targets and incurring overtime? (iii) How do the key parameters of our stochastic-robust policy impact its performance?

6.2. Assessing the Performance of Different Scheduling Policies

We evaluate our stochastic-robust policy against three benchmark policies in terms of the clinical and surgical overtimes (i.e., objective function) as well as the clinical and surgical access times using the RHP proposed in §5. These four policies are summarized below.

- **Stochastic-robust policy**. This policy is derived by solving the IMSDRO-APRX model (25a)-(25d), in which the uncertainty in the surgery duration is modeled by an ambiguity set, and the uncertainty in the number of appointment requests is modeled by a scenario tree.
- **Stochastic policy**. This policy is obtained by solving the MS-MIP model (1)-(14), in which the uncertainty in the number of appointment requests is modeled by a scenario tree, and the surgery durations are set to their empirical mean values.

- **Deterministic policy**. This policy is obtained by solving the deterministic version of the MS-MIP model (1)-(14), in which both the surgery durations and the number of appointment requests are set to their empirical mean values.
- **Current policy**. This heuristic policy mimics the current/existing policy used by the surgical clinic. As discussed above, for each appointment request, this policy suggests the surgeon with the earliest clinic appointment availability. On the clinic appointment date, if the patient requires a surgery, it offers the earliest surgical appointment with the same surgeon. Therefore, this policy does not incorporate the access targets and thus does not stratify patients by class.

To further evaluate our results, we employ two other instances A and B for the CAS problem in addition to our case study (base case). The differences between the case study and these instances are (i) the number of surgeons is 4 (instance A) and 10 (instance B) as opposed to 8 (case study), and (ii) the number of scenarios for arrival is 20 (instance A) and 10 (instance B) as opposed to 14 (case study). Other parameters are similar to the case study.

Evaluation of objective function values (overtime). For this analysis, 60 sample paths of a length 5 working weekdays for the arrival horizon are randomly drawn that include the realized (i) number of appointment referrals, (ii) surgery needs, and (iii) surgery durations. For each sample path, the RHP (Algorithm 2) is implemented for each of the above policies by solving their models with a 5-day arrival horizon (i.e., $T = |\mathcal{T}| = 5$) and rolling the horizon forward to cover 10 days. We use this roll-out window for demonstration only. Our approach is not limited to a 10-day roll-out window and one may continue rolling the horizon forward for as long as needed.

We calculate the mean ($\tilde{Z}$) and standard deviation ($\tilde{\sigma}$) of the objective functions (clinical and surgical overtimes) over all sample paths as the output of the data-driven RHP to assess these policies. Note that we start our analyses with long-run average system state and further use a 10-day burn-in period so that our results and findings are not affected by the initial system status.

The empirical results for our case study and test instances A and B are reported in Table 3. The optimal objective values (Z^*) are calculated by solving the IMSDRO-APRX, MS-MIP and deterministic models (without using the RHP). Moreover, $\tilde{Z}_{\max}$, $\tilde{Z}$, and $\tilde{\sigma}$ are the maximum, the mean and the standard deviation, respectively, of the true objective function values obtained by using the data-driven RHP on the 60 sample paths described above. The *out-of-sample stability error* is then calculated as follows:

$$\text{Out-of-sample Stability Error} = \frac{|\text{Mean objective value } (\tilde{Z}) - \text{Optimal objective value } (Z^*)|}{\text{Optimal objective value } (Z^*)} \times 100\%.$$

We do not include the current policy in this analysis. This is because the current policy does not incorporate the access target constraints in its decision, and there is no optimal objective value for this policy. We do, however, incorporate it in the clinical and surgical access times analyses.

Statistics	Stochastic-Robust Policy	Stochastic Policy	Deterministic Policy
The case study:			
Optimal objective value (Z^*)	4,451	4,325	4,311
Mean objective value ($\tilde{Z}$)	4,317	4,525	4,788
Max objective value ($\tilde{Z}_{\max}$)	5,285	5,389	5,528
Standard deviation ($\tilde{\sigma}$)	378	416	445
Out-of-sample error	3.10%	4.42%	9.96%
Test instance A:			
Optimal objective value (Z^*)	6,375	6,211	6,125
Mean objective value ($\tilde{Z}$)	6,254	6,348	6,633
Max objective value ($\tilde{Z}_{\max}$)	6,835	6,992	7,025
Standard deviation ($\tilde{\sigma}$)	656	712	695
Out-of-sample error	1.93%	2.16%	7.66%
Test instance B:			
Optimal objective value (Z^*)	2,274	2,175	2,165
Mean objective value ($\tilde{Z}$)	2,199	2,257	2,379
Max objective value ($\tilde{Z}_{\max}$)	2,868	2,912	3,012
Standard deviation ($\tilde{\sigma}$)	295	318	342
Out-of-sample error	2.99%	3.63%	9.00%

Table 3 The out-of-sample stability analysis of the stochastic-robust, stochastic, and deterministic policies. Numbers are the total clinical and surgical overtime aggregated over all 8 surgeons and the 5-day horizon.

As reported in Table 3, the *optimal objective function value* of the stochastic-robust policy is slightly larger (2.9% more overtime) than the stochastic policy. This is because the stochastic-robust policy accounts for the uncertainty of surgery duration by considering a whole range of possibilities for the probability distribution of the surgery duration and optimizes the *worst-case performance* as opposed to the stochastic policy that only considers the surgery duration mean and optimizes the *mean performance*. Also, the deterministic policy has the smallest optimal objective function value, because it optimizes for a single scenario in which the number of patient referrals and surgery durations are both set to their empirical means.

However, when we simulate and exploit these scheduling polices for the 60 sample paths, we observe from Table 3 that the stochastic-robust policy yields both the smallest *mean objective function value* (i.e. the lowest surgeon mean overtime) and the smallest variability around the overtime (i.e. the lowest standard deviation) relative to the stochastic and deterministic policies. In particular, the mean objective function value (i.e., overtime) of the stochastic-robust policy is 4.6% and 9.8% less than the stochastic and deterministic policies, respectively. This observation illustrates that even though the stochastic-robust policy hedges against the worst-case and makes *more conservative* decision strategies, the worst-case situation may not necessarily occur for all possible surgeries in practice. This can subsequently leads to a smaller overtime mean and variability for the stochastic-robust policy relative to the other policies. Limiting the variability is of paramount importance in healthcare operations as having consistent performance allows the surgical clinic to better plan for and manage their resources. On the other hand, the stochastic policy

makes scheduling decisions based on the surgery duration mean scenario; thus, it results in a *more compact* schedule compared to the stochastic-robust policy. However, the surgery duration mean scenario does not necessarily happen for all possible surgeries in practice, which subsequently leads to more overtime relative to the stochastic-robust policy. The deterministic policy has the poorest performance with respect to both mean objective value and variability. This is because it deploys a policy that was optimized for only one single scenario of patient arrival mean and surgery duration mean, for many sample paths. Results are similar for the test instances A and B.

Moreover, the *out-of-sample stability* (Kaut and Wallace 2003) guarantees that the mean objective function value $\tilde{Z}$ obtained from implementing the optimal scheduling policy by using the data-driven RHP is approximately the same as the optimal objective value Z^* of the optimization models. As reported in Table 3, the stochastic-robust policy has the *smallest* out-of-sample stability error (3.10%) compared to the stochastic policy (4.42%) and the deterministic policy (9.96%) in the case study and similarly in the two test instances A and B. The small difference between the mean objective function value and the optimal objective value further confirms the validity and reliability of the IMSDRO-APRX model as a reasonable approximation method.

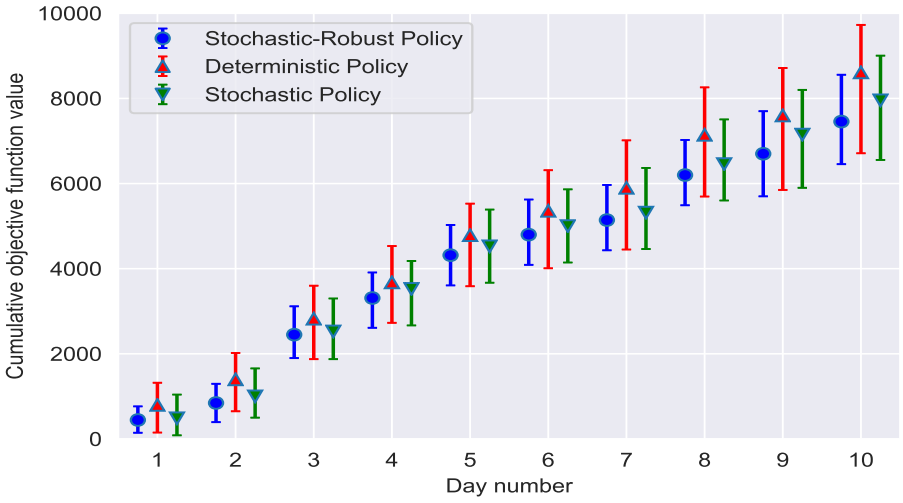


Figure 6 The comparison of the stochastic-robust, stochastic, and deterministic policies over 10 business days implemented by the RHP in terms of mean, 25%-QT and 75%-QT cumulative overtimes for the case study.

To further assess the stochastic-robust, stochastic, and deterministic policies, Figure 6 graphs the cumulative clinic and surgery overtime of these policies (aggregated over all 8 surgeons) over 10 business days for the case study. We observe that the stochastic-robust policy is better than both the stochastic and deterministic policies in terms of cumulative overtime mean and variability, and this is consistent through the end of the horizon. In particular, by day 10, the stochastic-robust policy incurs a statistically smaller cumulative overtime compared to the stochastic policy (p-value

< 0.001) and the deterministic policy (p-value $= 0.006$). Thus, the stochastic-robust policy offers an excellent performance with a much lower variability, which is critical for implementation in everyday practice. Results were similar for test instances A and B and are not shown here.

Evaluation of clinical and surgical access times. We next compare the scheduling policies in terms of the clinical/surgical access times. To do so, we define a new performance metric for earliness or tardiness that is "the number of days between the scheduled clinic and surgery appointment date and the maximum wait time target date." We call it the *access measure with respect to the target*. Since there are different patient classes with various access delay target windows, this metric helps us better understand how much earlier or later, with respect to the maximum wait target, the policies schedule the appointments. The negative (positive) value implies how much earlier (later) a policy schedules the clinic and surgery appointment with regard to the maximum wait target. In this analysis, we use the RHP for implementing the stochastic-robust, stochastic, and deterministic policies obtained by solving the IMSDRO-APRX, MS-MIP and deterministic models, respectively. Note that the current policy is also included in this analysis. We calculate the access earliness or tardiness measure with respect to the target for each patient whose referral is received within the roll-out window for the case study as well as the test instances A and B. We compute the mean, worst-case, and standard deviation (SD) of these access measures.

Statistics	Clinical Access Measure with respect to the Target				Surgical Access Measure with respect to the Target			
	Stochastic-Robust Policy	Stochastic Policy	Deterministic Policy	Current Policy	Stochastic-Robust Policy	Stochastic Policy	Deterministic Policy	Current Policy
The case study:								
Mean	−2.42	−2.67	−2.97	−4.55	−2.78	−2.83	−2.97	4.65
Worst-case	0	0	0	0	0	0	0	7
SD	0.89	1.12	1.21	1.69	1.07	1.15	1.18	1.47
Test instance A:								
Mean	−1.78	−1.85	−1.97	−3.25	−1.85	−1.93	−2.25	8.85
Worst-case	0	0	0	0	0	0	0	11
SD	0.76	0.89	0.91	1.49	0.76	0.93	0.98	1.28
Test instance B:								
Mean	−2.98	−3.12	−3.25	−4.96	−3.52	−3.75	−3.98	3.85
Worst-case	0	0	0	0	0	0	0	5
SD	1.16	1.21	1.47	2.01	1.28	1.55	1.62	1.85

Table 4 The statistical performance comparison of scheduling policies in terms of mean, worst-case, and SD for the clinical and surgical access measures with respect to the maximum wait target (in days).

Table 4 demonstrates empirical results of comparing various policies in terms of clinical/surgical access measures with respect to the target. We summarize the system performance by averaging across all classes. We observe that the stochastic-robust, stochastic, and deterministic policies all yield negative clinical and surgical access delay measures. This is because the three policies are able to grant the predefined priority-based access targets to all patients. However, the current policy performs quite differently. While it performs better than the other three policies in terms

of providing early access to a clinic consultation appointment, it often fails to provide the crucial surgical appointment within the safe time window, thus compromising health outcomes especially for acute patients. This is because, unlike the model-based policies, the current policy does not consider wait time targets and the uncertainty about the number of appointment requests, probability of surgery need, and surgery duration. It simply assigns the patient to the surgeon with the earliest clinic appointment availability.

We next graph the daily mean of surgical access measures with respect to the maximum wait target by the day of referral arrival over 10 days for the case study in Figure 7. Again, we summarize the system performance by averaging over all patient classes.

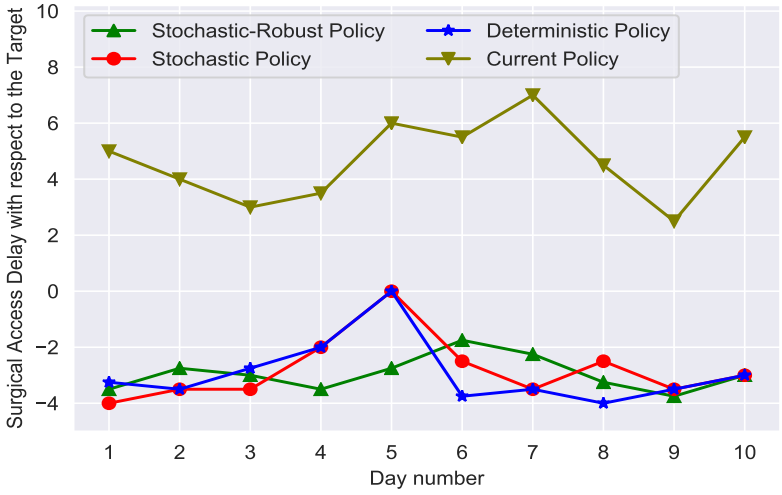


Figure 7 The comparison of the surgical access measure with respect to the maximum wait target (averaged across all classes) by the day of referral arrival obtained by different policies over the 10-day horizon by the RHP.

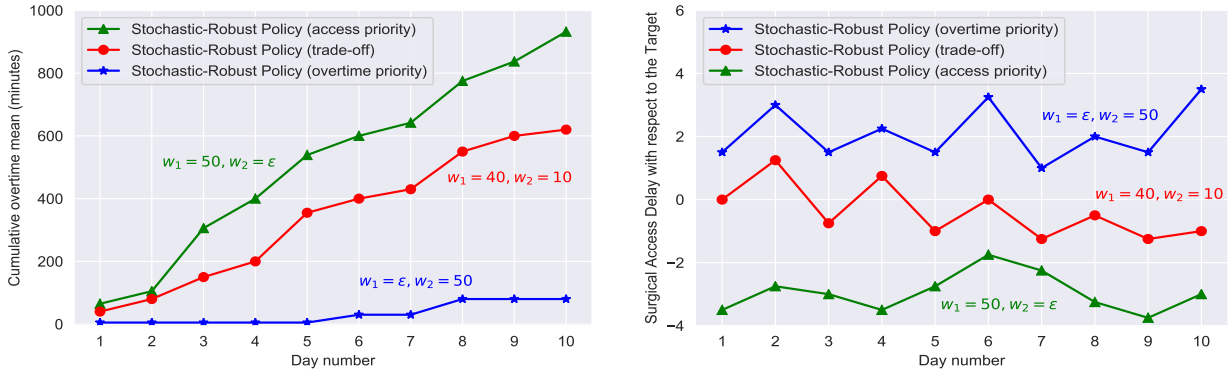
Figure 7 shows that the current policy consistently yields significantly higher wait time to surgery compared to all other polices. This implies a major drawback of the current policy that patients often need to wait a long time to receive a surgical visit, which can deteriorate their condition. The other three policies, however, uniformly provide on-time (often early) access to surgical procedure. It is worth noting that robustifying surgery duration in the stochastic-robust policy adds little in terms of computational complexity compared to the stochastic policy.

In conclusion, our coordinated stochastic-robust policy obtains the lowest overtime along with the smallest variability while respecting both clinical and surgical access limits (through imposing the clinical and surgical access constraints (2)-(6)) so that it finds safe clinical and (as needed) surgical appointments within the target window.

6.3. Access Delay versus Overtime Trade-off Analysis

As emphasized in §3, the IMSDRO-APRX model ensures *patient-centered care* by providing 100% service level in terms of granting access targets to all patients while optimizing the overtime. In Appendix C, we however formulate an alternative model that establishes a trade-off between meeting access delay targets and incurring overtime. This is a bi-objective optimization model, which minimizes (i) the expected penalty due to not meeting clinical and surgical access delay targets (weighted by w_1), and (ii) the maximum expected penalty due to incurring overtime (weighted by w_2). Here, we investigate this balance through implementing the RHP by the stochastic-robust policy obtained from solving this alternative model.

We calculate the cumulative overtime mean of each surgeon, and the mean of surgical access measures with respect to the maximum wait target by each day. We consider three possible scenarios: (i) the "stochastic-robust policy (access priority)", which puts a large penalty on not meeting the access delay targets ($w_1 = 50, w_2 = \epsilon$), (ii) the "stochastic-robust policy (overtime priority)", which puts a large penalty on the overtime incurred ($w_1 = \epsilon, w_2 = 50$), and (iii) the "stochastic-robust policy (trade-off)", which aims at striking a balance between these two objectives ($w_1 = 40, w_2 = 10$). Figure 8 demonstrates the results for the case study.



(a) Cumulative overtime mean of each surgeon. (b) Surgical access delay mean with respect to the target.

Figure 8 The illustration of trade-off between not meeting access delay targets and incurring overtime for the case study. The cumulative surgeon overtime mean and surgical access delay mean with respect to the target are obtained by three different stochastic-robust policies over a 10-day roll-out window by the RHP for the case study.

As seen in Figure 8, while the stochastic-robust policy (access priority) has the highest cumulative overtime mean per surgeon, it provides patients with the fastest surgical access compared to the other two stochastic-robust polices. It is also worth noting that the stochastic-robust policy (access priority) yields a surgical access measure mean of -2.1 days with respect to the maximum wait target (i.e., 2.1 days earlier than the deadline). This is about 55% better than the current policy, which yields a surgical access measure mean of 4.65 days (see Table 4). This occurs because unlike the current policy, which is a heuristic, the stochastic-robust policy (access priority) solves an optimization model to make appointment decisions.

6.4. Sensitivity Analysis Results

In this section, we evaluate (i) the in-sample stability, (ii) the importance of modeling the probability of needing a surgery, and (iii) the importance of number of days in an arrival horizon. In Appendix D, we provide additional analyses for our models/algorithms.

In-sample stability analysis. There are two essential criteria, (i) in-sample and (ii) out-of-sample stability, to evaluate the efficiency of a scenario tree construction method (Kaut and Wallace 2003). In §6.2, we show that the out-of-sample errors are 3.10% (case study), 1.93% (test instance A) and 2.99% (test instance B). Here, we evaluate the in-sample stability. If $|J|$ scenario trees ξ_j, $j \in J$ are generated by using our scenario tree construction method (see Appendix B), and we then solve the IMSDRO-APRX model for each of these scenario trees to calculate the optimal decision vector x_j^* with objective function $f(x_j^*, \xi_j)$ for scenario tree $j \in J$, then *in-sample stability* implies $f(x_j^*, \xi_j) \approx f(x_u^*, \xi_u)$, $\forall j, u \in J$. To evaluate the in-sample stability, we generate different scenario fans with 100 scenarios for the number of appointment requests by the Latin Hypercube Sampling method. Then, a forward scenario construction approach is applied to construct a scenario tree by using different values for the parameter $\boldsymbol{\zeta}_{rel}$ (see Appendix B). Recall that $\boldsymbol{\zeta}_{rel}$ represents a reduction scale of the scenario tree compared with the scenario fan. For each instance with different scenario trees and various values of $\boldsymbol{\zeta}_{rel}$, the *in-sample stability error* is calculated by:

$$\text{In-sample Stability Error} = \frac{\text{Max of objective values} - \text{Min of objective values}}{\text{Average of objective values}} \times 100\%.$$

The number of scenarios decreases as the value of $\boldsymbol{\zeta}_{rel}$ increases.

Test Instance	$\boldsymbol{\zeta}_{rel} = 0.8$			$\boldsymbol{\zeta}_{rel} = 0.7$		
	# of Scenarios	Objective fun.	in-sample error	# of Scenarios	Objective fun.	in-sample error
The case study	7	4,625	4.64%	13	4,545	3.29%
	8	4,561		14	4,451	
	10	4,474		15	4,478	
	11	4,687		17	4,578	
Test instance A	9	6,625	3.03%	17	6,488	2.21%
	10	6,737		18	6,536	
	11	6,536		20	6,421	
	13	6,585		22	6,485	
Test instance B	10	2,265	1.99%	17	2,254	1.85%
	11	2,235		19	2,289	
	13	2,280		20	2,247	
	14	2,280		21	2,265	

Table 5 The in-sample stability analysis for the scenario tree construction approach.

Table 5 shows the empirical results of the in-sample stability analysis. The difference between the objective function values with different scenario trees is smaller (i.e., smaller in-sample error) when $\boldsymbol{\zeta}_{rel} = 0.7$. More importantly, the lack of any substantial difference between the optimal objective function values indicates a very good in-sample stability of our scenario tree construction approach.

Importance of modeling the probability of needing a surgery. In the case study, about one in three appointment requests will end up requesting a surgical procedure after the clinic visit. Hence, the expected probability of needing surgery is 0.33, based on which we draw the sample paths. To assess the probability of surgery need, we implement the RHP for the case study under two scenarios: (i) considering the surgery need, and (ii) ignoring the possibility of surgery need. We then investigate how the stochastic-robust policy performs under these two scenarios by calculating the mean and variability of the cumulative overtime values over 10 days for one surgeon. Figure 9 illustrates the results for the comparison of considering versus ignoring the surgery need.

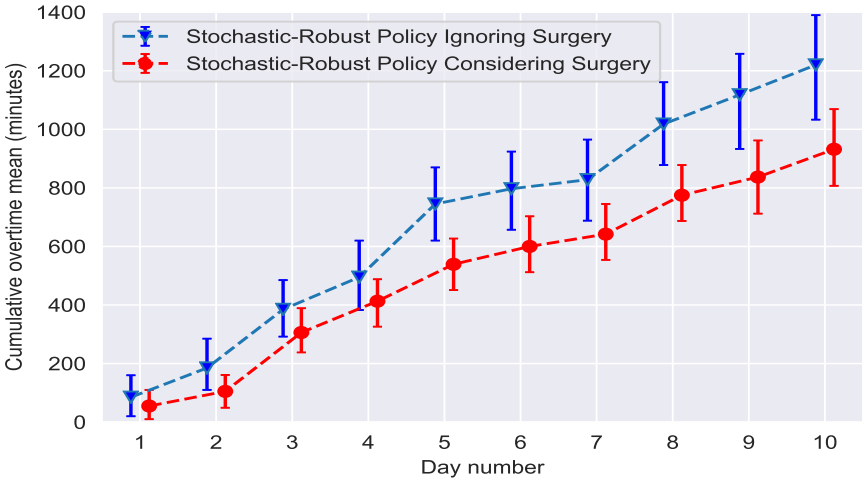


Figure 9 **A comparison of cumulative overtime means for the stochastic-robust policy in the case study when the probability of needing a surgery is considered vs. ignored.**

In Figure 9, the blue curve only considers the already booked surgeries over the next periods; it ignores the likelihood of future possible surgeries. On the other hand, the red curve does account for the probability of future surgeries when making scheduling decisions. As seen in Figure 9, the stochastic-robust policy performs significantly better (i.e., less overtime) when the probability of future surgeries are taken into account. In particular, the stochastic-robust policy considering surgery probabilities yields a daily overtime mean of 90 minutes for a surgeon, which is about 26% better (less overtime) than the stochastic-robust policy ignoring the probability of surgeries with a daily overtime mean of 122 minutes for a surgeon.

It is worth noting that we defined the "regular time" for each surgeon as 6 hours per day and called any OR time beyond that "overtime" as a conservative approach to scheduling ORs. This definition is not meant to reflect the regular and overtime shifts that the hospital uses to determine provider compensation. We employed this conservative definition of regular time to encourage our models to limit the OR time to the ideal level of 6 hours per day. The models will then incur a penalty for going over this ideal level. Thus, a mean overtime of 90 minutes observed in Figure 9

implies that the surgeons will work, on average, 7.5 hours per day under the stochastic-robust policy that considers future surgery needs. To sum up, we demonstrate that the idea of care coordination can help to achieve less overtime by considering the uncertainty around future surgery needs.

Importance of number of days in the arrival horizon. In the case study, we consider a 5-day arrival horizon. In other words, we account for the uncertainty around the number of appointment requests and the surgery durations of the next 5 days when making appointment decisions. In this analysis, we investigate how the number of days considered in the arrival horizon affects the quality of the stochastic-robust policy. Figure 10 demonstrates the performance of the stochastic-robust policy by computing the cumulative overtime mean for a surgeon obtained from solving the IMSDRO-APRX model by using the RHP for 10 days in our case study. We consider three different arrival horizons with $T = |\mathcal{T}| = 3$, 5 and 7 days.

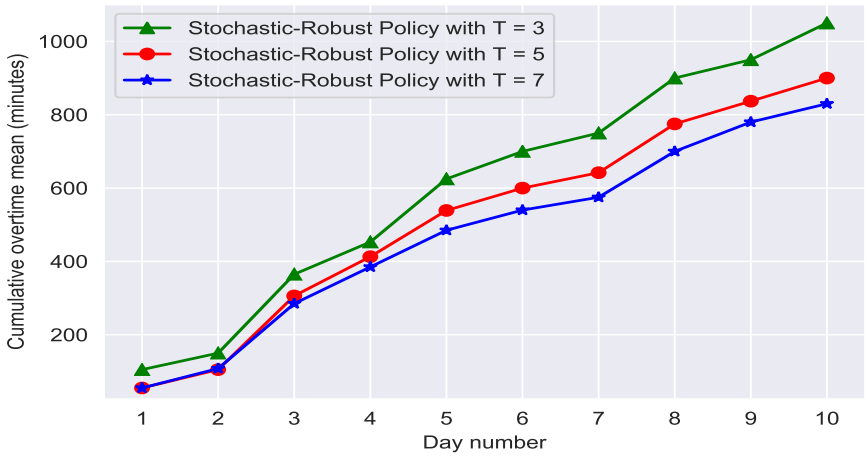


Figure 10 Importance of the number of days for the arrival horizon $\mathcal{T}$: the performance of the stochastic-robust policies with $T = 3$, 5, and 7 days in terms of cumulative overtime mean per surgeon over 10 days for the case study obtained by solving the IMSDRO-APRX model by using the RHP.

From Figure 10, we observe that as the number of days in the arrival horizon increases, the performance of the stochastic-robust policy gradually improves (i.e., less overtime is incurred) as we roll forward in the arrival horizon. The longer the arrival horizon, the less myopic the policy. This is because longer-term uncertainty about the number of appointment requests and surgery duration is taken into account when the stochastic-robust policy makes the clinical/surgical decisions. As seen in Figure 10, reducing the length of arrival horizon from 5 to 3 days increases the overtime mean per surgeon by about 15 minutes on day 10. However, increasing the length of arrival horizon from 5 to 7 days only reduces the overtime mean per surgeon by 6 minutes on day 10 (the cumulative overtime means are 1050, 900 and 840 minutes by day 10, for $T = 3, 5$ and 7, respectively, which are equivalent to 105, 90 and 84 minutes per surgeon per day). This suggests that while in general including *additional days* in the arrival horizon is helpful, increasing the horizon from 5 to 7 days has little benefit and may not worth the additional computational burden.

7. Practical Implications and Insights

In §6, we demonstrated the application of our IMSDRO approach to coordinate clinic and surgery visits in a highly specialized surgical unit. We showed that our model can provide access to surgery within a safe time frame, especially for acute patients who will most suffer from a long wait time, while minimizing the overtime. We summarize the insights and practical implications below.

First and foremost, surgical divisions that offer surgical procedures after a clinic consultation appointment should consider leveraging optimization algorithms that coordinate the clinic and surgery appointments when scheduling new appointments. Simple heuristic scheduling protocols, such as scheduling new appointment requests with the surgeon who has the earliest availability (i.e., the "current policy" in our case study), often result in prolonged wait times for patients with acute conditions like cancer. The lengthy wait times to receive a surgical procedure may result in adverse events and poor patient outcomes. In contrast, through minimizing the overtime, our proposed coordinated stochastic-robust policy achieved both clinical and surgical access targets, which were stratified into five classes based on patients' acuity level. Simple heuristics may allocate clinic and surgery appointment dates for a patient several days beyond the acceptable wait time target windows. This is not clinically safe for patients and may lead to additional complications. The success of our algorithmic, optimization-based method indicated that it is not always effective to offer the earliest available appointment slot to a new patient as commonly done in current practice. If the patient has an acute condition, consideration of the likelihood of surgery and availability of providers is key to ensure timely access to surgery.

Moreover, even though our model considered overtime to meet the priority-based access to care targets, our empirical results showed that the mean overtime per surgeon is around 90 minutes. It should be noted that we defined the regular time for surgeons as 6 hours per day in the case study. Thus, the mean overtime of 90 minutes per day means that a typical day for surgeons lasts 7.5 hours, on average. We also demonstrated that our stochastic-robust policy achieves the lowest overtime and the smallest variability among the four policy we investigated, while respecting both clinical and surgical access limits. The average workday of 7.5 hours, together with granting access targets of the stochastic-robust policy, confirm that the appointment scheduling plans obtained from our IMSDRO approach are feasible and implementable in practice. Our optimization models provide generality over a broader range of operation systems and parameters than most heuristics, which do not readily extend to new settings. Our analytic approach allows the decision maker to modify the parameters of the system to find an acceptable optimal policy. For instance, if the amount of overtime suggested by our model is not desirable, the decision maker can relax the priority-based access targets to reduce the required overtime. If new surgeons are hired or new procedures are offered, the model can be easily extended to accommodate the new conditions.

Modeling care coordination in our coordinated stochastic-robust policy results in better utilization of scarce resources, including surgeon time and operating rooms. We saw in Figure 9 that the policy that takes uncertain future surgeries into consideration outperforms the policy that ignores the uncertainty of the need for surgery (26% less overtime). Moreover, we demonstrated in multiple ways that the stochastic-robust policy achieves much lower variability in surgeon overtime and patient access time compared to alternative policies. This is extremely important in healthcare setting since avoiding extreme scenarios and achieving a reliable performance will allow the hospital management to better control patient flow and manage their resources and processes.

Our research promotes patient-centered care by stratifying patients into different priority classes based on what is known about the patient at the time the patient referral is received (e.g., the indication of disease), which are then translated into appropriate and safe maximum wait time targets. Surgical divisions should also take the uncertainty in appointment request arrival, surgical demand, and surgery durations into account when scheduling clinic consultation and surgery appointments. Our models provide a creative way to do so using data that are commonly available in the patient's electronic health records and the clinic's datasets, and do not rely on assumptions on the probability distribution of surgeries. Further, the proposed data-driven rolling horizon procedure introduces an innovative way of making use of the latest data that is revealed as time progresses, and adjusting the decisions in practice for stochastic optimization problems.

Furthermore, we provided two optimization models derived by our IMSDRO approach. The focus of the first (main) one was on ensuring patient-centered care by providing 100% service level in terms of meeting access delay targets to all patients while minimizing the surgeon overtime. The second optimization model considered two competing objectives, namely, meeting access delay targets and incurring overtime. As illustrated in Figure 8, this model allows decision makers to establish a trade-off between providing timely access to care to patients and asking surgeons to work overtime hours.

Our coordinated stochastic-robust policy improves the surgical access times by about 160%, on average, compared to the current policy (see Table 4). Intuitively, this is because our method takes into account the wait time target windows as well as various inherent sources of uncertainty, including the number of appointment requests, probability of surgery need, and surgery duration while coordinating clinic and surgery appointments. Also, the current policy ignores the valuable indication of disease that is available in the patient's electronic health records when a new appointment request is received. Unlike the current policy that operates based on a first-come first-serve idea, our model often defers the surgery of low-priority patients in order to preserve the near future capacity to serve high-priority patients that may arrive later. This approach helps meet the desired service level with minimum overtime.

8. Conclusion, Limitations, and Future Research

In this paper, we studied a new class of appointment scheduling problems called the "coordinated clinic and surgery appointment scheduling (CAS)" in which patients are stratified into different classes, with limits on the allowable access delay from request to appointment dates. We introduced the concept of care coordination in the sense of setting appointments for pairs of sequential clinic and (if needed) surgery visits that together achieve timely access to care. Methodologically speaking, our integrated multi-stage stochastic and distributionally robust optimization (IMSDRO) is the first optimization approach that can jointly incorporate different types of uncertainty in the number of patient appointment requests by a scenario tree, and in surgery durations by a moment-based ambiguity set for distributional robustness. Using the special structure of the CAS problem, we proposed a constraint generation algorithm for efficiently solving this problem. We then developed a new data-driven rolling horizon procedure to implement the decisions made by the IMSDRO approach in practice. This allows healthcare practitioners to make efficient use of data that are obtained as time unfolds, and so adjust their decisions in a rolling horizon framework. In a sense, our methods can be applied in an online (or real-time) fashion. We tested the validity of our models in a case study of scheduling clinic consultation and surgery appointments, and demonstrated that a significant improvement could be achieved by employing our proposed policies. We provided a number of practical insights from our empirical analyses as well.

This study has a few limitations. First, in our models, we do not consider patient no-shows and cancellations as well as the potential seasonality in demand as they rarely happen in a highly-specialized surgical suite. Patient preferences are also not part of our models and algorithms. Clearly, in many healthcare environments, the patient can prioritize the selection of the provider with whom they feel most comfortable. Our scope is; however, limited to the important class of environments in which the patients typically accept the provider offering the earliest access. Second, the allocation of resources, including operating rooms, to surgeons is not the main focus in our paper. We also assumed that each patient's surgical need follows a Bernoulli distribution with a success probability that only depends on the patient class. Alternative approaches to modeling this uncertainty can be studied in future research. Finally, given that a tractable system state can be defined, approximate or robust dynamic programming approaches may be used to solve the CAS problem. These ideas could be promising future research directions.

Acknowledgment. The authors thank the departmental editor Professor Edward Anderson, the anonymous senior editor, and the anonymous referees for their constructive and detailed comments, which have helped us significantly improve both the content and the exposition of this paper.

References

Ahmadi-Javid, Amir, Zahra Jalali, Kenneth J Klassen. 2017. Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research* **258**(1) 3–34.

Alden, Jeffrey M, Robert L Smith. 1992. Rolling horizon procedures in nonhomogeneous markov decision processes. *Operations Research* **40**(3-supplement-2) S183–S194.

Bandi, Chaithanya, Diwakar Gupta. 2020. Operating room staffing and scheduling. *Manufacturing & Service Operations Management* **22**(5) 958–974.

Bertsekas, D. 2005. Rollout algorithms for constrained dynamic programming. *Lab. for Information and Decision Systems Report* **2646**.

Bertsekas, Dimitri P, David A Castanon. 1999. Rollout algorithms for stochastic scheduling problems. *Journal of Heuristics* **5**(1) 89–108.

Bertsekas, Dimitri P, John N Tsitsiklis, Cynara Wu. 1997. Rollout algorithms for combinatorial optimization. *Journal of Heuristics* **3**(3) 245–262.

Browne, Cameron B, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, Simon Colton. 2012. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games* **4**(1) 1–43.

Deglise-Hawkinson, Jivan, Jonathan E Helm, Todd Huschka, David L Kaufman, Mark P Van Oyen. 2018. A capacity allocation planning model for integrated care and access management. *Production and operations management* **27**(12) 2270–2290.

Denton, Brian, Diwakar Gupta. 2003. A sequential bounding approach for optimal appointment scheduling. *IIE transactions* **35**(11) 1003–1016.

Denton, Brian, James Viapiano, Andrea Vogl. 2007. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health care management science* **10**(1) 13–24.

Diamant, Adam, Joseph Milner, Fayez Quereshy. 2018. Dynamic patient scheduling for multi-appointment health care programs. *Production and Operations Management* **27**(1) 58–79.

Dupačová, Jitka. 1995. Multistage stochastic programs: The state-of-the-art and selected bibliography. *Kybernetika* **31**(2) 151–174.

Erdogan, S Ayca, Brian Denton. 2013. Dynamic appointment scheduling of a stochastic server with uncertain demand. *INFORMS Journal on Computing* **25**(1) 116–132.

Gelly, Sylvain, Levente Kocsis, Marc Schoenauer, Michele Sebag, David Silver, Csaba Szepesvári, Olivier Teytaud. 2012. The grand challenge of computer go: Monte carlo tree search and extensions. *Communications of the ACM* **55**(3) 106–113.

Gocgun, Yasin, Martin L Puterman. 2014. Dynamic scheduling with due dates and time windows: an application to chemotherapy patient appointment booking. *Health care management science* **17**(1) 60–76.

Grant, Benjamin, Itai Gurvich, R Kannan Mutharasan, Jan A Van Mieghem. 2021. Optimal dynamic appointment scheduling of base and surge capacity. *Manufacturing & Service Operations Management* .

Gupta, Diwakar, Brian Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE transactions* **40**(9) 800–819.

Gupta, Diwakar, Lei Wang. 2008. Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research* **56**(3) 576–592.

He, Shuangchi, Melvyn Sim, Meilin Zhang. 2019. Data-driven patient scheduling in emergency departments: A hybrid robust-stochastic approach. *Management Science* **65**(9) 4123–4140.

Hernández-Lerma, O, JB Lasserre. 1990. Error bounds for rolling horizon policies in discrete-time markov control processes. *IEEE Transactions on Automatic Control* **35**(10) 1118–1124.

Jiang, Ruiwei, Siqian Shen, Yiling Zhang. 2017. Integer programming approaches for appointment scheduling with random no-shows and service durations. *Operations Research* **65**(6) 1638–1656.

Jung, Kyung Sung, Michael Pinedo, Chelliah Sriskandarajah, Vikram Tiwari. 2019. Scheduling elective surgeries with emergency patients at shared operating rooms. *Production and Operations Management* **28**(6) 1407–1430.

Kaplan, Garry, Marianne Hamilton Lopez, J Michael McGinnis. 2015. Transforming health care scheduling and access: Getting to now. *Washington DC: Institute of Medicine* .

Kaut, Michal, Stein W Wallace. 2003. Evaluation of scenario-generation methods for stochastic programming.

Kazemian, Pooyan, Mustafa Y Sir, Mark P Van Oyen, Jenna K Lovely, David W Larson, Kalyan S Pasupathy. 2017. Coordinating clinic and surgery appointments to meet access service levels for elective surgery. *Journal of biomedical informatics* **66** 105–115.

Keyvanshokooh, Esmaeil, Cong Shi, Mark P Van Oyen. 2021. Online advance scheduling with overtime: A primal-dual approach. *Manufacturing & Service Operations Management* **23**(1) 246–266.

Kong, Qingxia, Chung-Yee Lee, Chung-Piaw Teo, Zhichao Zheng. 2013. Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operations research* **61**(3) 711–726.

Lemay, Brian, Amy Cohn, Marina Epelman, Stephen Gorga. 2017. New methods for resolving conflicting requests with examples from medical residency scheduling. *Production and Operations Management* **26**(9) 1778–1793.

Liu, Nan, Stacey R Finkelstein, Margaret E Kruk, David Rosenthal. 2017. When waiting to see a doctor is less irritating: Understanding patient preferences and choice behavior in appointment scheduling. *Management Science* **64**(5) 1975–1996.

Liu, Nan, Van-Anh Truong, Xinshang Wang, Brett Anderson. 2019a. Integrated scheduling and capacity planning with considerations for patients' length-of-stays. *Production and Operations Management* .

Liu, Nan, Peter M van de Ven, Bo Zhang. 2019b. Managing appointment booking under customer choices. *Management Science* **65**(9) 4280–4298.

Liu, Nan, Serhan Ziya, Vidyadhar G Kulkarni. 2010. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing & Service Operations Management* **12**(2) 347–364.

Macario, Alex. 2010. What does one minute of operating room time cost? *Journal of clinical anesthesia* **22**(4) 233–236.

Mak, Ho-Yin, Ying Rong, Jiawei Zhang. 2014. Appointment scheduling with limited distributional information. *Management Science* **61**(2) 316–334.

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

Mandelbaum, Avishai, Petar Momčilović, Nikolaos Trichakis, Sarah Kadish, Ryan Leib, Craig A Bunnell. 2020. Data-driven appointment-scheduling under uncertainty: The case of an infusion unit in a cancer center. *Management Science* **66**(1) 243–270.

Markit, IHS. 2017. The complexities of physician supply and demand: Projections from 2015 to 2030 .

May, Jerrold H, William E Spangler, David P Strum, Luis G Vargas. 2011. The surgical scheduling problem: Current research and future opportunities. *Production and Operations Management* **20**(3) 392–405.

Mays, Glen P, Sharla A Smith, Richard C Ingram, Laura J Racster, Cynthia D Lamberth, Emma S Lovely. 2009. Public health delivery systems: evidence, uncertainty, and emerging research needs. *American Journal of Preventive Medicine* **36**(3) 256–265.

McCormick, Garth P. 1976. Computability of global solutions to factorable nonconvex programs: Part i—convex underestimating problems. *Mathematical programming* **10**(1) 147–175.

Morrice, Douglas J, Jonathan F Bard, Luci K Leykum, Susan Noorily. 2018. The impact of a patient-centered surgical home implementation on preoperative processes in outpatient surgery. *IISE Transactions on Healthcare Systems Engineering* **8**(2) 155–166.

Munos, Rémi, et al. 2014. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning* **7**(1) 1–129.

Oudhoff, JP, DRM Timmermans, DL Knol, AB Bijnen, G Van der Wal. 2007. Waiting for elective general surgery: impact on health related quality of life and psychosocial consequences. *BMC Public Health* **7**(1) 164.

Parvin, Hoda, Shervin Beygi, Jonathan E Helm, Peter S Larson, Mark P Van Oyen. 2018. Distribution of medication considering information, transshipment, and clustering: Malaria in malawi. *Production and Operations Management* **27**(4) 774–797.

Patrick, Jonathan, Martin L Puterman, Maurice Queyranne. 2008. Dynamic multipriority patient scheduling for a diagnostic resource. *Operations research* **56**(6) 1507–1525.

Saure, Antoine, Jonathan Patrick, Scott Tyldesley, Martin L Puterman. 2012. Dynamic multi-appointment patient scheduling for radiation therapy. *European Journal of Operational Research* **223**(2) 573–584.

Turkcan, Ayten, Bo Zeng, Mark Lawley. 2012. Chemotherapy operations planning and scheduling. *IIE Transactions on Healthcare Systems Engineering* **2**(1) 31–49.

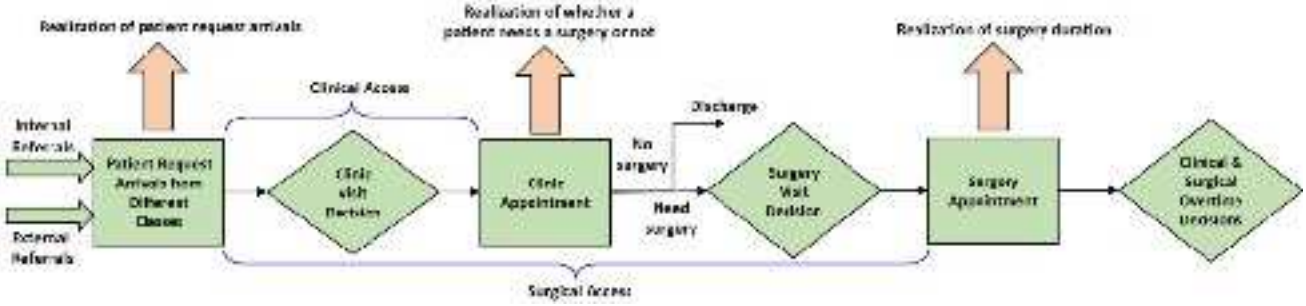
Wang, Dongyang, Douglas J Morrice, Kumar Muthuraman, Jonathan F Bard, Luci K Leykum, Susan H Noorily. 2018. Coordinated scheduling for a multi-server network in outpatient pre-operative care. *Production and Operations Management* **27**(3) 458–479.

Yang, XQ, CJ Goh. 1997. A method for convex curve approximation. *European Journal of Operational Research* **97**(1) 205–212.

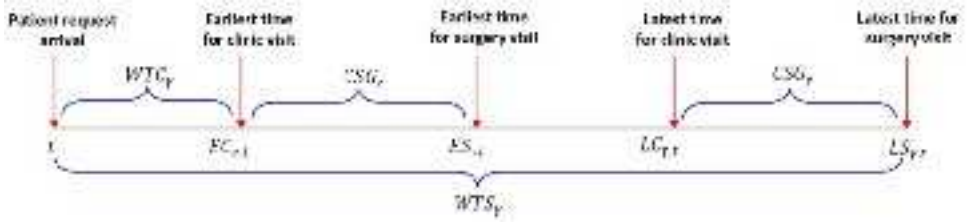
Yu, Siyun, Vidyadhar G Kulkarni, Vinayak Deshpande. 2020. Appointment scheduling for a health care facility with series patients. *Production and Operations Management* **29**(2) 388–409.

Zacharias, Christos, Tallys Yunes. 2020. Multimodularity in the stochastic appointment scheduling problem with discrete arrival epochs. *Management science* **66**(2) 744–763.

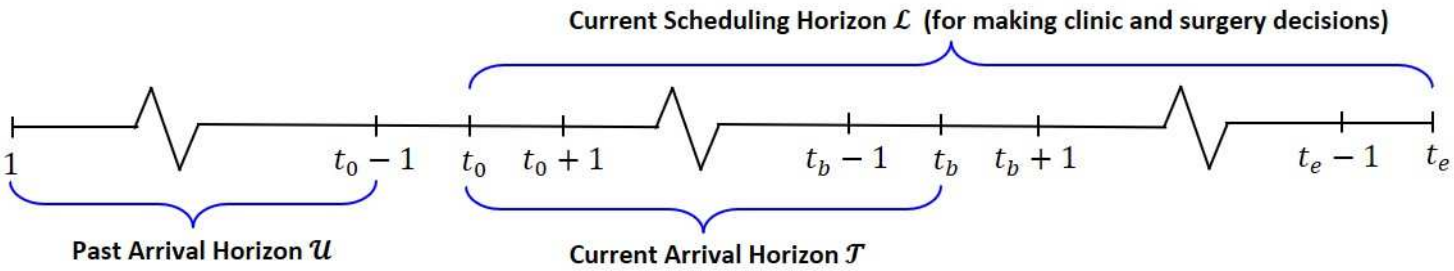
Zhou, Yun, Mahmut Parlar, Vedat Verter, Shannon Fraser. 2021. Surgical scheduling with constrained patient waiting times. *Production and Operations Management* .



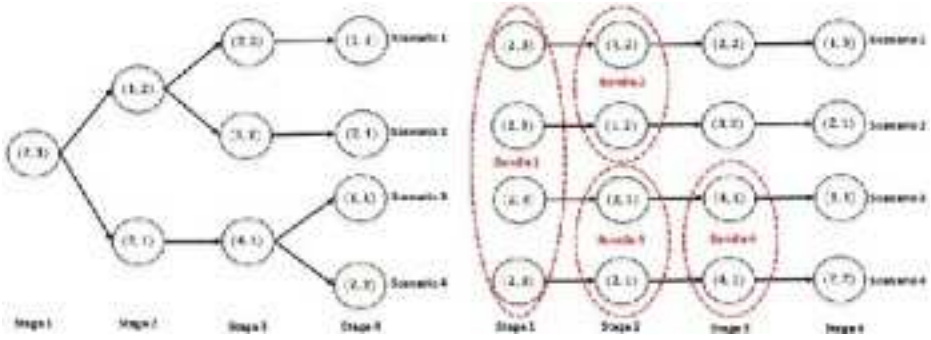
poms_13628_f1.jpg



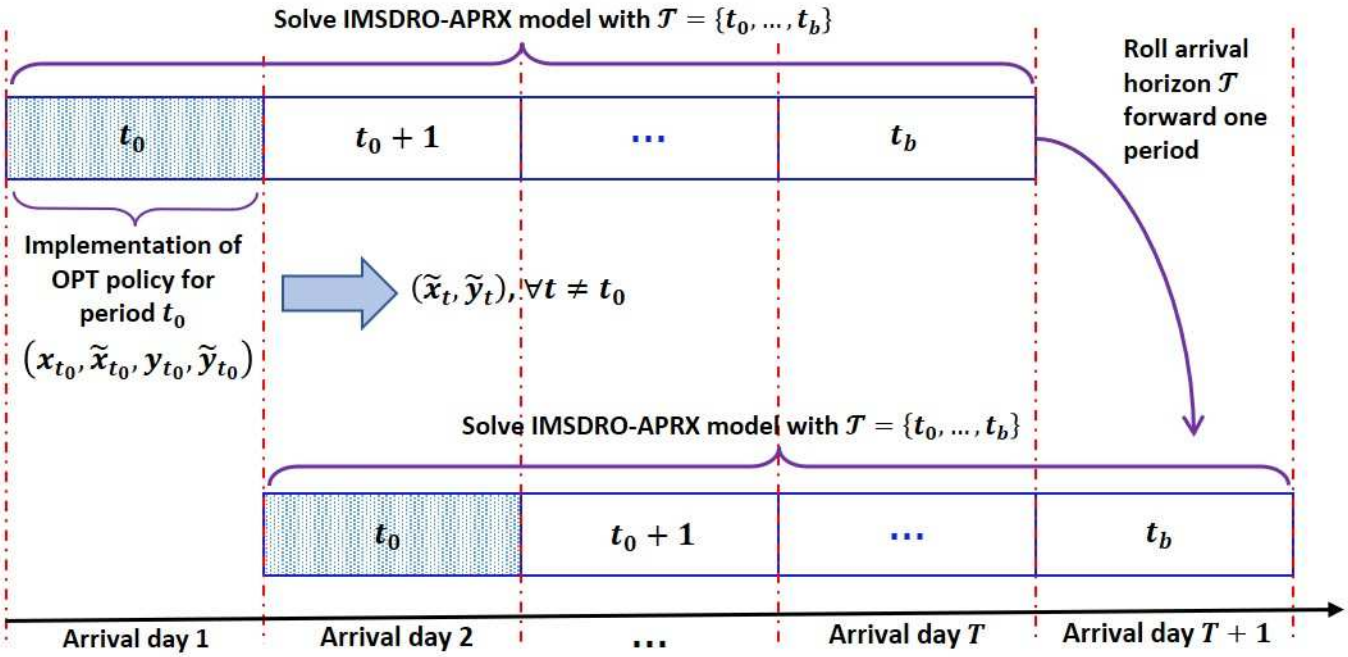
poms_13628_f2.jpg



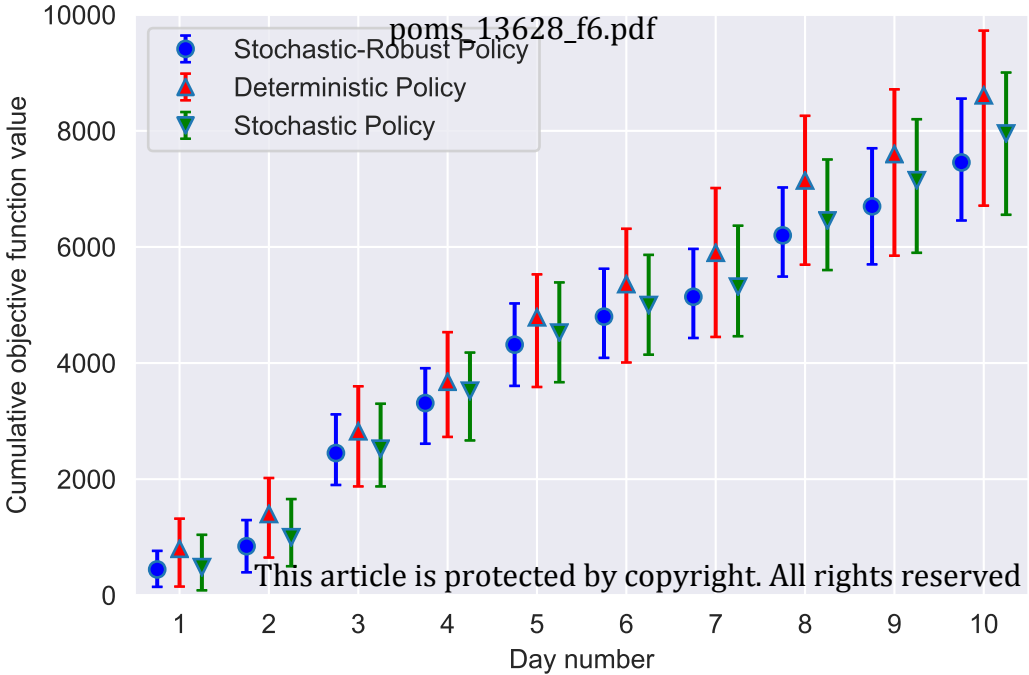
poms_13628_f3.jpg

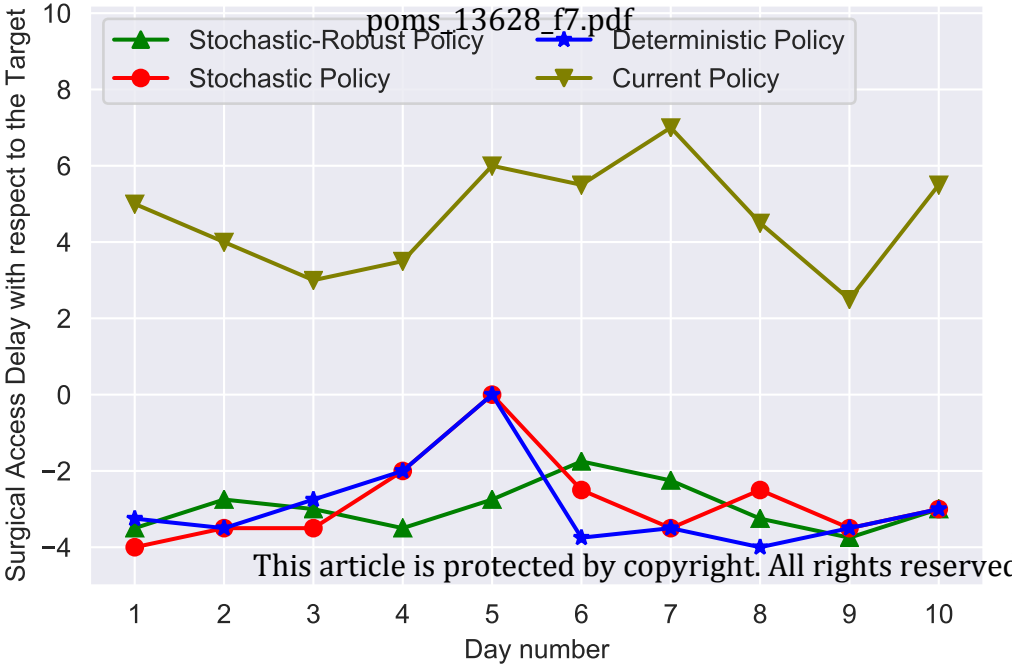


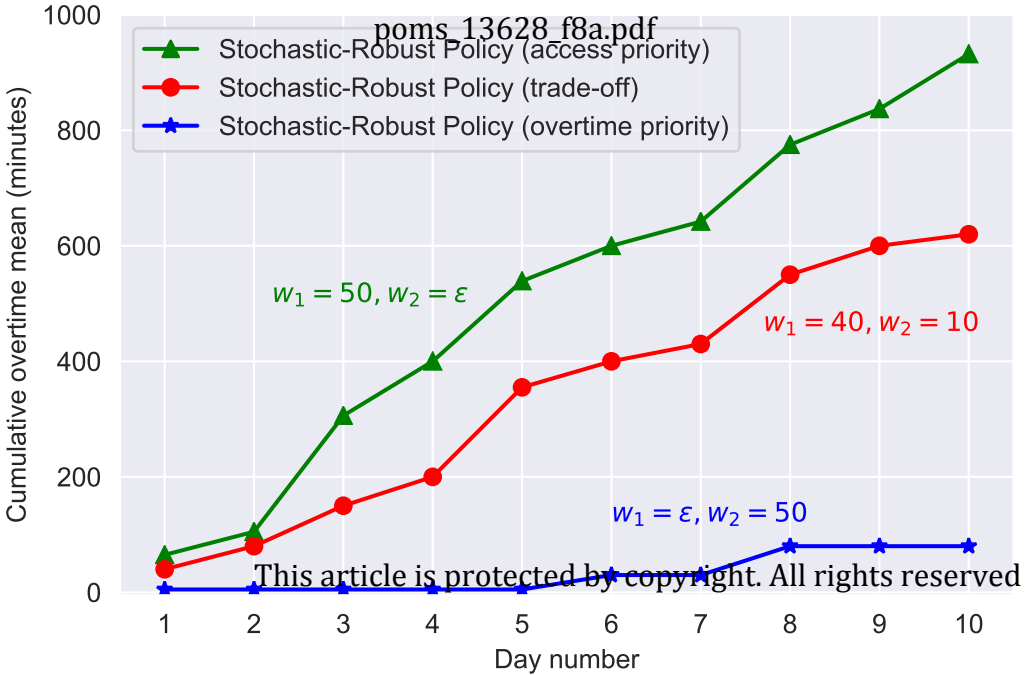
poms_13628_f4.jpg

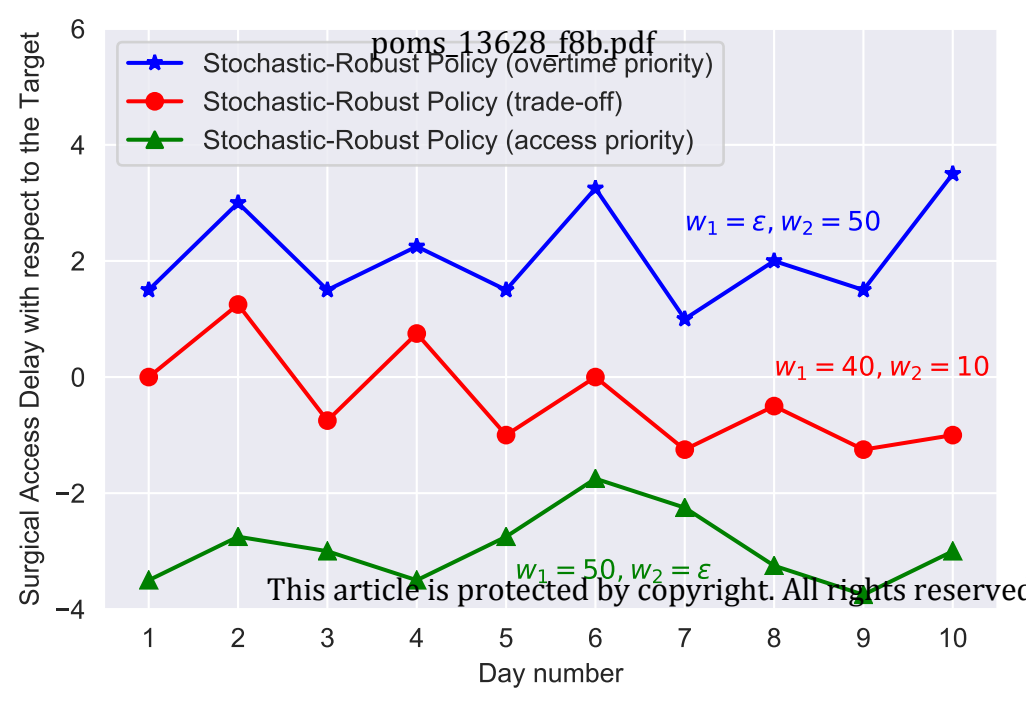


poms_13628_f5.jpg




poms_13628_f7.pdf
Stochastic-Robust Policy
Deterministic Policy
Stochastic Policy
Current Policy
Surgical Access Delay with respect to the Target
10
8
6
4
2
0
−2
−4
1
2
3
4
5
6
7
8
9
10
Day number
This article is protected by copyright. All rights reserved




poms_13628_f8b.pdf
Stochastic-Robust Policy (overtime priority)
Stochastic-Robust Policy (trade-off)
Stochastic-Robust Policy (access priority)
$w_1 = \varepsilon, w_2 = 50$
$w_1 = 40, w_2 = 10$
$w_1 = 50, w_2 = \varepsilon$
Surgical Access Delay with respect to the Target
Day number
6
4
2
0
−2
−4
1
2
3
4
5
6
7
8
9
10
This article is protected by copyright. All rights reserved