Received: 5 November 2019 | Revised: 5 June 2020 | Accepted: 20 June 2020 | Published on: 29 July 2020

DOI: 10.1002/nav.21933

RESEARCH ARTICLE

WILEY

Using stochastic programming to solve an outpatient appointment scheduling problem with random service and arrival times

Karmel S. Shehadeh[1] | **Amy E. M. Cohn**[2] | **Ruiwei Jiang**[2]

[1]Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, Pennsylvania

[2]Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan

Correspondence
Karmel S. Shehadeh, Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA.
Email: kshehadeh@lehigh.edu, ksheha@umich.edu

Funding information
Agency for Healthcare Research and Quality, Grant/Award Number: P30HS024385. National Science Foundation, Grant/Award Number: CMMI-1662774. University of Michigan Center for Healthcare Engineering and Patient Safety.

Abstract

We study a stochastic outpatient appointment scheduling problem (SOASP) in which we need to design a schedule and an adaptive rescheduling (i.e., resequencing or declining) policy for a set of patients. Each patient has a known type and associated probability distributions of random service duration and random arrival time. Finding a provably optimal solution to this problem requires solving a multistage stochastic mixed-integer program (MSMIP) with a schedule optimization problem solved at each stage, determining the optimal rescheduling policy over the various random service durations and arrival times. In recognition that this MSMIP is intractable, we first consider a two-stage model (TSM) that relaxes the nonanticipativity constraints of MSMIP and so yields a lower bound. Second, we derive a set of valid inequalities to strengthen and improve the solvability of the TSM formulation. Third, we obtain an upper bound for the MSMIP by solving the TSM under the feasible (and easily implementable) *appointment order* (AO) *policy*, which requires that patients are served in the order of their scheduled appointments, independent of their actual arrival times. Fourth, we propose a Monte Carlo approach to evaluate the relative gap between the MSMIP upper and lower bounds. Finally, in a series of numerical experiments, we show that these two bounds are very close in a wide range of SOASP instances, demonstrating the near-optimality of the AO policy. We also identify parameter settings that result in a large gap in between these two bounds. Accordingly, we propose an alternative policy based on neighbor-swapping. We demonstrate that this alternative policy leads to a much tighter upper bound and significantly shrinks the gap.

KEYWORDS

appointment scheduling, mixed-integer programming, Monte Carlo optimization, OR in health services, stochastic arrival, stochastic programming

1 | INTRODUCTION

Outpatient clinics (OPCs) are quickly growing as a central component of the health care system. OPCs offer a variety of medical services and benefits such as avoiding inpatient hospitalization, high patient safety, and low costs of care (Ahmadi-Javid, Jalali, & Klassen, 2017). However, they also introduce new challenges for appointment planning and scheduling primarily due to the heterogeneity and variability in patient characteristics. Multiple (heterogeneous) patient types may require different service times. Even patients of the same type may vary in the required service time, and such variation is hard to predict in advance, which may contribute to patient delay and provider overtime (Ahmadi-Javid et al., 2017; Cayirli, Veral, & Rosen, 2008; Cayirli & Yang, 2014; Gupta & Denton, 2008).

Another common phenomenon among OPC patients (as well as the customers in many other service industries with an appointment scheduling system) is the lack of punctuality, which introduces additional complexity into clinical operations and increases operational costs (Alexopoulos, Goldsman, Fontanesi, Kopald, & Wilson, 2008; Cayirli,

 © 2020 Wiley Periodicals LLC

Veral, & Rosen, 2006; Deceuninck, Fiems, & De Vuyst, 2018; Fetter & Thompson, 1966; Glowacka et al., 2017; Jiang, Tang, & Yan, 2019; Klassen & Yoogalingam, 2014; Williams, Chambers, Dada, McLeod, & Ulatowski, 2014). Late arrivals often create provider idling (and thus poor utilization of the clinic resources) and, along with the random service durations, may increase the waiting time of the subsequent appointments and provider overtime through the propagation of the delay, if there is no buffer in the schedule to absorb it. Very early arrivals, on the other hand, may require the provider to make challenging queuing decisions as whether to serve them ahead of the scheduled time. For example, serving early arrivals (arriving outside their scheduled turn and time) can result in service delay of the next on-time patient (Deceuninck et al., 2018; Glowacka et al., 2017; Samorani & Ganguly, 2016).

Accounting for uncertainty and the possibility of resequencing (and/or rescheduling a late patient to another day) in the scheduling decision process has the potential of mitigating these adverse outcomes. Most of the existing literature on stochastic appointment scheduling (since Welch & Bailey, 1952), however, focuses on addressing the variability in service duration only. Within the limited literature accounting for the stochastic arrival time, most studies ignore the sequencing decision and only consider homogeneous patients (i.e., all service durations follow the same distribution) or heterogeneous patients in a sequence found by heuristic approaches (Deceuninck et al., 2018; Klassen & Yoogalingam, 2014; Zacharias & Yunes, 2018). In addition, most studies adopt the appointment order (AO) policy, that is, the provider will serve the patients in their scheduled AO (as opposed to the actual arriving order) and decline a patient if he/she arrives later than the scheduled time by a certain amount of time (termed the *grace period*). Although this policy is easy to implement and effective under many circumstances, to the best of our knowledge, no literature has investigated its optimality (or suboptimality) in a rigorous manner.

In this paper, we study a stochastic outpatient appointment scheduling problem (SOASP) for OPC scheduling under stochastic arrival times and service durations. We consider for an OPC manager who needs to design an appointment schedule and a rescheduling policy for a single provider and a set of patients, where each patient has a known type and associated probability distributions of service duration and arrival time deviation (the difference between the scheduled and the actual arrival times). The objective is to minimize the expected total cost of patient waiting time, provider idle time, and provider overtime.

In theory, SOASP could be formulated as a *multistage* stochastic mixed-integer program (MSMIP). The first stage of the MSMIP pertains to deciding the initial sequence and schedule (i.e., a sequence of scheduled arrival times) of the patients, which are communicated to the patients before the start of the day. In each stage (the arrival of one of the scheduled patients) after the first stage, and for every branch in the scenario tree, the MSMIP optimizes a policy that reschedules (resequences or declines) the appointments according to the service durations and stochastic arrival times realized up to each new arrival. That is in each stage, a new appointment sequence is determined by solving a MSMIP from "now until the end" conditional on information revealed up to the current stage.

As we discuss in Section 3.1 formulating and solving SOASP in the form of MSMIP is challenging, primarily because of the enormous size of the scenario tree, randomness and ambiguity of the number of the MSMIP stages, and the mixed-integer recourse variables (see Shapiro, Dentcheva, & Ruszczyński, 2009; Birge & Louveaux, 2011 and references therein for a thorough discussion on the difficulty of formulating and solving MSMIP). Even if one could solve this MSMIP, the resulting rescheduling policy may not be implementable in practice. In recognition of these challenges, we first present two approximations based on *two-stage* stochastic mixed-integer models (TSM). The first TSM relaxes the nonanticipativity constraints of MSMIP (i.e., the OPC manager has the perfect information of all uncertain parameters when making *re*scheduling decisions) and so yields a lower bound of the optimal value of MSMIP. We derive a set of valid inequalities to strengthen and improve the solvability of this TSM formulation. The second approximation implements the (feasible) AO policy, enforcing nonanticipativity, and yielding an upper bound of the optimal value of MSMIP.

Second, we propose a Monte Carlo approach to evaluate the relative gap between the MSMIP upper and lower bounds. Finally, in a series of numerical experiments, we show that these two bounds are very close in a wide range of realistic SOASP instances, demonstrating the near-optimality of the AO policy. We also identify parameter settings that result in a large gap in between these two bounds. Accordingly, we propose an alternative policy based on neighbor-swapping (NS). We demonstrate that this alternative policy leads to a much tighter upper bound and significantly shrinks the gap (i.e., improves the solution of such instances). By doing so, we show the art and power of TSM approximation in evaluating the optimality of feasible (and implementable) solutions to challenging real-world MSMIP problems.

To the best of our knowledge, and according to the recent review of outpatient appointment systems by Ahmadi-Javid et al. (2017) and the literature review in Section 2, our paper contributes to the existing literature on single-server stochastic appointment sequencing and scheduling (SASS) with the first stochastic programming approach that considers: (1) random (and heterogeneous) service and arrival times, (2) optimizing both the initial appointment sequencing and scheduling decisions, and (3) the possibility of rescheduling (i.e., resequencing or declining) and investigating the optimality of the AO policy.

The remainder of the paper is structured as follows. In Section 2, we review the relevant literature. In Section 3, we

formally define SOASP as a MSMIP and present the two TSM approximations. In Section 4, we propose a Monte Carlo approach to evaluate the relative gap between the MSMIP upper and lower bounds. In Section 5, we report numerical results on (1) the optimality and suboptimality of the AO policy, (2) sequencing and scheduling patterns with stochastic service and arrival times, and (3) the value of incorporating the stochastic arrival times and service durations in the appointment scheduling problem. Finally, we draw conclusions in Section 6.

2 | LITERATURE REVIEW

Outpatient appointment scheduling problems have been an area of intense investigation in the field of operations research since the seminal work of Welch and Bailey (1952). Comprehensive surveys of the studies since then include Ahmadi-Javid et al. (2017), Cayirli and Veral (2003), and Gupta and Denton (2008). Within this literature, most studies assume patient punctuality and focus on homogeneous patients. In the following, we review the papers that consider patient heterogeneity and stochastic arrival times.

Cayirli et al. (2006, 2008), Klassen and Rohleder (1996), and Rohleder and Klassen (2000) demonstrated how using patient characteristics, either for appointment sequencing or scheduling, can improve the clinic performance and reduce the operational costs. More recent studies (see, eg, Ahmed, Chou, & Shehadeh, 2014; Berg, Denton, Erdogan, Rohleder, & Huschka, 2014; Chen & Robinson, 2014; Salzarulo, Mahar, & Modi, 2016; Shehadeh, Cohn, & Epelman, 2019 and references therein) also demonstrated the relevance of recognizing different patient types in appointment sequencing. Salzarulo et al. (2016), for example, described how data on patient characteristics and past appointments could be used to predict patient service durations and illustrated how to incorporate this information into effective scheduling decisions.

Patient stochastic arrival (often called patient unpunctuality) is a common phenomenon in OPC and introduces complexity to the design and analysis of appointment scheduling. Several studies have shown that both early and late arrivals disrupt clinic service operations (Alexopoulos et al., 2008; Fetter & Thompson, 1966; Glowacka et al., 2017), increase inefficiencies and delays (Deceuninck et al., 2018; Okotie, Patel, & Gonzalez, 2008), decrease the service quality for punctual patients (Glowacka et al., 2017), and increase clinic operational costs (eg, the provider overtime cost) and implicit costs related to the perceived quality of service (eg, decreased patient satisfaction) and hence the reputability of the clinic (Cayirli & Veral, 2003; Deceuninck et al., 2018; Glowacka et al., 2017; Kocas, 2015; Osuna, 1985). Stochastic arrival times can also lead to adverse operational outcomes and challenging queueing issues (Cayirli & Veral, 2003; Deceuninck et al., 2018; Jiang et al., 2019; Klassen & Yoogalingam, 2014). For example, not serving an early arrival (or waiting for a chronically late patient) can result in service delay of the next on-time patient (Deceuninck et al., 2018; Glowacka et al., 2017; Samorani & Ganguly, 2016). Therefore, it is worthwhile to design appointment schedules and resequencing policies that mitigate the impacts of both early and late arrivals.

Most empirical studies on stochastic arrival times suggest that patients tend to arrive early instead of being late (Brahimi & Worthington, 1991; Cox, Birchall, & Wong, 1985; Fetter & Thompson, 1966; Klassen & Rohleder, 1996; Klassen & Yoogalingam, 2014; Williams et al., 2014). Several studies used the normal, uniform, and exponential (for late patients) distributions to model stochastic arrival deviation (see Alexopoulos et al., 2008; Cheong, Bitmead, & Fontanesi, 2013; Cox et al., 1985; Tai & Williams, 2012; White & Pike, 1964 and references therein for a thorough discussion on the distribution of stochastic arrivals). It is worth noting that all of these studies observe that the stochastic arrival time deviation is independent of appointment time. In addition, Williams et al. (2014) and Gneezy, Meier, and Rey-Biel (2011) studied the impacts of strategies that intend to motivate patient punctuality and discussed when and why some strategies might fail. Interventions to reduce tardiness, for example, might potentially lead to increasing patient earliness. Therefore, as pointed out by Deceuninck et al. (2018), it is better to incorporate the uncertainty of patient arrival time in the schedule optimization process and develop policies to resequence appointments instead of trying to change the patients' arrival behavior.

Most of the existing literature on stochastic appointment scheduling focuses on the uncertainty pertaining to service durations and/or patient no-shows. Within the few papers accounting for stochastic arrival times, there are three primary approaches. The first approach employs heuristics (often in conjunction with simulation approaches) to develop and adjust the scheduling pattern (see, eg, Cayirli & Yang, 2014; Cayirli et al., 2006; Cox et al., 1985; Fetter & Thompson, 1966; Glowacka et al., 2017; Klassen & Yoogalingam, 2014; White & Pike, 1964). Although these pioneering studies provide easy-to-implement scheduling rules, they lack performance guarantee and often ignore the possibility of rescheduling (resequencing and declining). In this paper, we incorporate rescheduling into our SOASP model and provide computable performance guarantees.

The second approach employs predetermined resequencing and rescheduling policies. For example, Deceuninck et al. (2018) adopted the AO policy with 15- and 30-minute grace periods (i.e., the provider will serve patients arriving within the grace period in the scheduled order, and decline or deem as a no-show any late patient coming after the expiration of his/her grace period past the scheduled time). In addition, Deceuninck et al. (2018) proposed a local search algorithm based on the Lindley recursion (Lindley, 1952) to find an appointment schedule based on a sequence found by a heuristic. Differently, Zhu, Chen, Leung, and Liu (2018) assumed

that the provider would serve a waiting patient regardless of the actual appointment time, and when multiple patients are waiting, the provider will serve them by the order of their appointment times. Samorani and Ganguly (2016) studied the problem of whether an available provider should see an early patient right away (preempt) or wait for the next scheduled patient. Based on simulation results, Samorani and Ganguly (2016) reported the conditions under which the AO policy outperforms the always-preempt policy (eg, high-service-level clinics with low variability in patient lateness, long service durations, and when patients tend to arrive early rather than late).

First-come-first-serve is another policy which designates that the provider serves the patients according to their actual arriving order, regardless of their initial scheduled order of arrival. This policy often raises the concern that it (partially) conflicts with the goal of appointment scheduling and can encourage the patients to arrive early and beat the appointment system (Cayirli & Veral, 2003; Deceuninck et al., 2018; Jiang et al., 2019). In this paper, we adopt the AO policy to obtain an upper bound on the optimal value of MSMIP. However, we do not assume homogeneous patients as in Samorani and Ganguly (2016) or a fixed appointment sequence as in Deceuninck et al. (2018). Furthermore, we rigorously demonstrate the near-optimality of this policy in a wide range of parameter configurations and propose an alternative and near-optimal policy where this policy is suboptimal.

The third approach employs stochastic mixed-integer programming (SMIP) but, to the best of our knowledge, only involves very few papers (Ahmadi-Javid et al., 2017). Jiang et al. (2019) proposed a stochastic linear program that considers stochastic arrival and adopts the AO policy. Assuming patient homogeneity, they demonstrated how the optimal appointment intervals change under different stochastic arrivals scenarios. Our paper aims to contribute to this line of research. In contrast to Jiang et al. (2019), our SMIP approach considers (1) patient heterogeneity, (2) optimizing both the initial appointment sequencing and scheduling decisions, and (3) the possibility of rescheduling (i.e., resequencing or declining).

3 | FORMULATIONS OF SOASP

In this section, we formally define SOASP as a MSMIP and present two approximations based on TSMs. The first approximation relaxes the nonanticipativity constraints of SOASP and yields a lower bound on the optimal value of SOASP. The second approximation applies a (feasible) rescheduling policy and yields an upper bound.

3.1 | Formal statement of the problem

We consider the problem of sequencing a set of patients (appointments) for a single provider and determining the associated scheduled appointment time for each appointment and an adaptive *rescheduling* (i.e., resequencing or declining) policy. Each patient has a known type and a random service duration that follows a known probability distribution associated with the patient type. The actual arrival time of each patient is random relative to their scheduled arrival time, and the arrival time deviation follows a known probability distribution that is independent of the scheduled arrival time.

The provider's response to a patient's arrival depends on three periods (parameters): a *grace period* past the scheduled time, a *rescheduling period* past the grace period, and a *decline period* past the rescheduling period (denoted by intervals G, R, and DE respectively in Figure 1). Patients arriving early (i.e., within interval E) or within the grace period (i.e., within interval G) are granted to be served either in the initial scheduled order or maybe resequenced to an earlier (and not to a later) position than their initial scheduled position. Additionally, if a patient arrives within interval E or G, then any patients initially scheduled after him/her cannot be resequenced to an appointment before him/her. Patients arriving within the rescheduling period (i.e., within interval R) are subject to either resequencing to a later position than the initial scheduled position (i.e., the provider will try to fit him/her within the remaining sequence of appointments if possible) or declining (rescheduled to another day). Patients arriving within the decline period (i.e., within interval DE) are automatically declined.

The performance metric of scheduling is the weighted sum of three components: (i) *patient waiting time* (the time from the later of the patient *scheduled* arrival time and *actual* arrive time to the actual start time of his/her appointment), (ii) *provider idle time* (the time from the end of one appointment to the start of the next), and (iii) *provider overtime* (time worked beyond the scheduled working hours). Note that we do not incorporate the part of the waiting time due to the early arrival of a patient. That is, if a patient arrives early (i.e., within interval E), his/her waiting time is measured from

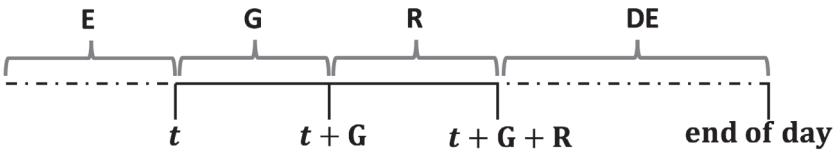


FIGURE 1 An illustration of the provider's response to patient's arrival. If a patient with scheduled arrival time t arrives any time before $t + G$, then s/he may be resequenced to an earlier position than their initial scheduled position and will not be resequenced to a later position. If s/he arrives in between $t + G$ and $t + G + R$, then s/he is subject to resequencing to a later position than the initial scheduled position or declining. If s/he arrives after $t + G + R$, then s/he will be declined. Notation: t is scheduled arrival time, E is early arrival period, G is grace period, R is rescheduling period, and DE is decline period

his/her scheduled time. This is consistent with the prior literature (see, eg, Gupta & Wang, 2012; Cayirli & Veral, 2003). Although declining a very late patient may be mathematically optimal in terms of collectively minimizing metrics (i)-(iii), it may impact patients' health and lead to patient dissatisfaction. While it is interesting to incorporate patient health outcomes and satisfaction, it is out of the scope of this paper, and we leave it as future work.

Table 1 summarizes the provider's rescheduling decisions and the corresponding waiting time calculations. We make the following assumptions:

A1. The set of patients (i.e., the total number of appointments and their types) are exogenously determined (a standard assumption in the offline appointment scheduling literature that mimics the OPC practice; see, eg, Ahmadi-Javid et al., 2017; Berg et al., 2014; Shehadeh, Cohn, & Jiang, 2020; Shehadeh, 2019; Zhu et al., 2018 and the references therein).

A2. The provider is always available at the start of the day, and immediately after the completion of each appointment (a common assumption in the appointment scheduling literature with random arrivals; see, eg, Berg et al., 2014; Cayirli et al., 2008; Deceuninck et al., 2018; Jiang et al., 2019; Shehadeh et al., 2019, 2020).

A3. The stochastic service duration and arrival time deviation (the difference between the scheduled and the actual arrival times) are independent of the scheduled time (a standard assumption in the appointment scheduling literature with random service and arrival times; see Cayirli et al., 2008; Cox et al., 1985; Berg et al., 2014; Deceuninck et al., 2018; Jiang et al., 2019; Klassen & Yoogalingam, 2014; Mancilla & Storer, 2012; Shehadeh et al., 2019). For the stochastic arrival time deviation, if a patient initially scheduled at 8 AM arrives at 8:20 AM (respectively, 7:40 AM) then the stochastic arrival time deviation equals 20 minutes (respectively, −20 minutes).

A4. The degree of the stochastic arrival time deviation of patient p, u_p, has a bounded support $[\underline{u}_p, \overline{u}_p]$, where $\underline{u}_p$ and $\overline{u}_p$ respectively are the lower and upper bounds on the stochastic arrival time deviation of patient p.

A5. The grace period and the rescheduling period are nonnegative, that is, $G \geq 0$ and $R \geq 0$. Note that the grace period represents the provider's tolerance for patient lateness (Deceuninck et al., 2018; Jiang et al., 2019), and so provider's response to those arriving within interval G (see Table 1) must be incorporated in any rescheduling policy.

In theory, SOASP could be formulated as a MSMIP. The first stage of the MSMIP pertains to deciding the initial sequence and schedule (i.e., a sequence of scheduled arrival times) of the patients, which are communicated to the patients before the start of the day. In each stage after the first stage, and for every branch in the scenario tree, the MSMIP optimizes a policy that reschedules (resequences or declines) the appointments according to the service durations and stochastic arrival times realized up to each new arrival. That is in each stage, a new appointment sequence is determined by solving a MSMIP for "from now until the end."

Unfortunately, formulating and solving this MSMIP is challenging for two reasons. First, even in a simplified setting in which all random variables are discretized, the scenario tree of the MSMIP can be enormous and grow exponentially in size with the number of patients. For example, suppose that each stage of the MSMIP represents a new patient arrival, then we need multiple random variables to describe the system status at this stage, including which patient the new arrival pertains to, whether this patient arrives in intervals E/G/R/DE, how many existing patients have entered/finished the service, if the provider is currently busy/idle, and so on. Second, each stage after the first stage involves a set of binary and continuous decision variables pertaining to the rescheduling (resequencing or declining) of appointments (see Birge & Louveaux, 2011; Shapiro et al., 2009 and references therein for a thorough discussion on the difficulty of formulating and solving MSMIP).

In recognition of the challenges of formulating and solving the MSMIP, we propose two approximations based on TSMs. Given the grace period, the first TSM in Section 3.2 relaxes the nonanticipativity constraints of MSMIP (i.e., determine the rescheduling decisions under perfect information of all uncertain parameters) and so yields a lower bound of the optimal value of MSMIP. Given the grace period, the second TSM in Section 3.3 implements the AO policy, enforcing nonanticipativity and yielding an upper bound of the optimal value of MSMIP. In Section 5, we show that the MSMIP bounds are

TABLE 1 The provider's response to patient's arrival within the intervals defined in Figure 1 and waiting time calculations

Arrival	Interval	Arrival time deviation	Provider's response	Waiting time
Early or punctual	E	$u \leq 0$	KEEP or RES-E	$s - t$
Late but within the grace period	G	$0 < u \leq G$	KEEP or RES-E	$s - a$
Late but within the rescheduling period	R	$G < u \leq G + R$	RES or DECL	$s - a$ (RES) or 0 (DECL)
Late but within the decline period	DE	$u > G + R$	DECL	0

Note: KEEP is keeping the initial scheduled position, RES is resequencing, RES-E is resequencing to an earlier position, DECL is declining, t is scheduled arrival time, a is actual arrival time, u is arrival time deviation, that is, $u = a - t$, and s is actual start time.

very close in many SOASP parameter settings, demonstrating the near-optimality of the AO policy. We also identify parameter settings that result in a large gap. Accordingly, we propose an alternative swap-based policy that improves the solution of such instances.

3.2 | TSM under perfect information

The first TSM relaxes the nonanticipativity constraints of the MSMIP. That is, the TSM assumes that we possess perfect information (i.e., the realizations of all service durations and stochastic arrival times) after deciding the initial sequence and schedule in the first stage. In this case, the TSM reschedules (i.e., resequences or declines) all appointments based on the perfect information in the second stage. Throughout this subsection, we use the term "appointment" and "position" interchangeably to refer to a position in the sequence, and for notational convenience, we suppress the scenario index ξ from the scenario-dependent variables and parameters.

Let binary decision variable $x_{p,i}$ represent the initial assignment of patient p to appointment i (equivalently, position i in the sequence), for all $i, p = 1, \ldots, P$, where P is the number of patients (and appointments). Let t_i represent the scheduled arrival time of appointment i for all $i = 1, \ldots, P$. For all $p = 1, \ldots, P$, let continuous parameters d_p and u_p respectively represent the realized service duration and arrival time deviation of patient p. For all $p = 1, \ldots, P$, let binary parameters e_p, τ_p, δ_p, and π_p respectively indicate the arrival of patient p before or at the scheduled time (i.e., $u_p < 0$ or $u_p = 0$), within interval E or G (i.e., $u_p < \mathrm{G}$), within the rescheduling period (i.e., $\mathrm{G} < u_p \le \mathrm{G} + \mathrm{R}$), and within the declined period (i.e., $u_p > \mathrm{R} + \mathrm{G}$).

Let binary decision variable $y_{i,j}$ represent the resequencing of appointment i to appointment j, for all $i, j = 1, \ldots, P$. Let binary decision variable z_j represent the decline of appointment j, for all $j = 1, \ldots, P$. For all $j = 1, \ldots, P$, let continuous variables a_j, $\widetilde{d}_j$, s_j, and w_j respectively represent the actual arrival time, actual service duration, actual start time, and waiting time of appointment j. Let continuous variables g_j represent the provider idle time before the start of appointment j, for all $j = 1, \ldots, P$. Let continuous variable w_{P+1} represent provider overtime. Let nonnegative parameters c_j^{w}, c_j^{g}, and c^{o} respectively represent the unit penalty costs of waiting, idling, and overtime, for all $j = 1, \ldots, P$. Table 2 summarizes the notation that we use in the following TSM under perfect information (TSM-PI).

$$v^{\mathrm{PI}} = \underset{x,t}{\text{minimize}} \quad \mathbb{E}_{\xi}[Q^{\mathrm{PI}}(x, t, \xi)] \tag{1a}$$

$$\text{Subject to}: \sum_{i=1}^{P} x_{p,i} = 1, \quad \text{for } p = 1, \ldots, P, \tag{1b}$$

$$\sum_{p=1}^{P} x_{p,i} = 1, \quad \text{for } i = 1, \ldots, P, \tag{1c}$$

$$0 \le t_i \le L, \quad \text{for } i = 1, \ldots, P, \tag{1d}$$

$$x_{p,i} \in \{0, 1\}, \quad \text{for } p, i = 1, \ldots, P, \tag{1e}$$

where for each ξ, $Q^{\mathrm{PI}}(x, t, \xi)$ is defined as

$$Q^{\mathrm{PI}}(x, t, \xi) = \underset{s,w,g,y,z}{\text{minimize}} \sum_{j=1}^{P} (c_j^{\mathrm{w}} w_j + c_j^{\mathrm{g}} g_j) + c^{\mathrm{o}} w_{P+1}, \tag{2a}$$

$$\text{Subject to}: z_j \ge \sum_{p=1}^{P} \pi_p x_{p,j}, \quad \text{for } j = 1, \ldots, P, \tag{2b}$$

$$z_j \le \sum_{p=1}^{P} (1 - \tau_p) x_{p,j}, \quad \text{for } j = 1, \ldots, P, \tag{2c}$$

$$\sum_{i=1}^{P} y_{i,j} = 1 - z_j, \quad \text{for } j = 1, \ldots, P, \tag{2d}$$

$$\sum_{j=1}^{P} y_{i,j} = 1 - z_i, \quad \text{for } i = 1, \ldots, P, \tag{2e}$$

$$\sum_{j=i+1}^{P} y_{i,j} \le \sum_{p=1}^{P} (1 - \tau_p) x_{p,i}, \quad \text{for } i = 1, \ldots, (P-1), \tag{2f}$$

$$\sum_{\ell=1}^{j} y_{k\ell} \le \sum_{p=1}^{P} (1 - \tau_p) x_{pi} - y_{ij} + 1, \quad \text{for } 1 \le j \le i \le P, \; k = i+1, \ldots, P, \tag{2g}$$

$$\widetilde{d}_j = \sum_{p=1}^{P} \sum_{i=1}^{P} d_p x_{p,i} y_{i,j}, \quad \text{for } j = 1, \ldots, P, \tag{2h}$$

$$a_j = \sum_{i=1}^{P} \left(t_i + \sum_{p=1}^{P} u_p x_{p,i} \right) y_{i,j}, \quad \text{for } j = 1, \ldots, P, \tag{2i}$$

$$s_j \ge a_j, \quad \text{for } j = 1, \ldots, P, \tag{2j}$$

$$s_j \ge s_{j-1} + \widetilde{d}_{j-1}, \quad \text{for } j = 2, \ldots, P, \tag{2k}$$

$$w_j \ge s_j - a_j - M_j \left(z_j + \sum_{p=1}^{P} \sum_{i=1}^{P} e_p x_{p,i} y_{i,j} \right), \quad \text{for } j = 1, \ldots, P, \tag{2l}$$

$$w_j \ge s_j - \sum_{i=1}^{P} t_i y_{i,j} - M_j \left(2 - (1 - z_j) - \sum_{p=1}^{P} \sum_{i=1}^{P} e_p x_{p,i} y_{i,j} \right), \quad \text{for } j = 1, \ldots, P, \tag{2m}$$

TABLE 2 Notation of the TSM formulations

Indices	
p	Patient index
i, j	Appointment indices
Parameters	
P	Number of patients
c_j^{w}	Unit waiting time cost of appointment j
c_j^{g}	Unit provider idle time cost after appointment j
c^{o}	Unit provider overtime cost
G	Grace period
R	Rescheduling period
L	Planned length of provider working hours
Scenario-dependent parameters	
d_p	Realized service duration of patient p
u_p	Realized arrival time deviation of patient p
e_p	$\begin{cases} 1 & \text{if patient } p \text{ arrives early (ie, within interval E) or at the scheduled time,} \\ 0, & \text{otherwise.} \end{cases}$
τ_p	$\begin{cases} 1 & \text{if patient } p \text{ arrives within interval E or G,} \\ 0, & \text{otherwise.} \end{cases}$
δ_p	$\begin{cases} 1 & \text{if patient } p \text{ arrives within the rescheduling period (ie, interval R),} \\ 0, & \text{otherwise.} \end{cases}$
π_p	$\begin{cases} 1 & \text{if patient } p \text{ arrives within the decline period (ie, interval DE),} \\ 0, & \text{otherwise.} \end{cases}$
K	Number of all patients who can be resequenced or declined, $K \equiv \sum_{p=1}^{P}(\delta_p + \pi_p)$
ξ	Random vector containing scenario-dependent parameters, $\xi = (d_1, \ldots, d_P, u_1, \ldots, u_P, e_1, \ldots, e_P, \tau_1, \ldots, \tau_P, \delta_1, \ldots, \delta_P, \pi, \ldots, \pi_P, K)$
Scenario-independent (first stage) variables	
$x_{p,i}$	$\begin{cases} 1 & \text{if patient } p \text{ is initially assigned to appointment } i, \\ 0, & \text{otherwise.} \end{cases}$
t_i	Scheduled arrival time of appointment i
Scenario-dependent (second-stage) variables	
$y_{i,j}$	$\begin{cases} 1 & \text{if appointment } i \text{ is resequenced to appointment } j, \\ 0, & \text{otherwise.} \end{cases}$
z_j	$\begin{cases} 1 & \text{if appointment } j \text{ is declined,} \\ 0, & \text{otherwise.} \end{cases}$
a_j	Actual arrival time of appointment j
s_j	Actual start time of appointment j
$\widetilde{d}_j$	Actual service duration of appointment j
w_j	Waiting time of appointment j
g_j	Provider idle time before the start of appointment j
w_{P+1}	Provider overtime

$$g_1 = s_1, \tag{2n}$$

$$g_j = s_j - (s_{j-1} + \widetilde{d}_{j-1}), \quad \text{for } j = 2, \ldots, P, \tag{2o}$$

$$w_{P+1} \geq s_P + \widetilde{d}_P - L, \tag{2p}$$

$$y_{i,j} \in \{0, 1\}, \quad \text{for } i, j = 1, \ldots, P, \tag{2q}$$

$$(g_j, s_j, w_j, w_{P+1}) \geq 0, \quad \text{for } j = 1, \ldots, P. \tag{2r}$$

Objective (1a) minimizes the expected total costs of patient waiting, provider idle time, and provider overtime. Constraints (1b) and (1c) ensure that initially each patient is assigned to one appointment and each appointment is assigned one patient, respectively.

Recall that in TSM-PI, we assume that we possess perfect information about the stochastic arrival time deviations and service durations. For modeling convenience, we choose to treat any patient with appointment declined (whether he/she arrives in interval DE, and thus is automatically declined, or in interval R and the model declines him/her) as a "ghost" patient arriving at time 0 (this ensures that the provider does

not wait for his/her grace period to expire before moving on to the next patient), and with zero service duration. We also force the waiting time of the "ghost" patient to be zero, as explained below.

Constraints (2b)–(2g) determine the new sequence. Specifically, constraint (2b) specifies that if a patient p initially assigned to position j (i.e., $x_{p,j} = 1$) arrives within the decline period (within interval DE and so $\pi_p = 1$) then this patient (and the corresponding appointment j) must be declined. Constraint (2c) specifies that if a patient p initially assigned to position j arrives any time before the end of the grace period (i.e., within interval E or G and so $\tau_p = 1$), then this patient (and the corresponding appointment) must not be declined. Constraint (2d) specifies that if a patient initially assigned to appointment j is declined, then no appointment can be resequenced to appointment j, that is, $y_{i,j} = 0 \ \forall i$ (recall that we still "treat" patient j, as a "ghost" with zero service duration). Constraint (2e) specifies that if the initial appointment i is declined, then this appointment cannot be resequenced to any other appointment.

Constraint (2f) specifies that if a patient p initially assigned to appointment i arrives early or within the grace period (i.e., within interval E or G and so $\tau_p = 1$ by definition), then he/she cannot be resequenced to an appointment later than i. Constraint (2g) specifies that if a patient initially assigned to appointment i arrives early or within the grace period, then any patients initially scheduled after him/her cannot be resequenced to an appointment before him/her. Note that, under this global resequencing rule, if the patient initially assigned to appointment i arrives within rescheduling period (i.e., interval R) and does not get declined, then he/she can still be resequenced to a position before i. For example, suppose that patients 1 and 2 arrive within interval E or G and patients 3 and 4 arrive within interval R. We are allowed to resequence them as 1-2-4-3 in this case. However, neither 3 nor 4 can be resequenced before 1 or 2. This is reflected in constraint (2g). Note that constraints (2d)–(2g) allow for the initial and the new positions to be the same. In this case, $j = i$ and so $y_{i,i} = 1$.

Constraint (2h) determines the actual service duration of appointment j based on the new sequence: if appointment j is declined (i.e., if $z_j = 1$ and so $y_{i,j} = 0$ for all i by constraint (2d)) then $\widetilde{d}_j = 0$; and if appointment j is not declined (i.e., if $z_j = 0$ and so $y_{i,j} = 1$ for some i) then $\widetilde{d}_j = d_p$ if patient p is initially assigned to appointment i and then resequenced to appointment j (i.e., $x_{p,i} = 1$ and $y_{i,j} = 1$). Note that constraint (2h) and some of the remaining constraints of this TSM are nonlinear because of the interaction terms $x_{p,i}y^n_{i,j}$ (see constraints (2h)–(2i) and (2l)–(2m)) and bilinear terms $t_i y^n_{i,j}$ (see constraint (2i)). In Section 4.2, we propose strategies to linearize these constraints.

Constraint (2i) determines the actual arrival time. Specifically, if patient p is initially assigned to position i (i.e., $x_{p,i} = 1$) and resequenced to position j (i.e., $y_{i,j} = 1$ and $z_j = 0$) then the actual arrival time of appointment j is $a_j = t_i + u_p$. In contrast, if appointment j is declined (i.e., $z_j = 1$ and so $y_{i,j} = 0$ for all $i = 1, \ldots, P$ by (2d)) then $a_j = 0$ (recall that this is a "ghost" patient arriving at time zero by design). Constraints (2j)–(2k) specify that the actual start time of appointment j, s_j, is no smaller than the actual arrival time, a_j, and no smaller than the completion time of the preceding appointment, that is, $s_{j-1} + \widetilde{d}_{j-1}$. Note that if appointment j is declined then the corresponding "ghost" patient will start and complete at the completion time of the preceding appointment $s_{j-1} + \widetilde{d}_{j-1}$ because $a_j = 0$ and $\widetilde{d}_j = 0$ for all declined appointments by (2i) and (2h).

Constraints (2l)–(2m) compute the waiting time based on the following four arrival time scenarios (see the first and last columns of Table 1) with M_j representing a sufficiently large constant. First, if patient p is initially assigned to appointment i, arrives early or at the scheduled time (i.e., within interval E), and is resequenced to appointment j (i.e., $x_{p,i} = y_{i,j} = 1$), then $\tau_p = 1$ by definition and so $z_j = 0$ by (2c), $e_p = 1$ by definition, and $\sum_{p=1}^{P} \sum_{i=1}^{P} e_p x_{p,i} y_{i,j} = 1$. It follows that constraint (2l) is relaxed and $w_j = \max\{s_j - t_i, 0\}$ by (2m) and (2r), that is, the waiting time of this patient is the difference between his/her actual start time and the scheduled arrival time. Recall from Section 3.1 that, consistent with the prior literature, we do not incorporate the part of the waiting time due to the early arrival of a patient (see, eg, Gupta & Wang, 2012; Cayirli & Veral, 2003).

Second, if patient p is initially assigned to appointment i arrives late but within the grace period (i.e., interval G), and is resequenced to appointment j, then $\tau_p = 1$ by definition and so $z_j = 0$ by (2c), $e_p = 0$ by definition, $\sum_{p=1}^{P} \sum_{i=1}^{P} e_p x_{p,i} y_{i,j} = 0$, and $a_j = t_i + u_p$ by (2i). It follows that constraint (2 m) is relaxed and $w_j = s_j - a_j$ by (2l), that is, the waiting time of this late patient is computed from his/her actual arrival time. Third, if patient p is initially assigned to appointment i arrives late but within the rescheduling period (i.e., interval R) but does not get declined, and resequenced to appointment j, then $z_j = 0$, $a_j = t_i + u_p$ by (2i), $e_p = 0$ by definition, constraint (2m) is relaxed, and $w_j = s_j - a_j$ by (2l). Fourth, if patient p initially assigned to appointment j arrives within the decline period (i.e., interval DE), then $\pi_p = 1$ by definition and so $z_j = 1$ by (2b), $y_{i,j} = 0 \ \forall i$ by (2d), and $\sum_{p=1}^{P} \sum_{i=1}^{P} e_p x_{p,i} y_{i,j} = 0$. It follows that both (2l)–(2m) are relaxed and $w_j = 0$ by (2r), that is, declined appointments yield zero waiting time.

Constraint (2n) computes the idle time before the first patient. Constraint (2o) compute the idle time between two consecutive appointments. Constraints (2p) and (2r) compute the provider overtime (if any) beyond the planned length of working hours. Finally, constraints (2q)–(2r) specify binary and nonnegative restrictions on the decision variables.

Proposition 1 *For a fixed grace period G and cost vectors c^w, c^g, and c^o, we have $v^{PI} \le v$, where v is the optimal value of the MSMIP.*

Proof Formulation (2) relaxes the nonanticaptivity constraints in the MSMIP. It follows that $v \geq v^{\mathrm{PI}}$. ∎

3.3 | TSM under the appointment order policy

The second TSM implements a fixed (feasible) rescheduling policy termed *AO policy*, which designates that (i) any patient arriving after his/her grace period (i.e., within interval R or DE) gets declined and (ii) the remaining patients are served in the order of their initial sequence regardless of their actual arrival times. Using the same notation in Table 2, we present the TSM under the appointment order policy (TSM-AO) as follows.

$$v^{\mathrm{AO}} = \underset{x,t}{\text{minimize}} \ \mathbb{E}_{\xi}[Q^{\mathrm{AO}}(x,t,\xi)] \tag{3a}$$

$$\text{Subject to:} \quad \text{(1b)–(1e)} \tag{3b}$$

where for each ξ, $Q^{\mathrm{AO}}(x, t, \xi)$ is defined as

$$Q^{\mathrm{AO}}(x,t,\xi) = \underset{s,w,g}{\text{minimize}} \sum_{j=1}^{P}(c_j^{\mathrm{w}} w_j + c_j^{\mathrm{g}} g_j) + c^{\mathrm{o}} w_{P+1}, \tag{3c}$$

$$\text{Subject to:} \ a_j = t_j + \sum_{p=1}^{P}(u_p\tau_p + G(1-\tau_p))x_{p,j}, \quad \text{for} j = 1, \ldots, P, \tag{3d}$$

$$s_j \geq a_j, \ \text{ for } \ j = 1, \ldots, P, \tag{3e}$$

$$s_j \geq s_{j-1} + \sum_{p=1}^{P} d_p\tau_p x_{p,j-1}, \ \text{ for } \ j = 2, \ldots, P, \tag{3f}$$

$$w_j \geq s_j - a_j - M_j\left(1 + \sum_{p=1}^{P}(e_p - \tau_p)x_{p,j}\right), \quad \text{for } \ j = 1, \ldots, P, \tag{3g}$$

$$w_j \geq s_j - t_j - M_j\left(2 - \sum_{p=1}^{P}(e_p + \tau_p)x_{p,j}\right), \quad \text{for } \ j = 1, \ldots, P, \tag{3h}$$

$$g_1 = s_1, \tag{3i}$$

$$g_j = s_j - \left(s_{j-1} + \sum_{p=1}^{P} d_p\tau_p x_{p,j-1}\right), \ \text{ for } \ j = 2, \ldots, P, \tag{3j}$$

$$w_{P+1} \geq s_P + \sum_{p=1}^{P} d_p\tau_p x_{p,P} - L, \tag{3k}$$

$$(g_j, s_j, w_j, w_{P+1}) \geq 0, \ \text{ for } \ j = 1, \ldots, P. \tag{3l}$$

Note that under the AO policy, the provider must wait for a patient till the end of his/her grace period before becoming eligible to treat a subsequent patient (Deceuninck et al., 2018). For modeling convenience, we choose to treat any patient arriving after the end of the grace period (i.e., within interval R or DE) as a "ghost" patient arriving at $t + G$ with zero service duration.

Objective (3a) minimizes the expected total costs of patient waiting, provider idle time, and provider overtime. Constraint (3d) determines the actual arrival time. Specifically, if patient p assigned to appointment j (i.e., $x_{p,j} = 1$) arrives early or within the grace period (i.e., within interval E or G) and so $\tau_p = 1$ by definition, then $a_j = t_j + u_p$. In contrast, if this patient arrives after the end of the grace period (i.e., within interval R or DE) and so $\tau_p = 0$ by definition, then this patient is declined and $a_j = t_j + G$, that is, this patients is treated as a "ghost" patient arriving at $t_j + G$ with zero service time, that is, $\sum_{p=1}^{P} d_p\tau_p x_{p,j} = 0$. Constraints (3e)–(3f) ensure that the actual start time is at least the arrival time of the patient and at least the completion time of the previous patient. Note that if appointment j is declined then the corresponding "ghost" patient p will start and complete at $s_j + \sum_{p=1}^{P} d_p\tau_p x_{p,j} = \max\left\{a_j = t_j + G, s_{j-1} + \sum_{p=1}^{P} d_p\tau_p x_{p,j-1}\right\} + 0$ because $\tau_p = 0$ and so $\sum_{p=1}^{P} d_p\tau_p x_{p,j} = 0$. Collectively, constraints (3d)–(3f) enforce the AO policy mathematically. For example, if patient 4 arrives earlier than $t_3 + G$ (i.e., the expiration time of patient 3's grace period), the *provider whenever becomes available will remain idle until one of the following two cases takes place*: (1) patient 3 arrives before $t_3 + G$ and the provider starts serving patient 3, or (2) $t_3 + G$ expires, and the provider declines patient 3 and starts serving patient 4.

Constraints (3g)–(3h) compute the waiting time based on the following three arrival time scenarios with M_j representing a sufficiently large constant. First, if patient p is assigned to appointment j (i.e., $x_{p,j} = 1$) and arrives early or at the scheduled time (i.e., within interval E), then $e_p = 1$ and $\tau_p = 1$ by definition, and thus $\sum_{p=1}^{P}(e_p - \tau_p)x_{p,j} = 0$, and $\sum_{p=1}^{P}(e_p + \tau_p)x_{p,j} = 2$. It follows that constraint (3g) is relaxed and $w_j = \max\{s_j - t_j, 0\}$ by constraints (3h) and (3l). Second, if patient p is assigned to appointment j (i.e., $x_{p,j} = 1$) and arrives within the grace period (i.e., within interval G), then $e_p = 0$ and $\tau_p = 1$ by definition, and so $\sum_{p=1}^{P}(e_p - \tau_p)x_{p,j} = -1$ and $\sum_{p=1}^{P}(e_p + \tau_p)x_{p,j} = 1$. It follows that constraint (3h) is relaxed and $w_j = s_j - a_j$ by constraint (3g). Third, if patient p assigned to appointment j arrives after the end of the grace period (i.e., within interval R or DE), then $e_p = \tau_p = 0$ by definition and so $\sum_{p=1}^{P}(e_p - \tau_p) = \sum_{p=1}^{P}(e_p + \tau_p)x_{p,j} = 0$. It follows that both (3g) and (3h) are relaxed and $w_j = 0$ by (3l), that is, declined appointments yield zero waiting time.

Constraint (3i) computes the idle time before the first patient. Constraint (3j) computes the idle time between two consecutive appointments. Constraints (3k) and (3l) compute the provider overtime beyond the planned length of

working hours. Finally, constraint (3l) specifies nonnegative restrictions on the decision variables.

Proposition 2 *For a fixed grace period G and cost vectors c^w, c^g, and c^o, we have $v \le v^{AO}$. Furthermore, $v^{PI} \le v \le v^{AO}$.*

Proof The AO policy is a feasible rescheduling policy to the MSMIP. It follows that $v \le v^{\mathrm{AO}}$, and so $v^{\mathrm{PI}} \le v \le v^{\mathrm{AO}}$ by Proposition 1. ∎

4 | SOLUTION APPROACH

There are two well-known difficulties in obtaining an (exact) optimal solution to the TSM-PI in (1) and TSM-AO in (3). First, evaluating the values of $\mathbb{E}_\xi[Q^{\mathrm{PI}}(x,t,\xi)]$ and $\mathbb{E}_\xi[Q^{\mathrm{AO}}(x,t,\xi)]$ involves taking multidimensional integrals. Second, formulation (2) of $Q^{\mathrm{PI}}(x, t, \xi)$ involves mixed-integer recourse variables, which makes $Q^{\mathrm{PI}}(x, t, \xi)$ non-convex and even discontinuous. In view of these two difficulties, we resort to approximation solution approaches.

In Section 4.1, we present a Monte Carlo approach to obtain near-optimal solutions to (1) and (3). In Section 4.2, we propose several strategies to linearize and strengthen the formulation (1), which helps improve the solution efficacy of the TSM-PI.

4.1 | Monte Carlo optimization

In the Monte Carlo approach, we replace the distribution of ξ with a (discrete) empirical distribution based on N independent and identically distributed (i.i.d.) samples of the service durations and arrival time deviations, and then we solve the sample average approximation (SAA) formulations of (1) and (3), which are denoted formulations (4) and (5), respectively, and presented below. Note that, in the SAA formulations (4) and (5), we associate all scenario-dependent parameters, variables, and constraints with a scenario index n for all $n = 1, \ldots, N$. For example, parameters d_p are replaced by d_p^n to represent the service durations realized in scenario n, and variables $y_{i,j}$ are replaced by $y_{i,j}^n$ to represent the resequencing decisions in scenario n. In addition, constraints (2b)–(2r) and (3d)–(3l) are incorporated in each scenario.

$$\text{(SAA-PI)} \quad v_N^{\mathrm{PI}} = \underset{x,t,s,w,g,y,z}{\text{minimize}} \ \widehat{f}_N^{\mathrm{PI}}$$

$$:= \frac{1}{N} \sum_{n=1}^{N} \left[\sum_{j=1}^{P} (c_j^{\mathrm{w}} w_j^n + c_j^{\mathrm{g}} g_j^n) + c^{\mathrm{o}} w_{P+1}^n \right] \tag{4a}$$

$$\text{Subject to}: \text{(1b)–(1e)} \tag{4b}$$

$$\text{(2b)–(2r)}, \ \text{for } n = 1, \ldots, N, \tag{4c}$$

$$\text{(SAA-AO)} \quad v_N^{\mathrm{AO}} = \underset{x,t,s,w,g}{\text{minimize}} \ \widehat{f}_N^{\mathrm{AO}}$$

$$:= \frac{1}{N} \sum_{n=1}^{N} \left[\sum_{j=1}^{P} (c_j^{\mathrm{w}} w_j^n + c_j^{\mathrm{g}} g_j^n) + c^{\mathrm{o}} w_{P+1}^n \right] \tag{5a}$$

$$\text{Subject to}: \text{(1b)–(1e)} \tag{5b}$$

$$\text{(3d)–(3l)}, \ \text{for } n = 1, \ldots, N. \tag{5c}$$

Note that the sample averages $\widehat{f}_N^{\mathrm{PI}}$ and $\widehat{f}_N^{\mathrm{AO}}$ are unbiased estimators of the expected values $f^{\mathrm{PI}} := \mathbb{E}_\xi[Q^{\mathrm{PI}}(x,t,\xi)]$ and $f^{\mathrm{AO}} := \mathbb{E}_\xi[Q^{\mathrm{AO}}(x,t,\xi)]$ in (1) and (3), respectively (Mak, Morton, & Wood, 1999; Shapiro, 2003). By the Law of Large Numbers and Shapiro (2003), we have $\widehat{f}_N^{\mathrm{PI}} \to f^{\mathrm{PI}}$ and $\widehat{f}_N^{\mathrm{AO}} \to f^{\mathrm{AO}}$ with probability one (w.p. 1) as $N \to \infty$ (Homem-de Mello & Bayraksan, 2014; Kleywegt, Shapiro, & Homem-de Mello, 2002; Linderoth, Shapiro, & Wright, 2006). It follows that $v_N^{\mathrm{PI}} \to v^{\mathrm{PI}}$ and $v_N^{\mathrm{AO}} \to v^{\mathrm{AO}}$ w.p. 1 as $N \to \infty$, that is, the optimal values of the SAA formulations (4) and (5) converge to those of TSM-PI and TSM-AO, respectively, as the sample size N grows to infinity. However, for a fixed sample of N scenarios, formulation (4) and (5) reduce to a mixed-integer nonlinear program (MINLP) and a mixed-integer linear program (MILP), respectively. Hence, one would expect the computational effort and solution time of solving the SAA formulations to increase as the sample size increases.

Algorithm 1 summarizes the Monte Carlo optimization (MCO) algorithm that determines an appropriate sample size N and obtain near-optimal solutions to the TSM-PI model based on SAA formulations, and the algorithm for solving the TSM-AO model is similar. This algorithm is based on the vanilla SAA method in Homem-de Mello and Bayraksan (2014), Kleywegt et al. (2002), and Molina-Pariente, Hans, and Framinan (2018) with some adaptations to our TSMs.

Starting with an initial candidate value of N, the MCO algorithm proceeds as follows. First, for $k = 1, \ldots, K$, we repeat the following steps. In step 1.1, we generate a sample of N i.i.d. scenarios of service durations and arrival time deviations. In step 1.2, we solve the SAA formulation of the TSM-PI with the scenarios generated in step 1.1 and record the corresponding optimal objective value v_N^k and optimal schedule $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{t}})_N^k$. In step 1.3, we evaluate the objective function value $v_{N'}^k$ via Monte Carlo simulation of the schedule $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{t}})_N^k$ with a sample of N' i.i.d. scenarios of service durations and arrival time deviations.

In step 2, we compute the average of v_N^k and $v_{N'}^k$ among the K replications as $\bar{v}_N = (1/K)\sum_{k=1}^K v_N^k$ and $\bar{v}_{N'} = (1/K)\sum_{k=1}^K \hat{v}_{N'}^k$, respectively. The statistical results in Mak et al. (1999) and Linderoth et al. (2006) infer that $\bar{v}_N$ and $\bar{v}_{N'}$ are respectively statistical lower and upper bounds of the optimal value of the TSM. In step 3, we compute the approximate

Algorithm 1: Monte Carlo Optimization (MCO) Method

Input: N_o is an initial sample size, K is number of replicates, N' is number of scenarios in the Monte Carlo Simulation step, and ϵ is a termination tolerance.

Output: N is sample size, $\bar{v}_N$ and $\bar{v}_{N'}$ are respectively statistical lower and upper bounds on the optimal value of the TSM, and AOI_N is approximate optimality index.

Initialization: $N := N_o$

Step 1. MCO Procedure

for $k = 1, ..., K$, **do**

Step 1.1 ***Scenario Generation***

- Generate N independent and identical distributed (i.i.d.) scenarios of service durations $\mathbf{d}_p = (d_p^1, ..., d_p^N)^\top$ and arrival time deviations $\mathbf{u}_p = (u_p^1, \ldots, u_p^N)^\top$ for all $p = 1, ..., P$.

Step 1.2 ***Solving the SAA formulation***

- Solve the SAA formulation in (4) with the scenarios generated in step 1.1 and record the corresponding optimal objective value v_N^k and optimal schedule $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{t}})_N^k$.

Step 1.3 ***Cost Evaluation using Monte Carlo Simulation***

- Generate N' i.i.d scenarios of service durations $\mathbf{d}'_p = (d_p^1, ..., d_p^{N'})^\top$ and arrival time deviations $\boldsymbol{u}'_p = (u_p^1, \ldots, u_p^{N'})^\top$ for all $p = 1, ..., P$.
- Use the schedule $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{t}})_N^k$ and parameters $(\mathbf{d}'_p, \boldsymbol{u}'_p)$ to compute $w_j^{n'}$, $g_j^{n'}$, and $w_{P+1}^{n'}$ for all $n' = 1, \ldots, N'$, and evaluate the objective function $\hat{v}_{N'}^k$ as follows:

$$\hat{v}_{N'}^k = \frac{1}{N'} \sum_{n'=1}^{N'} \sum_{j=1}^{P} \left(c_j^{\mathrm{w}} w_j^{n'} + c_j^{\mathrm{g}} g_j^{n'} \right) + c^{\mathrm{o}} w_{P+1}^{n'}$$

end

Step 2. Compute the average of $\hat{v}_N^k$ and $\hat{v}_{N'}^k$ among the K replications

$$\overline{v}_N = \frac{1}{K} \sum_{k=1}^{K} v_N^k \qquad \overline{v}_{N'} = \frac{1}{K} \sum_{k=1}^{K} \hat{v}_{N'}^k$$

Step 3. Compute the Approximate Optimality Index

$$AOI_N = \frac{\overline{v}_{N'} - \overline{v}_N}{\overline{v}_{N'}}$$

Step 4. If AOI_N satisfies a predetermined termination tolerance (i.e., $|AOI_N| < \epsilon$), terminate and output N, $\overline{v}_N$, $\overline{v}_{N'}$, and AOI_N. Otherwise, update $N \leftarrow 2N$, and go to step1.

optimality index $|\ AOI_N = (\bar{v}_{N'} - \bar{v}_N)/\bar{v}_{N'}\ |$ as a point estimate of the relative optimality gap between these two statistical bounds, and so AOI_N serves as an approximate estimate of the optimality gap. Finally, if AOI_N satisfies a predetermined termination tolerance (i.e., $|AOI_N| < \epsilon$), the algorithm terminates and outputs N, $\bar{v}_N$, $\bar{v}_{N'}$, and AOI_N. Otherwise, we increase the sample size (i.e., $N \leftarrow 2N$), and go to step 1. Note that N' can be either fixed to a very large number (eg, 10 000-50 000) or updated dynamically in step 4 such that $N' >> N$.

Finally, we can "warm start" Algorithm 1 from solved instances. That is, we can store the output sample size N in a solved instance (i.e., N leads to a near-optimal solution to the TSM via its SAA formulation) and set the initial sample size $N_o = N$ when solving similar instances. This can significantly speed up Algorithm 1.

We evaluate the gap between the two TSM approximations of SOASP, that is, Gap $:= \frac{v^{\mathrm{AO}} - v^{\mathrm{PI}}}{v^{\mathrm{PI}}}$, where v^{PI} and v^{AO} represent the optimal values of TSM-PI and TSM-AO, respectively. As $\bar{v}_N^{\mathrm{PI}}$ is a statistical lower bound of v^{PI} and $\bar{v}_{N'}^{\mathrm{AO}}$ is a statistical upper bound of v^{AO}, we conservatively approximate Gap by using the approximate (statistical) relative gap ARG $:= \frac{\bar{v}_{N'}^{\mathrm{AO}} - \bar{v}_N^{\mathrm{PI}}}{\bar{v}_N^{\mathrm{PI}}}$.

4.2 | Strengthening the TSM-PI

As compared to the TSM-AO, the SAA formulation (4) of the TSM-PI is significantly more difficult to solve for three main reasons. First, formulation (4) is a MINLP because of the interaction terms $x_{p,i} y_{i,j}^n$ (see constraints (2h)–(2i) and (2l)–(2m)) and bilinear terms $t_i y_{i,j}^n$ (see constraints (2i) and (2m)). Second, as we consider rescheduling (resequencing and declining) in the TSM-PI, formulation (4) involves integer recourse variables $y_{i,j}^n$ and z_j^n. Third, the search space of the rescheduling variables increases with the sample size (i.e., the

number of binary variables $y^n_{i,j}$ and z^n_j grows linearly with the sample size N). In the following subsections, we strengthen the formulation (4) by linearization and valid inequalities.

4.2.1 | SMILP reformulation

Let $\mathcal{N}^{\text{EG}}$ be the set of all scenarios in which all patients arrive early or within the grace period (i.e., within interval E or G and so $\tau^n_p = 1$ and $\delta^n_p = 0$, for all $p = 1, \dots, P$ and $n \in \mathcal{N}^{\text{EG}}$). In this set of scenarios, the provider must follow the initial AO and not resequence any patient. Accordingly, we have $y^n_{j,j} = 1$ and $z^n_j = 0$ for all $j = 1, \dots, P$ and $n \in \mathcal{N}^{\text{EG}}$. Therefore, the nonlinear constraints (2h)–(2p) reduce to linear constraints (3d)–(3k) for all $n \in \mathcal{N}^{\text{EG}}$.

To linearize formulation (4) for all other scenarios $n \notin \mathcal{N}^{\text{EG}}$, we define $\alpha^n_{i,j} = t_i y^n_{i,j}$, $\forall i,j \in [P]$ and $\beta^n_{p,i,j} = x_{p,i} y^n_{i,j}$, $\forall (p, i, j) \in [P]$ (where $[H] \coloneqq \{h \in \mathbb{N} : 1 \le h \le H\}$ for all $H \in \mathbb{N}$). We also incorporate the following McCormick inequalities (6a)–(6b) and (6c)–(6d) for variables $\alpha^n_{i,j}$ and $\beta^n_{p,i,j}$, respectively.

$$\alpha^n_{i,j} - t^{\min}_i y^n_{i,j} \ge 0, \quad \alpha^n_{i,j} - t_i + t^{\max}_i (1 - y^n_{i,j}) \ge 0, \tag{6a}$$

$$\alpha^n_{i,j} - t^{\max}_i y^n_{i,j} \le 0, \quad \alpha^n_{i,j} - t_i + t^{\min}_i (1 - y^n_{i,j}) \le 0, \tag{6b}$$

$$\beta^n_{p,i,j} \ge 0, \quad \beta^n_{p,i,j} - y^n_{i,j} - x_{p,i} + 1 \ge 0, \tag{6c}$$

$$\beta^n_{p,i,j} - x_{p,i} \le 0, \quad \beta^n_{p,i,j} - y^n_{i,j} \le 0, \tag{6d}$$

where $t^{\min}_i = 0$ and $t^{\max}_i = L$ are the minimum and maximum possible values of the scheduled time of appointment i, respectively (see constraint (1d)).

4.2.2 | Valid inequalities

We derive two families of valid inequalities to strengthen formulation (4). We summarize these inequalities in the following proposition.

Proposition 3

$$\sum_{j=\max\{i-K^n,1\}}^{i} y^n_{i,j} \ge \sum_{p=1}^{P} \tau^n_p x_{p,i}, \quad \forall n \in [N], \ \forall i \in [P], \tag{7a}$$

$$y^n_{i,i} \ge 1 - \sum_{k=1}^{i} \sum_{p=1}^{P} (1 - \tau^n_p) x_{pk}, \quad \forall n \in [N], \ \forall i \in [P]. \tag{7b}$$

where $K^n \equiv \sum_{p=1}^{P} (\delta^n_p + \pi^n_p)$.

Proof Inequalities (7a) specify that if a patient p initially assigned to appointment i arrives early or within the grace period in scenario n (i.e., within interval E or G and so $\tau^n_p = 1$ by definition), then he/she can only be resequenced to an appointment j among $\max\{i - K^n, 1\}, \dots, i$. This is because in each scenario n there are at most K^n many appointments before i that can either be declined or be resequenced to be after i. Inequalities (7b) specify that if all patients in the first i appointments arrive early or within the grace period, then appointment i must stay as i. This is because (1) the patient in appointment i cannot swap with any patients before him/her as none of them arrive after the grace period, and (2) he/she cannot swap with any patients after him/her as he/she arrives before the start of the rescheduling period. ∎

4.2.3 | Tight estimation of Big-M coefficients

We derive a tight estimation of the Big-M coefficients involved in constraints (2l)–(2m) and (3g)–(3h) in the following proposition, whose proof is relegated to Appendix A (Supporting information). The tight estimation strengthens the SAA formulations of both TSM-PI and TSM-AO.

Proposition 4 *$M^n_j = L + \max_{p=1,\dots,P}\{u^n_p\} + (j-1)\max_{p=1,\dots,P}\{d^n_p\}$ for $j = 1, \dots, P$ and $n \in [N]$ is a valid value for the Big-M constant in constraints* (2l)–(2m) *and* (3g)–(3h).

4.2.4 | Symmetry breaking inequalities

We apply symmetry breaking inequalities (see Berg et al., 2014; Denton, Miller, Balasubramanian, & Huschka, 2010; Ostrowski, Linderoth, Rossi, & Smriglio, 2011; Shehadeh et al., 2019) to further strengthen the TSM formulations. Specifically, we aggregate the patients into classes, each consisting of patients with a common patient type and a common distribution of arrival time deviations. Suppose that we have Q classes and P_q be the set of patients in class q, $q = 1, \dots, Q$. Without loss of generality, we can assume that patients within each P_q are numbered sequentially. Accordingly, we add the following symmetry breaking constraints to the TSM-PI and TSM-AO (and their SAA formulations):

$$x_{p,i} - \sum_{j=i+1}^{P} x_{p+1,j} \le 0, \quad \forall i = 1, \dots, P, \ \forall p : p, p+1 \in P_q, \ \forall q = 1, \dots, Q, \tag{8}$$

indicating that, if patients p and $p+1$ are of the same class then p is scheduled before $p+1$.

5 | COMPUTATIONAL RESULTS

In this section, we numerically study the statistical relative gap $\text{ARG} \coloneqq \frac{\overline{v}^{\text{AO}}_{N'} - \overline{v}^{\text{PI}}_{N}}{\overline{v}^{\text{PI}}_{N}}$ between SOASP statistical upper $\overline{v}^{\text{AO}}_{N'}$ and lower $\overline{v}^{\text{PI}}_{N}$ bounds and the optimality of the AO policy under

a range of parameter settings. In Section 5.1, we describe the set of SOASP instances that we use in our experiments and discuss other experimental setups. In Section 5.2, we obtain a near-optimal solution to the TSM-PI and TSM-AO via their SAA formulations (specifically, tight statistical lower $\overline{v}_N^{\mathrm{PI}}$ and upper $\overline{v}_{N'}^{\mathrm{AO}}$ bounds on the optimal values of these TSMs). In Section 5.3, we discuss the computation time of TSM-AO and TSM-PI. In Section 5.4, we evaluate ARG between the statistical bounds on the optimal values of these TSMs, demonstrating the near-optimality of the AP policy in a wide range of SOASP parameter settings. We also identify parameter settings that result in a large gap and accordingly propose an alternative rescheduling policy that significantly shrinks the gap in Section 5.5. In Section 5.6, we provide an analysis of the optimal scheduling patterns. Finally, in Section 5.7, we demonstrate the benefits of incorporating uncertainty in outpatient appointment scheduling.

5.1 | Description of experiments

We use assumptions and parameter settings made in prior outpatient appointment scheduling literature to construct 60 SOASP instances characterized by four patient mixes, five distributions of stochastic arrival time deviation, and three choices of the grace period. Each instance consists of 12 patients based on a typical outpatient scheduling problem studied in Deceuninck et al. (2018). Table 3 summarizes the characteristics and parameters of these instances.

Heterogeneous instances (Hetero1 and Hetero2) consist of two types of patients; newly referred ("new") patients and follow-up ("return") patients (a typical patient mix in OPCs; see Deceuninck et al., 2018; Zhu et al., 2018 and references therein). Hetero1 is based on the problem studied in Deceuninck et al. (2018), where 75% of the patients are return, and the remaining 25% are new. Hetero2 represents clinics where 50% of the patients are return, and the remaining 50% are new. Homogeneous instances (Homo1 and Homo2) represent clinics where patients are identical (i.e., of the same type).

The distributions for stochastic service duration of each patient type follow a lognormal distribution, $d(\text{type}) \sim \text{LogN}(\mu_d, \sigma_d^2)$ with mean μ_d and standard deviation σ_d as in Deceuninck et al. (2018). The lognormal distribution is a typical distribution for service duration in the appointment scheduling literature, and several empirical studies show how it accurately describes the shape of service duration distributions in a variety of service systems (see Cayirli et al., 2006; Gul, Denton, Fowler, & Huschka, 2011; Klassen & Rohleder, 1996; Klassen & Yoogalingam, 2014 and references therein).

Consistent with the literature, we model the stochastic arrival time deviation u by: (1) normal distribution, $u \sim N(\mu_u, \sigma_u^2)$, with mean $\mu_u = -15$ minutes (Cayirli et al., 2006), 0 minutes and standard deviation $\sigma_u = 10$ minutes, 20 minutes (see, eg, Deceuninck et al., 2018; Klassen & Yoogalingam, 2014); and (2) uniform distribution $U[\underline{u}, \overline{u}]$ with two intervals $u \sim U[-40, 20]$ and $u \sim U[-40, 40]$ (uniform as in Deceuninck et al., 2018). We select the value of the grace period, G, from the set {10,15,20} (eg, $G = 15$ minutes in Deceuninck et al., 2018) and, unless stated otherwise, we consider a rescheduling period of 60 minutes in the TSM-PI (i.e., the model automatically decline a late patient arriving after $G + 60$ minutes past his/her scheduled time).

Finally, we consider two different cost structures for the objective function: (1) Cost1: $c^{\mathrm{w}} = c^{\mathrm{g}} = c^{\mathrm{o}}$; and (2) Cost2: $c^{\mathrm{w}} = 1$, $c^{\mathrm{g}} = 5$, $c^{\mathrm{o}} = 7.5$. For the first cost structure, each of the three objectives is equally important (this is a classical assumption in the domain of outpatient appointment scheduling, see, eg, Berg et al., 2014; Deceuninck et al., 2018; Shehadeh et al., 2019). The second cost structure fixes the $c^{\mathrm{o}}/c^{\mathrm{g}}$ ratio to 1.5 as in Deceuninck et al. (2018), based on OPC practice (see Cayirli et al., 2006; Deceuninck et al., 2018 for detailed discussions). We also adopt the classical assumption of identical waiting and idle time costs, that is, $c^{\mathrm{w}} = c_j^{\mathrm{w}}$ and $c^{\mathrm{g}} = c_j^{\mathrm{g}}$ for all $j = 1, \ldots, P$ (see Ahmadi-Javid et al., 2017; Deceuninck et al., 2018; Shehadeh et al., 2019 and the references therein).

We implemented our TSMs and the MCO approach using the AMPL2016 Programming language calling CPLEX V12.6.2 as a solver with default settings (our experiments showed no consistent benefits in any parameter tuning). We ran all experiments on an HP workstation running Windows Server 2012 with two Intel E5-2620-v4 processor, each with 8-Cores (16 total), 2.10 GHz CPUs, and 128 GB shared RAM.

5.2 | Obtaining near-optimal solutions to the TSMs

For each SOASP instance and cost structure, we implemented the MCO algorithm in Section 4.1 with $N_o = 5$, $N' = 10\,000$, $K = 20$, and $\epsilon = 0.1$ (i.e., Algorithm 1 terminates with N, $\overline{v}_N$, and $\overline{v}_{N'}$ whenever $|\text{AOI}_N| < 0.1$, see step 4).

Tables B.1 and B.2 in Supporting information present the approximate optimality index ($\text{AOI}_N = \frac{\overline{v}_{N'} - \overline{v}_N}{\overline{v}_{N'}}$) and the 95% confidence intervals (95% CI) of the statistical lower and upper bounds ($\overline{v}_N$ and $\overline{v}_{N'}$, receptively) on the objective values of TSM-PI and TSM-AO for each SOASP instance under Cost1 and Cost2, respectively, at the termination of Algorithm 1. In Section 5.3, we present and analyze the ranges of solution time of the TSM-PI and TSM-AO formulations for each SOASP instance. Herein we summarize the key findings of the MCO algorithm.

Clearly, $N = 100$ and $N = 200$ scenarios of service durations and arrival time deviations are sufficient to obtain a near-optimal solution to the TSM-PI and TSM-AO via their SAA formulations. First, AOI_N at $N = 100{,}200$ ranges from 0.00 to 0.09 and from 0.00 to 0.07 for the TSM-PI and TSM-AO, respectively. In fact, only 2 and 6 of the 60 SOASP instances terminated with AOI = 0.09 and AOI = 0.07, respectively, and the majority of AOI_N at $N = 100{,}200$ are less than or equal to 0.05. Second, the 95%CI of $\overline{v}_N^{\mathrm{PI,AO}}$ and $\overline{v}_{N'}^{\mathrm{PI,AO}}$

TABLE 3 Characteristics and parameters of the SOASP instances

Parameter	Value/distribution
Number of patients, P	12 patients
Patient mix	Hetero1: (9 return, 3 new), Hetero2: (6 return, 6 new), Homo1: (12 return), Homo2: (12 new)
Planned provider service hours, $[0, L]$	[0, expected duration for the patient mix]
Service time distributions	$d(\text{return}) \sim \text{Log}\, N(20, 16^2)$, $d(\text{new}) \sim \text{Log}\, N(30, 24^2)$
Arrival time deviation distributions	$N(0, 10^2)$, $N(-15, 10^2)$, $N(-15, 20^2)$, $U[-40, 20]$, $U[-40, 40]$
Grace periods, G	10, 15, 20
Rescheduling period, R	60 min

with N = 100, 200 are very tight (i.e., have a small variance). These results qualify $\bar{v}_N^{\text{PI}}$ and $\bar{v}_{N'}^{\text{AO}}$ as tight statistical estimates for the lower and upper bounds on the optimal value of each SOASP instance, respectively. In Section 5.4, we compute and analyze the statistical relative gap (ARG $= \frac{\bar{v}_{N'}^{\text{AO}} - \bar{v}_N^{\text{PI}}}{\bar{v}_N^{\text{PI}}}$) between these two bounds.

Recall that during this experiment, we use $K = 20$ as the number of replication in the MCO Algorithm (see step 1). The tightness of the 95% CI of $\bar{v}_N$ and $\bar{v}_{N'}$ suggests that $K = 20$ replications are sufficient to get a tight confidence interval for both bounds. In other words, the standard error estimates are small enough so that (statistically) we have high confidence in our conclusions about these bounds.

5.3 | Computation time of the TSM-AO and TSM-PI

Table C.1 in Supporting information presents the range of the average solution time for the TSM-AO and TSM-PI (reported as average solution time with $N = 5$ scenarios-200 scenarios) across 20 random instances for each combination of patient mix, stochastic arrival distribution, and grace period. First, we note that the TSM-AO enjoys an outstanding computational behavior independent of patient mix (equivalently, distributions of service durations), distribution of arrival time deviations, length of the grace period, and the cost structure. Solution time of the TSM-AO ranges from 0.01 seconds (with $N = 5$ scenarios) to 66 seconds (with $N = 200$ scenarios), demonstrating its implementability (i.e., can be easily translated into standard optimization software packages, not requiring customized algorithmic development or tuning). Note that larger sample sizes than $N = 200$ increased the solution times of this TSM without consistent and significant improvement in the AOI_N and the approximate relative gap that we evaluate in Section 5.4. We refer interested readers to Appendix D (Supporting information) for an example on the scalability of the TSM-AO with larger P and N.

Second, the computational behavior of the TSM-PI depends on patient mix, distribution of arrival time deviations, and length of the grace period. For example, the TSM-PI takes longer times to solve the heterogeneous (Hetero) instances than homogeneous (Homo) instances. The TSM-PI solution time (in seconds) ranges from 0.01 (with $N = 5$ scenarios) to 7200 (with $N = 200$ scenarios) and from 0.01 (with $N = 5$ scenarios) to 3700 (with $N = 200$ scenarios) for the Hetero and Homo instances, respectively. Moreover, the TSM-PI solution time increases as the variability of arrival time deviation increases and as the grace period shortens. For example, for Hetero1 and a grace period of 10 minutes, increasing the variability from $u \sim N(-15, 10^2)$ to $u \sim N(-15, 20^2)$ widens the range of the TSM-PI solution time from 0.2-13 to 15-5312 seconds, respectively. In contrast, increasing the grace period from 10 to 20 minutes under $u \sim N(-15, 20^2)$ narrowed the range of the TSM-PI solution time from 15-5312 to 0.2-17 seconds.

We attribute the longer solution time of the TSM-PI for the instances with heterogeneous patient types, higher variability in arrival time deviations, and shorter grace periods to the following. First, heterogeneous instances require a larger number of binary decision variables representing the initial appointment sequencing, which leads to larger MIP models. Second, under higher variability in arrival time deviation, a larger portion needs to be rescheduled (i.e., arrive within interval R or DE). This results in a larger search space of the rescheduling decisions in the second-stage formulation of the TSM-PI.

Finally, we note that, without strengthening the TSM-PI (see Section 4.2), we were unable to solve any of the 60 TSM-PI instances or even smaller instances with 3 patients and 10 scenarios. As we solve the TSM-PI only to obtain a statistical lower bound for SOASP and evaluate the (sub-)optimality of the AO policy, we did not attempt to further improve the solution efficacy of TSM-PI under challenging parameter settings.

5.4 | Evaluating ARG and the optimality of the AO policy

Table 4 presents ARG% $= \frac{\bar{v}_{N'}^{\text{AO}} - \bar{v}_N^{\text{PI}}}{\bar{v}_N^{\text{PI}}} \times 100\%$. Recall that ARG is a conservative approximation of the Gap $= \frac{v^{\text{AO}} - v^{\text{PI}}}{v^{\text{PI}}}$ between the two TSM approximations of SOASP. Hence, small ARG% indicates that the AO policy is near-optimal. We observe the following about ARG% and the optimality of the AO policy.

First, the AO policy is near-optimal in a wide range of parameter settings. Specifically, the ARG% is small if (i) the

TABLE 4 ARG% for each SOASP instance under two cost structures: (1) Cost1: $c^w = c^g = c^o$; and (2) Cost2: $c^w = 1$, $c^g = 5$, $c^o = 7.5$

	$N(0,10^2)$		$N(-15,10^2)$		$N(-15,20^2)$		$U[-40,20]$		$U[-40,40]$	
G	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2
Hetero1										
10	**15%**	**16%**	3%	1%	**15%**	**12%**	**14%**	**28%**	**42%**	**65%**
15	0.3%	2%	3%	0.1%	2%	4%	7%	7%	**35%**	**43%**
20	1%	2%	2%	0.4%	3%	2%	1%	6%	**23%**	**29%**
Hetero2										
10	**16%**	**15%**	3%	1%	**10%**	**11%**	**22%**	**28%**	**25%**	**64%**
15	6%	2%	3%	1%	3%	4%	7%	5%	**28%**	**47%**
20	5%	3%	3%	0.9%	3%	0.7%	7%	5%	**25%**	**17%**
Homo1										
10	**10%**	**20%**	6%	2%	**14%**	**19%**	**14%**	**21%**	**29%**	**59%**
15	2%	5%	4%	3%	6%	8%	3%	6%	**26%**	**37%**
20	0.3%	1%	3%	0.3%	1%	6%	1%	1%	**16%**	**31%**
Homo2										
10	**12%**	**18%**	4%	3%	**13%**	**18%**	**13%**	**20%**	**20%**	**73%**
15	4%	7%	4%	3%	6%	7%	7%	8%	**29%**	**54%**
20	1%	3%	4%	3%	4%	4%	5%	1%	**25%**	**35%**

Note: Results are based on the average across 20 random instances for each combination of patient mix, stochastic arrival distribution, grace period and cost structures. ARG% values greater than or equal to 10% are marked in bold.

patients arrive on time or early on average with a small variability in arrival time or (ii) the grace period is relatively long. For example, ARG% ranges from 0.1% to 6% when $u \sim N(-15, 10^2)$ and a majority of ARG% are less than 7% when G = 15, 20. This makes sense because, under these parameter settings, the majority of the patients are on-time or early (i.e., arrive within interval E or G) and so should be served in their scheduled order (i.e., according to AO policy). Indeed resequencing may seem meaningless under these conditions, but we are able to mathematically model it and rigorously prove the near-optimality of AO.

Second, the AO policy becomes suboptimal under high variability in arrival time and short grace periods. For example, with $G = 10$ minutes, increasing the variability from $u \sim N(-15, 10^2)$ to $u \sim N(-15, 20^2)$ increased the range of ARG% from 1%-6% to 10%-19%. In addition, ARG% is significantly larger when the patients are likely to arrive after their grace periods (eg, with $G = 10$ and $u \sim U[-40, 40]$, $\mathbb{P}(u > G) \approx 0.38$; see columns 10 and 11 in Table 4). This is because, in this case, the AO policy declines all patients arriving in interval R or DE, which yields unnecessary provider idling. In contrast, TSM-PI may reschedule some of the late patients arriving within interval R.

Third, in all parameter settings, the AO policy becomes closer to optimal (i.e., ARG% decreases) as the grace period lengthens. This is particularly significant when there is a high degree in patient lateness (eg, when $u \sim U[-40, 40]$) and when idling is costly (see, eg, the last column of Table 4). Intuitively, the longer the grace period, the lower the probability of arriving beyond the grace period (i.e., within interval R or DE) and thus the less the number of patients to be rescheduled. For example, with $u \sim U[-40, 40]$ and as G increases from 10 to 20, the probabilities of arriving beyond the grace period decreases from $\mathbb{P}(u > 10) \approx 0.38$ to $\mathbb{P}(u > 20) = 0.25$.

Finally, it is interesting to note that we obtain similar results on the (sub)optimality of the AO policy with a larger number of patients in each mix. For example, when we increase the number of patients from $P = 12$ to 16 and 20 in Hetero2 ($0.5P$ return, $0.5P$ new), the AO policy remains near-optimal under $u \sim N(-15, 10^2)$ and Cost2 with ARG% ranges from 0.3% to 4% and from 1% to 2%, respectively (see Appendix D in Supporting information for the sample size and solution times). Additionally, we obtain similar results on the (sub)optimality AO-policy when we relax assumption A3, that is, when arrival deviations depend on the scheduled time or patient type. For example, when return patients (who are familiar with the clinic policy) are punctual and the arrival time deviations of new patients (who are less familiar with the clinic policy) follow $N(-15, 10^2)$ and $U[-40, 40]$, ARG% for a mix of Hetero2 with $P = 12$ ranges from 0.07% to 1% and from 10% to 32%, respectively. Note that to test such instances of the problem, we do not need to modify TSM-AO or TSM-PI. Instead, we adjust parameter u based on patient type. When the arrival deviations depend on the scheduled time, we adjust u based on the order of the scheduled time.

5.5 | Alternative rescheduling policy

We propose and evaluate the performance of an alternative rescheduling policy, which we call Neighbor Swapping (NS), for those parameter settings under which the AO policy is suboptimal. The NS policy has two schemes; a priority queueing scheme (in Algorithm 2) and a serving scheme (in Algorithm 3). Given the initial schedule, a patient receptionist can implement the

Algorithm 2: Neighbor Swapping Policy–Priority Queueing Scheme

Input: scheduled arrival times $\{t_i\}_{i=1}^{P}$, grace period (G), rescheduling period (R)

for *(all patients arriving at the clinic)* **do**
  Denote the set of the initial positions of patients currently waiting as $\mathcal{W}$.
  Denote the initial position of the current arrival (patient) as i.
  if $a_i > t_i + G + R$ **then** /* if the patient i arrives in interval DE. */
    Decline this patient.
  else
    if $a_i \leq t_i + G$ **then** /* if the patient i arrives in interval E or G. */
      This patient i has a lower priority than all patients in $\{j \in \mathcal{W} : j \leq i-1\}$, and a higher priority than all patients in $\{j \in \mathcal{W} : j \geq i+1\}$. /* respects the appointment order of on-time patients. */
    end
    if $t_i + G < a_i \leq t_i + G + R$ *and* $a_i \leq a_{i+1}$ **then** /* if the patient i arrives in interval R and earlier than patient $(i+1)$. */
      This patient i has a lower priority than all patients in $\{j \in \mathcal{W} : j \leq i-1\}$, and a higher priority than all patients in $\{j \in \mathcal{W} : j \geq i+1\}$./* keep the order of i and $i+1$. */
    end
    if $t_i + G < a_i \leq t_i + G + R$ *and* $a_i > a_{i+1}$ **then** /* if the patient i arrives in interval R and later than patient $(i+1)$. */
      This patient i has a lower priority than all patients in $\{j \in \mathcal{W} : j \leq i-1\} \cup \{i+1\}$, and a higher priority than all patients in $\{j \in \mathcal{W} : j \geq i+2\}$. /* swap the order of i and $i+1$. */
    end
  end
end

NS policy as follows: (i) using the priority queueing scheme, s/he maintains a priority queue of patients waiting for service and updates the priority whenever a patient arrives, and (ii) using the serving scheme, s/he decides whether the provider, whenever idle, should stay idle or start serving the patient with the highest priority in the queue.

Note that the NS policy respects the AO of early and punctual arrivals and de-prioritizes or even declines late patients. These two properties are particularly demanded in both the OPC practice and the appointment scheduling literature (see, eg, Deceuninck et al., 2018; Glowacka et al., 2017, and the references therein).

We evaluate the performance of the NS policy for those parameters settings under which the AO policy is sub-optimal (i.e., SOASP instances with large ARG% values, as marked in bold in Table 4). We did not formulate another TSM to find the initial schedule required to implement the NS policy. Instead, we first fix the initial appointment sequence and the corresponding scheduled arrival time to the optimal ones to the TSM-AO for two reasons. First, to guarantee a fair comparison between the performance of the NS and AO policy (given that the actual arrival time equals the scheduled time plus the arrival time deviation). Second, in contrast to the TSM-AO, which is easy to solve, implementing the two schemes of the NS policy in a third TSM is challenging and would result in a complex TSM to solve (due to the need for new binary second-stage variables and constraints to employ the priority queueing, swapping, and serving rules). While it is interesting to study the optimality of the NS policy rigorously, we leave it as future work.

Second, for each $R \in \{5,10,0.5G,\ G,\ 2G,\ 3G\}$ we simulate appointment system under the NS policy for $N_{\mathrm{NS}} = 10\,000$ scenarios of service durations and arrival time deviations (d_p^n, u_p^n) for all $p = 1, \ldots, P$ and $n = 1, \ldots, N_{\mathrm{NS}}$. For each $n = 1, \ldots, N_{\mathrm{NS}}$, we compute w^n, g^n, w_{P+1}^n, and $v^{\mathrm{NS}} = \frac{1}{N_{\mathrm{NS}}} \sum_{n=1}^{N_{\mathrm{NS}}} \sum_{j=1}^{P} (c_j^{\mathrm{w}} w_j^n + c_j^{\mathrm{g}} g_j^n) + c^{\mathrm{o}} w_{P+1}^n$. We repeat this simulation for 20 times, each time with a new sample of N_{NS} scenarios. Finally, we take the average of the 20 objective function values to obtain $\bar{v}^{\mathrm{NS}}$.

Note that the NS policy is a feasible rescheduling policy to SOASP. Therefore, $\bar{v}^{\mathrm{NS}}$ serves as a statistical upper bound on the optimal value of SOASP. Accordingly, we compute ARG% under NS as $\frac{\bar{v}^{\mathrm{NS}} - \bar{v}_N^{\mathrm{PI}}}{\bar{v}_N^{\mathrm{PI}}}$.

For those parameter settings under which the AO policy is suboptimal (i.e., SOASP instances with large ARG% values,

Algorithm 3: Neighbor Swapping Policy–Serving Scheme

Input: scheduled arrival times $\{t_i\}_{i=1}^{P}$, grace period (G), rescheduling period (R)

while *(provider is idle)* **do**
  Denote the set of the initial positions of patients currently waiting as $\mathcal{W}$.
  Denote i^* as the initial position of the patient with the highest priority in $\mathcal{W}$.
  Denote the current clock time as T.
  if $\mathcal{W} = \emptyset$ **then** /* no patients waiting for service. */
    The provider should wait.
  else
    if $\mathcal{W} \neq \emptyset$ *and* $T \geq t_j + G$ *for all* $j = 1, \ldots, (i^* - 1)$ **then** /* all patients j, with $j \leq i^* - 1$, either are late or have been served. */
      The provider should start serving patient i^*. BREAK.
    else
      The provide should wait.
    end
  end
end

as marked in bold in Table 4), Figure 2 compares ARG% under the AO and NS policies for the Hetero and Homo instances. In addition, Table E.1 in Supporting information reports the best selection of rescheduling period R^* from {5,10,0.5G, G, 2G, 3G} associated with the minimum (i.e., tightest) $\overline{v}^{NS}$ in each parameter setting.

We observe from Figure 2 that the NS policy significantly and consistently shrinks the ARG% in all challenging parameter settings (i.e., those with high variability in arrival time and short grace periods). For example, the performance of the NS policy is very close to the optimal policy when $G = 10$ and $u \sim \{N(0, 10^2), N(-15, 20^2), U[-40, 20]\}$ with ARG% ranging from 2% to 15% as compared to 10%–28% under the AO policy. In addition, when $u \sim U[-40, 40]$, NS cuts the ARG% by half on average from those obtained under AO. This indicates that the NS policy can effectively reschedule the appointments and avoid unnecessary idling under these settings. As shown in Table F.1 in Supporting information, the average number of declined appointments per day under NS is approximately zero when $G = 10$ and $u \sim \{N(0, 10^2), N(-15, 20^2), U[-40, 20]\}$, and is significantly less than under the AO policy when $u \sim U[-40, 40]$. This is another important property of the NS policy because a declined patient will have to be rescheduled for another day, which is inconvenient to the patient and has a cost to the system.

Finally, we note that we repeat the same simulation using the optimal schedule to the TSM-PI as the initial schedule. The obtained results on the NS policy are similar. We also obtained similar results on the NS policy when arrival time deviations depend on patient type. For example, when return patients are punctual and the arrival time deviations of new patients follow $U[-40, 40]$, the NS policy shrinks the ranges of ARG% for a mix of Hetero2 from 10%–32% (under AO policy) to 5%-11%. Additionally, we use the same simulation steps to evaluate the performance of the NS policy for those parameter settings under which AO policy is near-optimal (i.e., SOASP instances with small ARG% values in Table 4). The obtained ARG% values under NS are very similar to ARG% values obtained under AO policy, demonstrating the near-optimality of NS under those parameter settings. This makes sense because under these settings the majority of patients are punctual or early, and so should be served in the AO according to both the priority queuing and serving schemes. These results motivate using NS as a global rescheduling policy. Practitioners should thus decide whether it is more appropriate to adopt the AO policy or NS policy in their particular OPC and can implement the TSM-AO to find the initial appointment schedule.

5.6 | Analysis of the optimal scheduling patterns

In this section, we study the sensitivity of the optimal schedule of the TSM-AO to SOASP parameters. First, we compare the optimal scheduling patterns of the TSM-AO with punctual (i.e., $u_p = 0$ for all $p = 1, \ldots, P$,) and random arrivals. For presentation brevity and illustration purposes, we present results for Hetero2 (6 return, 6 new). Figures 3 and 4 present the optimal schedules under Cost1 and Cost2, respectively. The point $(x, y) = (i, \text{alloted time}_i)$ of every schedule in each subfigure corresponds to the optimal time assigned for each appointment $i = 1, \ldots, 12$. We observed the following about these optimal schedules.

Under all combinations of cost structures, distributions of arrival time deviations, and grace periods the optimal sequence always starts with a block of consecutive return patients (appointments 1-6) followed by a block of

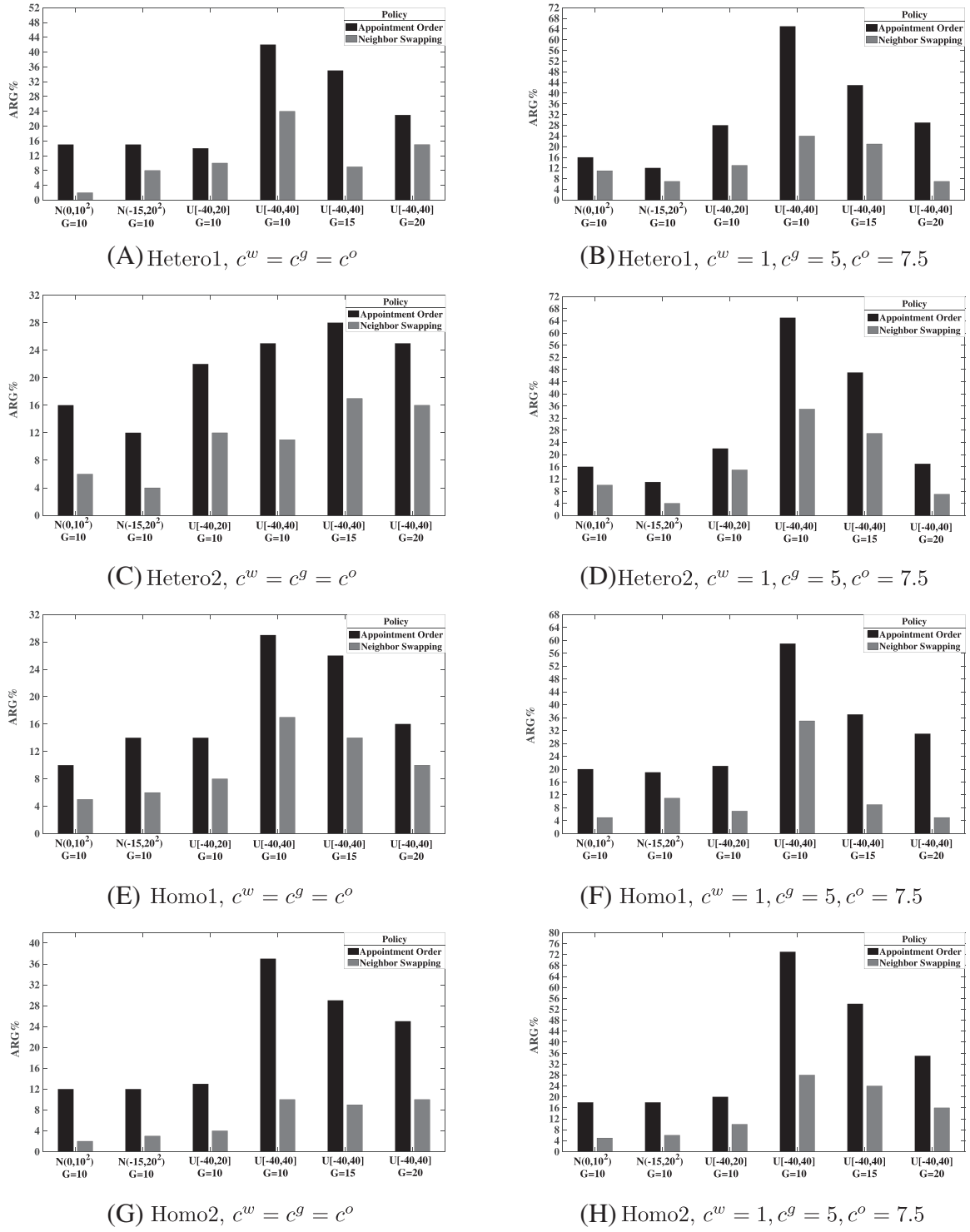


FIGURE 2 Comparisons of the ARG% values under appointment order and neighbor swapping policies

consecutive new patients (appointments 7-12). In addition, more time is assigned to new patients than return patients. Intuitively, by scheduling return patients first and assigning more time for new patients, the TSM-AO attempt to mitigate delays that may accumulate due to the longer service time of new patients. Under Cost1 and punctual arrivals, the optimal time allowances in each block form a "dome-shape" (i.e., time allowances between appointments first increase then decrease) with the last appointment scheduled at the end of the day.

As compared to the optimal schedules of punctual arrivals, the optimal time allowances between random arrivals form a "zigzag-shape," that is, time allowances increase and decrease alternatively between successive appointments in each block. When $u \sim N(-15,\ 10^2)$ and $G = 15$ and 20, however, the optimal scheduled times are very close to those when $u = 0$ because the majority of patients arrives before the end of the grace period under these settings. Under higher variability in arrival times, the optimal scheduled times in each block are very different. For example, increasing the variability from

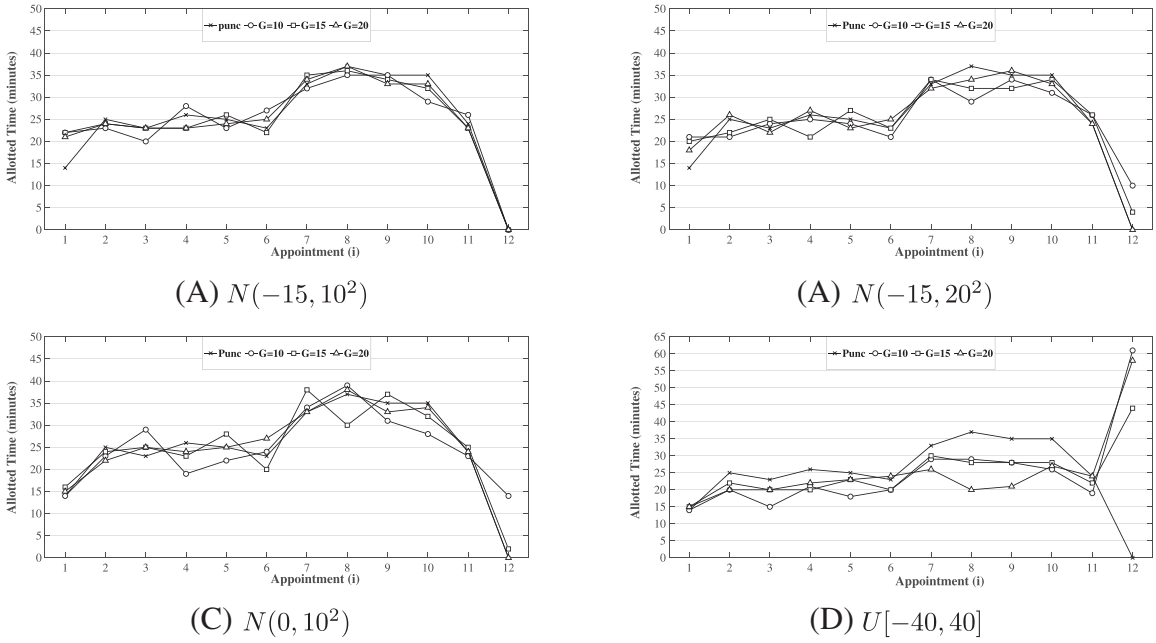


FIGURE 3 Optimal appointment schedules under Cost1. Appointments 1-6 are return, and 7-12 are new

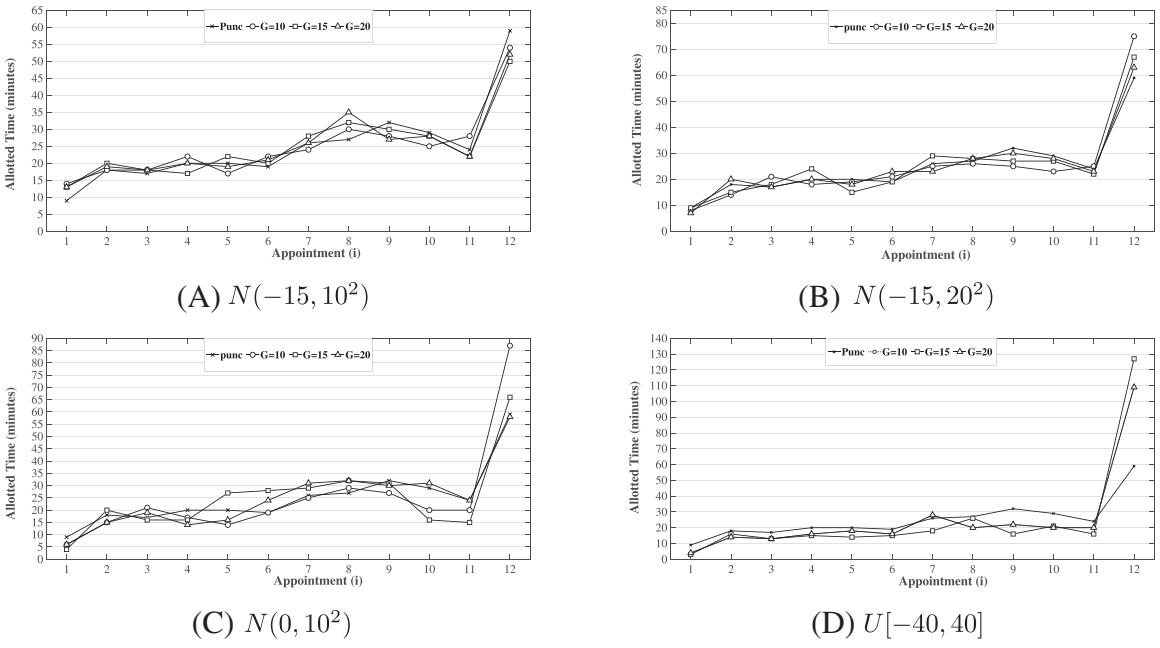


FIGURE 4 Optimal appointment schedules under Cost2. Appointments 1-6 are return, and 7-12 are new

$u \sim N(-15,\ 10^2)$ to $u \sim N(-15,\ 20^2)$ with $G = 20$ changes the time allocated for appointments 1-4 from 22-24-23-23 to 18-26-22-27. When $u \sim U[-40,\ 40]$, the TSM-AO allocates shorter time for each appointment and a longer time for the last appointment than with other distributions of u. This makes sense because when $u \sim U[-40, 40]$, patients are likely to arrive after the grace period, and so scheduling a long time between appointments may lead to provider idling.

The optimal scheduling patterns under Cost2, in which provider time is more important than patient time, are similar to those under Cost1. However, a shorter time is assigned per appointment and a longer time is assigned for the last appointment than under Cost1. Additionally, to minimize the risk of idling, the first two appointments are scheduled very close to each other when there is high variability in arrival times and patients are likely to arrive late, that is, under $u \sim \{N(-15,\ 20^2),\ N(0,\ 10^2),\ U[-40,\ 40]\}$. Finally, under this cost structure, the last patient is not scheduled at the end of the day (i.e., at L) as under Cost1.

Second, we investigate the effect of heterogeneous distributions of arrival time deviations on the optimal sequencing decisions. For this analysis, we assume that all patients are of the same type but have different arrival behavior. Specifically, we consider a class of Homo1 instances (i.e., 12 return patients with $d_p \sim (20,\ 16^2)$ for all $p = 1,\ \ldots,\ 12$) in which half of the patients have a "low variance" arrival distribution

TABLE 5 Optimal sequencing patterns with heterogenous random arrivals

		Cost1: $c^w = c^g = c^o$											
		Appointment											
Dist of H		1	2	3	4	5	6	7	8	9	10	11	12
$N(0, 10^2)$	10	L	L	L	H	H	H	H	L	H	H	L	L
	20	L	H	H	H	H	L	H	H	L	L	L	L
$N(0, 20^2)$	10	L	L	L	L	H	H	H	H	H	H	L	L
	20	L	L	H	H	L	L	H	H	H	H	L	L
$N(0, 30^2)$	10	L	L	H	H	L	H	H	L	H	H	L	L
	20	L	L	H	H	H	L	L	H	H	H	L	L
$N(-15, 20^2)$	10	L	L	L	H	H	H	H	H	L	L	L	H
	20	L	L	L	L	H	H	H	H	H	H	L	L
$U[-40, 40]$	10	L	L	H	H	H	H	H	H	L	L	L	L
	20	L	L	H	H	H	H	H	H	L	L	L	L
		Cost2: $c^w = 1, c^g = 5, c^o = 7.5$											
		Appointment											
Dist of H	G	1	2	3	4	5	6	7	8	9	10	11	12
$N(0, 10^2)$	10	L	L	H	H	H	H	H	L	L	L	H	L
	20	L	L	H	H	H	H	L	H	L	L	L	L
$N(0, 20^2)$	10	L	L	H	H	L	L	H	H	H	H	L	L
	20	L	L	H	H	L	L	H	H	H	H	L	L
$N(0, 30^2)$	10	L	L	L	H	H	H	H	L	H	H	L	L
	20	L	L	H	H	H	H	L	L	H	H	L	L
$N(-15, 20^2)$	10	L	L	L	H	H	L	H	H	H	H	L	L
	20	L	L	L	H	H	L	H	H	H	H	L	L
$U[-40, 40]$	10	L	L	L	H	H	H	H	H	H	L	L	L
	20	L	L	L	H	H	H	H	H	H	L	L	L

Note: L is low variance and H is high variance.

($u \sim N(-15, 10^2)$) and the other half have the following "high variance" arrival distributions: $N(0, 10^2)$, $N(0, 20^2)$, $N(0, 30^2)$, $N(-15, 20^2)$, and $U[-40, 20]$. The optimal sequencing decisions in Table 5 show that to minimize costs it is better to schedule patients who are early on average and have low variability in arrival time at the beginning of the day to reduce the risk of idling and at the end of the day to reduce the risk of overtime.

Finally, we note that when return and new patients have different distributions of u and when arrival deviations depend on the time of the day, the optimal sequence is similar to when the distribution of u is the same for all patients and is independent of the scheduled time, that is, return appointments are scheduled first.

5.7 | Value of modeling stochastic arrivals

In this section, we demonstrate the benefit of modeling stochastic arrivals in outpatient appointment scheduling. For each SOASP instance, we solve the TSM-AO twice, with one assuming punctual arrivals (i.e., zero arrival time deviations) and the other considering stochastic arrivals. We denote the obtained schedules as S^0 and S^u, respectively. We then evaluate the cost $E[Q^{AO}(x, t, \xi)]$ of S^0, denoted as C^0, by reevaluating the schedule S^0 under a set of 10 000 scenarios with stochastic arrivals, while C^u denotes the corresponding cost of S^u. We compute the relative increase in cost $RC^{(punc)} := \frac{C^0 - C^u}{C^u} \times 100\%$.

We also compare S^u with the well-known and widely employed deterministic schedule S^m that assigns appointment slot to each patient by the mean service duration of his/her type. For all heterogeneous instances, we use the optimal sequence to the TSM-AO to assign patient types to appointments in S^m and fix the length of each appointment slot to the mean service duration of the assigned type to it. We then obtain the cost C^m of ignoring uncertainty by reevaluating S^m under a set of 10 000 scenarios with stochastic arrivals and service durations. We compute the relative increase in cost $RC^{(mean)} := \frac{C^m - C^u}{C^u} \times 100\%$.

Table 6 presents $RC^{(punc)}$ and $RC^{(mean)}$, including the average and maximum among all combinations of patient mix and distribution of stochastic arrival time deviation. From this table, we observe that the cost of S^u (i.e., considering stochastic arrivals) is significantly lower than that of S^0 (i.e., ignoring stochastic arrivals). Overall, the average and

TABLE 6 The (average, maximum)% relative cost gaps $RC^{(punc)}$ and $RC^{(mean)}$ as functions of patient mix and stochastic arrival distributions

	$N(0,10^2)$		$N(-15,10^2)$		$N(-15,20^2)$		$U[-40,20]$		$U[-40,40]$	
	Cost1	**Cost2**	**Cost1**	**Cost2**	**Cost1**	**Cost2**	**Cost1**	**Cost2**	**Cost1**	**Cost2**
Hetero1										
$RC^{(punc)}$	(9, 24)%	(16, 23)%	(7, 22)%	(8, 22)%	(10, 26)%	(8, 22)%	(10, 22)%	(5, 13)%	(28, 32)%	(45, 52)%
$RC^{(mean)}$	(27, 44)%	(12, 28)%	(33, 57)%	(9, 23)%	(29, 56)%	(17, 32)%	(27, 47)%	(21, 32)%	(29, 39)%	(32, 43)%
Hetero2										
$RC^{(punc)}$	(11, 20)%	(13, 26)%	(7, 26)%	(6, 18)%	(9, 24)%	(9, 26)%	(8, 16)%	(10, 20)%	(21, 27)%	(64, 70)%
$RC^{(mean)}$	(21, 30)%	(24, 43)%	(26, 55)%	(9, 24)%	(22, 44)%	(13, 37)%	(22, 29)%	(17, 32)%	(24, 33)%	(70, 80)%
Homo1										
$RC^{(punc)}$	(8, 23)%	(8, 21)%	(7, 23)%	(8, 24)%	(8, 17)%	(9, 26)%	(7, 18)%	(8, 21)%	(19, 24)	(31, 36)%
$RC^{(mean)}$	(13, 36)%	(8, 21)%	(23, 53)%	(9, 25)%	(18, 34)%	(29, 36)%	(16, 31)%	(15, 26)%	(10, 15)%	(48, 57)%
Homo2										
$RC^{(punc)}$	(10, 20)%	(13, 29)%	(6, 21)%	(12, 28)%	(7, 17)%	(10, 22)%	(8, 19)%	(11, 26)%	(24, 29)%	(58, 63)%
$RC^{(mean)}$	(19, 24)%	(21, 42)%	(19, 44)%	(9, 26)%	(18, 35)%	(13, 32)%	(18, 35)%	(18, 41)%	(26, 37)%	(81, 92)%

Note: Results are based on 20 random instances for each parameter setting.

maximum $RC^{(punc)}$ range in 7%-64% and 13%-70%, respectively (see Tables G.1-G.2 in Supporting information for the improvement in scheduling metric). In addition, we observe that the cost of S^u (i.e., considering stochasticity) is significantly lower than that of S^m (i.e., ignoring stochasticity). The average and maximum $RC^{(mean)}$ range in 8%-81% and 15%-92%.

6 | CONCLUSION

In this paper, we studied SOASP for OPC scheduling under stochastic arrival times and service durations. We consider for an OPC manager who needs to design an appointment schedule and a rescheduling policy for a single provider and a set of patients, where each patient has a known probability distribution of arrival time deviations and service durations. The objective is to minimize the expected total cost of patient waiting time, provider idle time, and provider overtime.

By deriving two-stage approximations of SOASP and testing them in extensive numerical experiments, we show that the AO policy is near-optimal in a wide range of realistic parameter settings. Specifically AO is near-optimal when (i) the patients arrive on time or early on average with a small variability in arrival time or (ii) the grace period is relatively long. We also identify parameter settings that result in sub-optimality of the AO policy. The AO policy becomes sub-optimal under high variability in arrival time and short grace periods. Accordingly, we propose an alternative policy based on NS that leads to a significantly better performance.

Our NS policy respects the AO of early and punctual arrivals and de-prioritizes or even declines late patients. These two properties are particularly demanded in the OPC practice and the appointment scheduling literature. Additionally, the average number of declined appointments per day under NS is significantly less than under the AO policy. This is another important property of the NS policy because a declined patient will have to be rescheduled for another day, which is inconvenient to the patient and incurs a cost to the system.

We suggest following areas for future research. First, our results give insights on when to use the AO or the NS policy under different distributions of arrival time deviations, grace, rescheduling, and declining periods as reported in the literature. Although practitioners can use these results to decide which policy to use, it may be interesting to optimize the value of these parameters in a more data-driven and tractable approach. Second, we would like to extend our approach to include additional sources of uncertainty, particularly variability in patient no-show behavior. Third, we are interested in studying trade-offs between scheduling metrics such as provider work load (i.e., number of scheduled appointments), patient access delays (i.e., the length of time a patient has to wait from the appointment request until a scheduled appointment become available to them), and impact on patient outcomes when declining (rescheduling) an appointment.

Fourth, our current approach assumes that the probability distributions of service durations and arrival times are known. We are interested in extending this work for the settings where such probability distributions are ambiguous. Fifth, we are interested in studying the (sub)optimality of the NS policy rigorously both for single-server and more complicated settings (eg, multiple servers). Sixth, our approach assumes a fixed number of patients with known types. We seek to use the results of this research to develop templates and policies for scheduling patients dynamically as they randomly request future appointments considering patient and provider preferences and random service duration, arrival time, and no-show. Finally, a major limitation of our TSM approach is that the TSM-PI is very challenging to solve when there is high variability in arrival time deviations, specifically when the majority of the patients arrive late and are reschedulable. And so, it may be interesting to investigate exact and

heuristic methods to improve the solvability of this TSM for large and challenging instances of SOASP that may appear in reality.

ACKNOWLEDGMENTS

This work is supported in part by the Agency for Healthcare Research and Quality (P30HS024385), the National Science Foundation (CMMI-1662774), and the University of Michigan Center for Healthcare Engineering and Patient Safety (CHEPS). We are also grateful for the insights of Prof Marina A. Epelman and our clinical collaborators Dr Jacob Kurlander and Dr Sameer Saini. Karmel S. Shehadeh dedicates her effort in this paper to every little dreamer in the whole world who has a dream so big and so exciting. Believe in your dreams and do whatever it takes to achieve them-the best is yet to come for you.

ORCID

Karmel S. Shehadeh https://orcid.org/0000-0001-7842-0951

REFERENCES

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

Ahmed, A., Chou, C.-A., & Shehadeh, K. S. (2014). Joint effect of appointment scheduling rules and resource selection on outpatient clinics performance. In *IIE Annual Conference. Proceedings* (p. 3729). Institute of Industrial and Systems Engineers (IISE).

Alexopoulos, C., Goldsman, D., Fontanesi, J., Kopald, D., & Wilson, J. R. (2008). Modeling patient arrivals in community clinics. *Omega*, *36*(1), 33–43.

Berg, B. P., Denton, B. T., Erdogan, S. A., Rohleder, T., & Huschka, T. (2014). Optimal booking and scheduling in outpatient procedure centers. *Computers & Operations Research*, *50*, 24–37.

Birge, J. R., & Louveaux, F. (2011). Introduction to stochastic programming. Springer Science & Business Media.

Brahimi, M., & Worthington, D. (1991). Queueing models for out-patient appointment systems – A case study. *Journal of the Operational Research Society*, *42*(9), 733–746.

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

Cayirli, T., Veral, E., & Rosen, H. (2006). Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science*, *9*(1), 47–58.

Cayirli, T., Veral, E., & Rosen, H. (2008). Assessment of patient classification in appointment system design. *Production and Operations Management*, *17*(3), 338–353.

Cayirli, T., & Yang, K. K. (2014). A universal appointment rule with patient classification for service times, no-shows, and walk-ins. *Service Science*, *6*(4), 274–295.

Chen, R. R., & Robinson, L. W. (2014). Sequencing and scheduling appointments with potential call-in patients. *Production and Operations Management*, *23*(9), 1522–1538.

Cheong, S., Bitmead, R. R., & Fontanesi, J. (2013). Modeling scheduled patient punctuality in an infusion center. *Lecture Notes in Management Science*, *5*, 46–56.

Cox, T. F., Birchall, J. P., & Wong, H. (1985). Optimising the queuing system for an ear, nose and throat outpatient clinic. *Journal of Applied Statistics*, *12*(2), 113–126.

Deceuninck, M., Fiems, D., & De Vuyst, S. (2018). Outpatient scheduling with unpunctual patients and no-shows. *European Journal of Operational Research*, *265*(1), 195–207.

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

Fetter, R. B., & Thompson, J. D. (1966). Patients' waiting time and doctors' idle time in the outpatient setting. *Health Services Research*, *1*(1), 66–90.

Glowacka, K. J., May, J. H., Goffman, R. M., May, E. K., Milicevic, A. S., Rodriguez, K. L., … Vargas, L. G. (2017). On prioritizing on-time arrivals in an outpatient clinic. *IISE Transactions on Healthcare Systems Engineering*, *7*(2), 93–106.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don') work to modify behavior. *Journal of Economic Perspectives*, *25*(4), 191–210.

Gul, S., Denton, B. T., Fowler, J. W., & Huschka, T. (2011). Bi-criteria scheduling of surgical services for an outpatient procedure center. *Production and Operations Management*, *20*(3), 406–417.

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

Gupta, D., & Wang, W.-Y. (2012). *Patient appointments in ambulatory care*. In Handbook of healthcare system scheduling (pp. 65–104). Berlin, Germany: Springer.

Homem-de Mello, T., & Bayraksan, G. (2014). Monte Carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science*, *19*(1), 56–85.

Jiang, B., Tang, J., & Yan, C. (2019). A stochastic programming model for outpatient appointment scheduling considering unpunctuality. *Omega*, *82*, 70–82.

Klassen, K. J., & Rohleder, T. R. (1996). Scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management*, *14*(2), 83–101.

Klassen, K. J., & Yoogalingam, R. (2014). Strategies for appointment policy design with patient unpunctuality. *Decision Sciences*, *45*(5), 881–911.

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

Kocas, C. (2015). An extension of Osuna's model to observable queues. *Journal of Mathematical Psychology*, *66*, 53–58.

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

Lindley, D. V. (1952). *The theory of queues with a single server*. In Mathematical Proceedings of the Cambridge Philosophical Society (Vol. *48*, pp. 277–289). Cambridge University Press.

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

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

Molina-Pariente, J. M., Hans, E. W., & Framinan, J. M. (2018). A stochastic approach for solving the operating room scheduling

problem. *Flexible Services and Manufacturing Journal*, *30*(1–2), 224–251.

Okotie, O. T., Patel, N., & Gonzalez, C. M. (2008). The effect of patient arrival time on overall wait time and utilization of physician and examination room resources in the outpatient urology clinic. *Advances in Urology*, *2008*, 1–3.

Ostrowski, J., Linderoth, J., Rossi, F., & Smriglio, S. (2011). Orbital branching. *Mathematical Programming*, *126*(1), 147–178.

Osuna, E. E. (1985). The psychological cost of waiting. *Journal of Mathematical Psychology*, *29*(1), 82–105.

Rohleder, T. R., & Klassen, K. J. (2000). Using client-variance information to improve dynamic appointment scheduling performance. *Omega*, *28*(3), 293–302.

Salzarulo, P. A., Mahar, S., & Modi, S. (2016). Beyond patient classification: Using individual patient characteristics in appointment scheduling. *Production and Operations Management*, *25*(6), 1056–1072.

Samorani, M., & Ganguly, S. (2016). Optimal sequencing of unpunctual patients in high-service-level clinics. *Production and Operations Management*, *25*(2), 330–346.

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

Shapiro, A., Dentcheva, D., & Ruszczyński, A. (2009). Lectures on stochastic programming: Modeling and theory. Philadelphia, PA: SIAM.

Shehadeh, K. (2019). *Stochastic optimization approaches for outpatient appointment scheduling under uncertainty*. (PhD thesis).

Shehadeh, K. S., Cohn, A. E., & Epelman, M. A. (2019). Analysis of models for the stochastic outpatient procedure scheduling problem. *European Journal of Operational Research*, *279*(3), 721–731.

Shehadeh, K. S., Cohn, A. E., & Jiang, R. (2020). A distributionally robust optimization approach for outpatient colonoscopy scheduling. *European Journal of Operational Research*, *283*(2), 549–561.

Tai, G., & Williams, P. (2012). Optimization of scheduling patient appointments in clinics using a novel modelling technique of patient arrival. *Computer Methods and Programs in Biomedicine*, *108*(2), 467–476.

Welch, J., & Bailey, N. J. (1952). Appointment systems in hospital outpatient departments. *The Lancet*, *259*(6718), 1105–1108.

White, M. B., & Pike, M. (1964). Appointment systems in out-patients' clinics and the effect of patients' unpunctuality. *Medical Care*, *2*, 133–145.

Williams, K. A., Chambers, C. G., Dada, M., McLeod, J. C., & Ulatowski, J. A. (2014). Patient punctuality and clinic performance: Observations from an academic-based private practice pain centre: A prospective quality improvement study. *BMJ Open*, *4*(5), e004679.

Zacharias, C., & Yunes, T. (2018). *Multimodularity in the stochastic appointment scheduling problem with discrete arrival epochs*. Technical report, Working paper. School of Business Administration, University of Miami, Coral.

Zhu, H., Chen, Y., Leung, E., & Liu, X. (2018). Outpatient appointment scheduling with unpunctual patients. *International Journal of Production Research*, *56*(5), 1982–2002.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Shehadeh KS, Cohn AEM, Jiang R. Using stochastic programming to solve an outpatient appointment scheduling problem with random service and arrival times. *Naval Research Logistics* 2021;68:89–111. https://doi.org/10.1002/nav.21933