informs
http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 66, No. 2, February 2020, pp. 667–686
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Managing Appointment-Based Services in the Presence of Walk-in Customers

Shan Wang,[a] Nan Liu,[b] Guohua Wan[a]

[a] Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 200030, China; [b] Operations Management Department, Carroll School of Management, Boston College, Chestnut Hill, Massachusetts 02467

Contact: wangshan_731@sjtu.edu.cn, http://orcid.org/0000-0002-4625-7720 (SW); nan.liu@bc.edu, http://orcid.org/0000-0001-7644-7341 (NL); ghwan@sjtu.edu.cn, http://orcid.org/0000-0002-9298-1023 (GW)

Received: August 5, 2016
Revised: September 21, 2017; June 28, 2018
Accepted: October 14, 2018
Published Online in Articles in Advance: May 21, 2019

https://doi.org/10.1287/mnsc.2018.3239

Abstract. Despite the prevalence and significance of walk-ins in healthcare, we know relatively little about how to plan and manage the daily operations of a healthcare facility that accepts both scheduled and walk-in patients. In this paper, we take a data-analytics approach and develop an optimization model to determine the optimal appointment schedule in the presence of potential walk-ins. Our model is the first known approach that can jointly handle general walk-in processes and heterogeneous, time-dependent no-show behaviors. We demonstrate that, with walk-ins, the optimal schedules are fundamentally different from those without. Our numerical study reveals that walk-ins introduce a new source of uncertainties to the system and cannot be viewed as a simple solution to compensate for patient no-shows. Scheduling, however, is an effective way to counter some of the negative impact from uncertain patient behaviors. Using data from practice, we predict a significant cost reduction (42%–73% on average) if the providers were to switch from current practice (which tends to overlook walk-ins in planning) to our proposed schedules. Although our work is motivated by healthcare, our models and insights can also be applied to general appointment-based services with walk-ins.

History: Accepted by Gad Allon, operations management.
Funding: Financial support from the National Natural Science Foundation of China [Grants 71125003, 71520107003, and 71421002, for which G. Wan is the principal investigator] is gratefully acknowledged.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2018.3239.

Keywords: service operations management • healthcare • appointment scheduling • walk-ins • analytics

1. Introduction

Making an appointment is a common way for customers to get service in many industries. Walk-in customers without appointments (or "walk-ins" for short), however, are often welcome and accepted as well. Providing service to walk-ins benefits a firm in a range of ways, such as increasing revenues, enlarging the customer pool, and building a good business image. To name a few examples, banks accept walk-ins for more business; hotels seldom reject requests from walk-ins if rooms are still available; restaurants rely on walk-ins to build the word of mouth; beauty salons always try to make walk-ins become their regular clients; tech support accepts walk-ins to attract more customers. As walk-ins arrive spontaneously without advance notice, they may interrupt the firm's daily operations, in particular, the service of *scheduled* customers who have set specific arrival times for services.

One industry that often sees the conflict between serving walk-ins and scheduled customers is healthcare. In the outpatient care setting, walk-ins without appointments are usually accepted and constitute a major stream of the customers. In the United States, walk-ins can range from 10% to 60% of the total daily visits to primary care practices; see, for example, Moore et al. (2001) and Cayirli et al. (2008). In the United Kingdom, 63% of genitourinary medicine clinics operate both appointment-based and walk-in services (Djuretic et al. 2001).

Despite the prevalence and significance of walk-ins, we know relatively little about how to plan and manage daily operations of a healthcare facility in the presence of walk-ins. Current practice of outpatient care deals with walk-ins by setting up daily schedule templates, which specify when to schedule an appointment and when, if ever, to intentionally leave open in anticipation for walk-ins. However, there is a lack of scientific understanding on how to set up such a daily template. Most extant literature develops models and insights that can only be applied to an environment free of walk-ins; managing a practice that accepts walk-ins requires fundamentally different tools and guidelines. Without careful planning for walk-ins, daily service operations may be interrupted, resulting in long patient waits, provider overtime work, and, ultimately, poor service quality.

The negative and potentially serious impact on the organization due to not carefully considering walk-ins becomes evident when we interact and collaborate with two large outpatient care systems in New York City (NYC). The first one is a community health center that provides comprehensive medical and dental care to the Central Harlem and Washington Heights areas. Being a Federally Qualified Health Center (FQHC), this facility has to serve all patients regardless of their ability to pay; as a result, more than 15% of the total patient visits to this center are walk-ins (see detailed data in Section 3). However, the administrative team of this center has informed us that walk-ins are "believed to be the main reason for long patient waits." The other organization we interact with is a large community healthcare network, made up of 11 FQHCs located across NYC. One physician told us that, "I know there are always many walk-ins at 10 a.m., but I can't take them. I have appointments [at that time]" (Berman 2016). Undoubtedly, walk-ins have presented a significant challenge in running both organizations, and how to deliver high-quality care services in the presence of these uncertain walk-ins becomes a critical operational issue.

In this paper, we take a data-analytics approach and develop decision models to inform the design of daily schedule templates in outpatient care practices where both scheduled and walk-in patients are accepted. Using a large data set obtained from our first collaborating organization, we find that patient walk-in processes and patterns vary across providers, even in one practice. More importantly, walk-ins may *not* arrive according to the classic (time-inhomogeneous) Poisson process, as often found in the previous literature (Kim and Whitt 2014). In particular, the "zero-event" probability—that is, the chance that no walk-ins arrive in a short time period—may be too large for the Poisson distribution. Motivated by these empirical findings, we develop optimization models that can accommodate general arrival patterns of walk-ins. Specifically, we consider a generic clinic session, with $T > 0$ appointment slots, for a single provider. Throughout the session, a random number of walk-ins may arrive for services according to some arrival process. We are concerned with determining a right number of appointments to schedule and scheduling them to the T slots simultaneously, in anticipation for potential walk-ins that may arrive over time. The objective is to minimize the expected total cost due to patient waiting, provider idling, and overtime.

Another important factor to consider when designing schedule templates in healthcare is patient no-show behavior. Patient no-shows occur when patients miss their booked appointments without early notice or cancellation. Patient no-show rate can range from 1% to 60% depending on practice and patient profiles, and not accounting for patient no-shows may lead to significant operational inefficiency; see Cayirli and Veral (2003), Kopach et al. (2007), Gupta and Denton (2008), and Liu (2016) for detailed discussions on the phenomenon and impact of patient no-shows. Our models and solution approaches can accommodate general patient no-show behaviors as well.

Our contributions in this work can be summarized as follows. To the best of our knowledge, we are among the first to develop analytical optimization models to determine the optimal appointment schedule in the presence of potential walk-ins. Our model is the first known approach that can jointly handle general walk-in processes and heterogeneous, time-dependent patient no-show behaviors. Because of this flexibility, our approach can incorporate almost any finding on these patient behaviors based on empirical data, thus presenting great value for practical use. In particular, we show that the objective function in our optimization model is *multimodular* in the decision variables when no-show probabilities are homogeneous and time-independent; this elegant property guarantees that a *local search yields a global optimum*. When no-show probabilities become heterogeneous and time-dependent, we propose an innovative variable transformation to reformulate the original challenging two-stage *nonlinear* optimization model into a stochastic *linear* programming model with simple structures, which can be directly solved by off-the-shelf optimization packages. In addition, this reformulation can leverage the multimodularity of the objective function, if this property holds, to further accelerate the solution process. To our knowledge, we are the first to propose such a reformulation, which may have broader applications in other contexts of optimization.

In addition to the above, our empirical investigation of walk-in patients also contributes to the relatively scant empirical literature on customer arrivals by revealing new temporal patterns and models for customer demand. Via extensive numerical experiments, we demonstrate that, with walk-ins, the optimal schedules are fundamentally different from those identified in the previous literature, which does not consider walk-ins. Our numerical study also reveals that walk-ins introduce a new source of uncertainty to the system and cannot be viewed as a solution to compensate for patient no-shows. Scheduling, however, is an effective way to counter some of the negative impact from uncertain patient behaviors. We show that adopting the schedules suggested by our models, which explicitly take walk-ins into account, can lead to a significant efficiency improvement in practice. Full benefit of our model can be realized with sufficient demand; even when demand is insufficient, our model can still be applied in an online fashion and deliver excellent performances.

The remainder of this paper is organized as follows. Section 2 briefly reviews the relevant literature.

Section 3 presents an exploratory analysis of walk-in patterns using our data from practice. Section 4 develops and analyzes our basic scheduling model with random walk-ins. In Section 5, we incorporate patient no-show behaviors into the basic model and present our reformulation approach. We discuss our numerical study and managerial insights in Section 6. Section 7 provides concluding remarks. All proofs of the analytical results can be found in the online appendix.

2. Literature Review

Our work is closely related to the (outpatient) appointment-scheduling literature that investigates how to schedule patients over time in a day. Extensive work has been focused on developing mathematical programming models to optimize the tradeoff between patient in-clinic waiting and provider utilization. Two types of decision variables have been considered. The first type of decision scenario is concerned with the exact appointment time for each patient (decision variables are continuous); see, for example, Denton and Gupta (2003), Hassin and Mendel (2008), Kong et al. (2013), Chen and Robinson (2014), and Jiang et al. (2017). The second type of decision scenario, like ours, divides a day into a certain number of appointment slots and determines the number of patients scheduled to each slot (decision variables are integers); see, for example, Kaandorp and Koole (2007), Robinson and Chen (2010), LaGanga and Lawrence (2012), Zacharias and Pinedo (2014), and Zacharias and Pinedo (2017). These previous studies have considered a variety of uncertainties in practice that may affect the design of appointment templates (such as patient no-shows and random service times). However, none of the above work explicitly considers walk-ins, an important phenomenon in healthcare, as discussed earlier. Besides the analytical work above, simulation-based models have been used to study appointment-scheduling decisions, and some consider the impact of walk-ins; see, for example, Cayirli and Gunes (2014). Our work complements and advances this prior literature by proposing new analytical models and solution approaches to optimize the appointment schedule in anticipation for random walk-ins.

Next, we draw close attention to a few articles that are most related to our work. All these studies, including ours, treat the number of patients scheduled to each appointment slot as the decision variable. In Kaandorp and Koole (2007), patients' no-show probabilities are homogeneous, and provider service times are exponentially distributed. They develop a local search procedure for the optimal schedule. Different from Kaandorp and Koole (2007), Robinson and Chen (2010) and Zacharias and Pinedo (2014) both assume deterministic service times for providers. Under this assumption, Robinson and Chen (2010) identify an important property—the "No Hole" property—for the optimal schedule (more on this below). Zacharias and Pinedo (2014) consider both offline and online scheduling, and develop structural properties and effective heuristics for the optimal schedule. Zacharias and Pinedo (2017) extend their earlier work to a multiserver setting.

Our work departs from the four studies above in several important ways. First, we explicitly take into account potential walk-ins during the day, and we allow the walk-in process to be general. We demonstrate that the resulting optimization problem is much more complicated, and those elegant properties that hold without walk-ins (e.g., the No Hole property) do not hold anymore when walk-ins are accepted. Second, the previous literature solves for the optimal schedule, either using local search or via enumeration (after characterizing the structural results). We are, however, able to provide the first two-stage stochastic linear programming formulation for the appointment-scheduling problem with both walk-ins and no-shows present. This formulation not only is amenable to many standard mixed-integer programming solvers, but also has a special structure, which allows us to develop a unified solution approach proven to be highly effective in numerical experiments. Third, our modeling framework and solution approaches are very flexible and can also accommodate heterogeneous, time-dependent patient no-show behaviors and random service times (of a certain distribution).

An important recent work by Zacharias and Yunes (2019) is studying a similar problem as ours. They aim to investigate the multimodularity of the objective function in a general setting and design a fast local search procedure, whereas we focus on reformulating a challenging problem to a tractable mathematical program. These two studies complement each other by investigating a similar, challenging problem from fundamentally different angles.

Our work is also connected to, but differs significantly from, the literature in service operations management that deals with walk-in customers. For instance, Bertsimas and Shioda (2003) develop methods to dynamically decide when, if at all, to seat an incoming party during the day of operations of a restaurant. This is an online decision problem in the restaurant industry, whereas our work focuses on an offline decision to determine the best schedule in a doctor's office. Alexandrov and Lariviere (2012) develop a game-theoretic model to study whether reservations are recommended for restaurants where walk-in customers are often allowed. Bitran and Gilbert (1996) and Gans and Savin (2007) study the reservation management problem with uncertain walk-in customers for hotels and rental firms, respectively. The last three studies focus on capacity-level decisions (e.g., how much capacity to reserve for walk-ins), rather than within-day operations investigated by us.

3. Exploratory Study of Walk-in Arrival Patterns

Although previous literature has a rich documentation on the volume of walk-ins [see, e.g., Moore et al. (2001) and Cayirli et al. (2008)], relatively little is known about the temporal pattern of walk-in arrivals. In this section, we use a data set obtained from a large community health center located in NYC to conduct an exploratory study on the temporal pattern of walk-ins. This simple study is based on data from a single organization and is by no means comprehensive; its main purpose is to motivate our analytical appointment-scheduling model that follows.

3.1. Data

The data were extracted from the electronic medical-record system of our collaborating health center. This center provides comprehensive medical and dental care to the local community and serves more than 25,000 patient visits every year. The data set spans 3 years ranging from January 2011 to January 2014, and contains 67,847 valid records of patient visits. In these records, more than 15% (10,402) are walk-ins. There are 38 providers (including physicians and nurse practitioners) in the data set; some providers have more than 50% of the patients they see as walk-ins. In this center, walk-ins are accepted throughout the office hours. When analyzing these data, we focus on three specialties, Nurse Practitioner, Internal Medicine, and Pediatrics, which serve more than 80% of the walk-in visits (with 5,076, 2,669, and 1,128 records, respectively). Then, we choose six providers who have the most walk-in records (four Nurse Practitioners, one Internist, and one Pediatrician) for analysis.

3.2. Statistical Analysis Framework

To study the arrival patterns of walk-ins, we adopt a Poisson regression framework to model the number of walk-ins in each hour. Specifically, for each of the six providers, we estimate and compare three regression models below (from simple to more comprehensive). In these models, walk-in arrivals in different time slots are assumed independent. This assumption is supported by our empirical observation that, for each of the six providers we study, the correlations of walk-in counts in different time slots are very weak (with fairly small correlation coefficients) and in most cases not statistically significant (see Section A in the online appendix for detailed data presentation).

Model 1 is a classic Poisson regression model, where Y_t, the number of walk-ins in hour t, has a Poisson distribution with mean λ_t, which depends on the hour t. That is,

$$Pr(Y_t = k) = \frac{\lambda_t^k e^{-\lambda_t}}{k!}, \quad k = 0, 1, 2, \ldots. \tag{1}$$

Using the logarithm as the canonical link function, Model 1 is specified as follows:

$$\log(\lambda_t) = \gamma_1 + \gamma_2 x_2 + \cdots + \gamma_T x_T, \qquad \text{(Model 1)}$$

where x_i is a dummy variable, which takes value 1 if $t = i$ and value 0 otherwise, $i = 2, 3, \ldots, T$ (note that we do not have x_1 in Model 1, because hour 1 is the base category whose effect is captured by γ_1). In Model 1, γ_t's are the statistical parameters we will estimate, and the hourly arrival rates can then be estimated as $\lambda_1 = e^{\gamma_1}$ and $\lambda_t = e^{\gamma_1+\gamma_t}$ for $t > 1$.

A close look at our data reveals that for some of the providers, there are an excessive number of zeros in hourly arrivals, which may make Model 1 not a good fit. To address this problem of excess zeros, we consider zero-inflated Poisson regression models (Lambert 1992), which first determine whether there are zero events or any events, and then use a Poisson distribution to determine the number of events, if there are any. That is, the number of walk-ins in hour t is modeled as follows:

$$Pr(Y_t = k) = \begin{cases} a + (1-a)e^{-\lambda_t} & \text{if } k = 0, \\ (1-a)\dfrac{\lambda_t^k e^{-\lambda_t}}{k!} & \text{if } k > 0, \end{cases} \tag{2}$$

where a is the zero-event probability and λ_t is the hourly arrival rate. Using the canonical link functions, the statistical specification of the above model can be written as follows:

$$\log\left(\frac{a}{1-a}\right) = b \ \text{ and } \ \log(\lambda_t) = \gamma_1 + \gamma_2 x_2 + \cdots + \gamma_T x_T, \qquad \text{(Model 2)}$$

where x_i is defined as in Model 1. Under Model 2, $a = 1 - 1/(e^b + 1)$, $\lambda_1 = e^{\gamma_1}$ and $\lambda_t = e^{\gamma_1+\gamma_t}$ for $t > 1$.

Model 2 assumes a constant zero probability a. A more comprehensive model, however, is to specify that the zero probability also depends on time t. That is, the number of walk-ins in hour t is modeled as below:

$$Pr(Y_t = k) = \begin{cases} a_t + (1-a_t)e^{-\lambda_t} & \text{if } k = 0, \\ (1-a_t)\dfrac{\lambda_t^k e^{-\lambda_t}}{k!} & \text{if } k > 0, \end{cases} \tag{3}$$

where a_t is the zero-event probability in hour t. Its corresponding statistical specification is

$$\log\left(\frac{a_t}{1-a_t}\right) = b_1 + b_2 x_2 + \cdots + b_T x_T \ \text{ and}$$
$$\log(\lambda_t) = \gamma_1 + \gamma_2 x_2 + \cdots + \gamma_T x_T, \qquad \text{(Model 3)}$$

where x_i is defined as in Model 1. Under Model 3, $a_1 = 1 - 1/(e^{b_1} + 1)$ and $a_t = 1 - 1/(e^{b_1+b_t} + 1)$ for $t > 1$; $\lambda_1 = e^{\gamma_1}$ and $\lambda_t = e^{\gamma_1+\gamma_t}$ for $t > 1$.

These three models increase in their generality. To assemble the data for analysis, for each provider, we count the number of patients who arrive between 30 minutes before an hour and 30 minutes after as those arriving for that hour. To arrive at the most parsimonious model that adequately describes the data, we conduct a series of statistical tests. Note that Model 2 is a reduced model of Model 3 (by specifying that $a_t = a$); we can use the likelihood-ratio test to test whether Model 3 makes a significant improvement over Model 2. Model 1 and Model 2, however, are not nested. So we use Vuong's closeness test to test whether Model 2 improves upon Model 1 significantly (Vuong 1989). For each provider, we adopt the simplest model, to which more complicated models cannot make a significant improvement, as our final model. We test the goodness-of-fit of the final model using the chi-square test.

3.3. Empirical Results

Table 1 summarizes the testing results of three fitted models for each provider. Providers' initials are used to protect their confidentiality. It is important and interesting to note that all three models have appeared as the final model for some provider. Specifically, for providers GED, KNI, and WAT, we find that the number of walk-ins in each slot follows the zero-inflated Poisson distribution rather than the classic Poisson distribution. For providers GED and KNI, the estimated zero event probabilities a are constant over time and they are both 0.14 in the final model. For provider WAT, the zero event probability depends on hour of day, and its estimated value is 0.99, 0.24, 0.42, 0.50, 0.44, 0.59, 0.39, 0.63, 0.33, 0.80, and 0.33 from 8 a.m. to 6 p.m., respectively. For providers ALD, GAR, and LOK, we find that the Poisson distribution is appropriate to model the number of walk-ins arriving in each hour, although its mean varies over time.

Table 1 makes an important implication that there is no one-size-fits-all model for walk-in processes. There is relatively scant literature that examines the arrival pattern of walk-ins using empirical data. The extant limited literature almost unanimously suggests that the unscheduled walk-in process follows a (nonhomogeneous) Poisson process; see, for example, Kim and Whitt (2014). Our exploratory study contributes to this literature by revealing new arrival patterns of walk-ins (i.e., zero-inflated Poisson process). Although the previous appointment-scheduling work that considers walk-ins (all are simulation-based, to the best of our knowledge) has predominantly used Poisson process to model walk-in arrivals (e.g., Cayirli et al. 2008), we suggest that appointment-scheduling models should be able to accommodate general arrival patterns of walk-ins. We develop one such optimization model in the following sections.

Figure 1 shows the expected number of hourly walk-ins for each provider. All providers except for GAR take a lunch break around 1 p.m. (but still may have a few walk-ins at that time), and thus we see a bimodal distribution of the arrival rates. In contrast, GAR does not take a lunch break and goes home earlier; the walk-in rate to this provider shows a unimodal pattern over the day. This observation suggests the potential endogeneity of walk-ins on the provider's work schedule. That is, if patients know that the provider has a lunch break (and does not serve patients during that time), patients will not come. On a strategic level, Alexandrov and Lariviere (2012) study such an issue. Specifically, they consider how a firm (restaurant) should make a reservation decision when customer walk-in behavior is influenced by such a decision. In contrast, our mathematical formulation below assumes that the provider's work schedule has already been fixed and announced to patients, and thus the walk-in distribution is *exogenously* determined. If the provider changes his or her work schedule, our model can be rerun based on the newly observed patient walk-in pattern after it is stabilized, to generate the optimal schedule under the new work schedule of the provider. It is, however, very interesting to study how to set a provider's work schedule taking into account the endogeneity of walk-ins, and we leave this topic for future research.

4. Basic Model

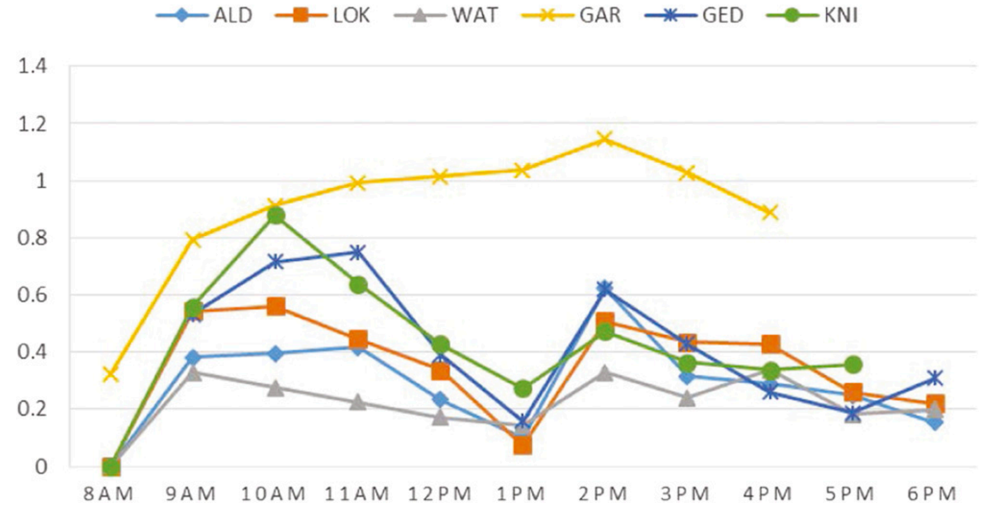
In this section, we develop a basic appointment-scheduling model with random walk-in patients. For now, we

Table 1. Summary of Statistical Analysis

Provider	Specialty	Sample size	Model 2 vs. 3	Model 1 vs. 2	Final model	Goodness of fit
ALD	Nurse practitioner	1,524	0.98	0.49 (0.63)	Model 1	0.74
GAR	Nurse practitioner	3,403	0.99	−0.00 (0.99)	Model 1	0.86
GED	Internal medicine	1,115	0.97	1.00 (0.32)	Model 2	0.18
KNI	Nurse practitioner	1,729	0.42	1.48 (0.14)	Model 2	0.70
LOK	Nurse practitioner	1,308	0.69	0.37 (0.71)	Model 1	0.27
WAT	Pediatrics	3,045	0.08	3.09 (0.00)	Model 3	0.43

Notes. (1) Providers ALD, LOK, and WAT work from 8 a.m. to 6 p.m.; GAR works from 8 a.m. to 4 p.m.; GED works from 9 a.m. to 6 p.m.; and KNI works from 8 a.m. to 5 p.m. (2) Column "Model 2 vs. 3" shows the p-value of the likelihood ratio test. If $p < 0.1$, Model 3 makes a significant improvement over Model 2. (3) Column "Model 1 vs. 2" shows the Vuong-test statistic value (p-value in parentheses). A positive Vuong-test statistic suggests that Model 2 is closer to the true model. Unless $p < = 0.5$, we still choose Model 1 as the final model. (4) Column "Goodness of fit" shows the p-value of the goodness-of-fit test ($p > 0.1$ indicates a good fit).

Figure 1. (Color online) Expected Number of Hourly Walk-ins for Each Provider



assume that all scheduled patients will show up at their appointment times. We will extend our modeling framework to incorporate patient no-show behavior in Section 5. Throughout, we will use lowercase (uppercase) Greek letters to denote random variables (calculated values), lowercase (uppercase) letters to denote variables (constants), and bold-faced lowercase (uppercase) letters to denote vectors (matrices). The dimensions of vectors or matrices should be evident from the context. We provide a summary of the notations in Table B.7 of Online Appendix B.

Consider a generic clinic session for a single provider. In practice, the length of a clinic session is often measured by the number of appointment slots, and patients are scheduled to arrive at the beginning of these slots (patients are rarely scheduled to arrive in the middle of a slot). Following this convention, we consider a clinic session with T appointment slots, where T is a prespecified number. The provider needs to schedule n patients in these slots, and n is a decision variable.

Besides these scheduled patients, a random number of patients may walk in for services. For tractability, we assume that walk-in patients always arrive at the beginning of each appointment slot.[1] Let $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_T)$ be a random vector with support on nonnegative integers, where β_t represents the number of walk-ins arriving at the beginning of slot t. That is, the arrival pattern of walk-ins may depend on time t. For now, we assume that β_t's are independent of each other. (In the next section, we will consider more general walk-in processes, e.g., those with correlations of walk-in counts at different times.) We note that $\boldsymbol{\beta}$ as a whole is exogenously given and is independent of other aspects of the model (see more discussions in Section 3.3).

We assume that the service time of each patient is exactly one appointment slot (normalized as one unit of time in our model). In practice, especially in primary care, the provider usually can control her consultation time with patients to be within the allotted time by adjusting the conversation content and speed (Gupta and Denton 2008). Indeed, deterministic service time is a reasonable assumption commonly made in the appointment-scheduling literature; see, for example, Robinson and Chen (2010), LaGanga and Lawrence (2012), and Zacharias and Pinedo (2014, 2017). Nevertheless, we note that our models and solution approaches can be easily extended to incorporate random service times with certain probability distributions (see Online Appendix F).

In the setting above, we need to determine n, the total number of patients to be scheduled, and also the number of patients scheduled in each slot. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_T)$ be our decision vector,[2] in which x_t is the number of patients scheduled at slot t. It is evident that $n = \sum_{t=1}^{T} x_t$. Following the previous literature—for example, Robinson and Chen (2010) and Zacharias and Pinedo (2014)—we assume that all scheduled patients are punctual for tractability.

A common optimization framework in the literature is to assign different cost rates to patient wait time, provider idle time, and overtime, and then to minimize the expected total weighted cost with these cost rates serving as the weights. We follow this framework, but note that this cost structure can be slightly simplified in our model without loss of generality. To see that, let C_S and C_W be the waiting cost for a scheduled patient and a walk-in patient per appointment slot of time, respectively. Let C_I and C_O be the provider's idling cost and overtime cost per appointment slot of time, respectively. For a given schedule, let Γ_S and Γ_W be the expected total wait time of scheduled patients and that of walk-in patients, respectively. Let Γ_I and Γ_O be the expected idle time and overtime of the provider. Thus, the expected total weighted cost is

$$C_S\Gamma_S + C_W\Gamma_W + C_I\Gamma_I + C_O\Gamma_O. \tag{4}$$

Let Γ_D be the expected duration of the whole clinic session—that is, the time from the beginning of the session to T or the time when the last patient leaves, whichever is later. It is clear that $\Gamma_O = \Gamma_D - T$. Let N_W be the expected number of walk-in patients—that is, $N_W = \mathbb{E}(\sum_{t=1}^{T}\beta_t)$. Then, $\sum_{t=1}^{T} x_t + N_W$ is the expected total consultation time that the provider spends with patients, and thus the difference between Γ_D and $\sum_{t=1}^{T} x_t + N_W$ is the expected idle time of the provider—that is, $\Gamma_I = \Gamma_D - \sum_{t=1}^{T} x_t - N_W$. We can rewrite the expected total weighted cost (4) as

$$C_S\Gamma_S + C_W\Gamma_W + C_I(\Gamma_D - \sum_{t=1}^{T} x_t - N_W) + C_O(\Gamma_D - T). \tag{5}$$

As N_W and T are constants, they can be omitted from the optimization process. Let $C_D = C_I + C_O$, and normalize C_S to be 1. The expected total weighted cost in our optimization objective can be simplified as follows:

$$\Gamma_S + C_W\Gamma_W + C_D\Gamma_D - C_I\sum_{t=1}^{T} x_t. \tag{6}$$

When deriving the optimal solution, we use (6) for simplicity; we will use (5) when the actual objective value is needed, such as calculating the expected total cost associated with a schedule.

To calculate Γ_S, Γ_W, and Γ_D, we first evaluate $\Pi_t(k)$, the probability of k patients waiting for services at the end of t. Let $p_t(b)$ be the probability of b walk-ins arriving at slot t—that is, $p_t(b) = Pr(\beta_t = b)$—and let $\overline{N_t}$ be a sufficiently large number so that it only suffices to consider at most $\overline{N_t}$ patients in the system at time t ($\overline{N_t}$ can be determined by truncating from above the distribution of walk-ins in slot t). Given a schedule $\boldsymbol{x}$, we can write $\Pi_t(k)$ recursively as

$$\Pi_t(k) = \sum_{j=0}^{k-x_t+1} \Pi_{t-1}(j)p_t(k - x_t - j + 1) + \begin{cases} \Pi_{t-1}(0)p_t(0) & \text{if } k = 0 \text{ and } x_t = 0, \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

for $k = 0,..,\overline{N_t}$ and $t = 1, \ldots, T$ with $\Pi_0(0) = 1$. The first term in the right-hand side (RHS) of Equation (7) calculates the joint probability that j patients wait at the end of $t - 1$, $k - x_t - j + 1$ walk-ins arrive at t, and one patient gets served at t. For $k = 0$, there is one more term, which is the joint probability that the system is empty at the end of $t - 1$, no walk-ins arrive at t, and no patients are served at t (this term is valid only if $x_t = 0$—that is, no scheduled patients arrive at t). For Γ_D, we have

$$\Gamma_D = T + \sum_{k=1}^{\overline{N_T}} k\Pi_T(k), \tag{8}$$

in which the second term is the expected number of patients waiting at the end of T.

Before analyzing patient wait time, we need to specify the priority order between scheduled patients and walk-ins. Although some walk-ins arrive due to acute care needs, their health conditions are, in general, stable. If indeed walk-ins have emergency issues that require immediate attention, they are often diverted to emergency rooms following the standard clinical protocol. Therefore, common practice usually gives walk-ins lower priority compared with scheduled patients (Berman 2016). The underlying cost structure adopted by practitioners implies that the waiting cost of walk-ins is no larger than that of scheduled patients. We follow this rationale and assume that $C_W \leq C_S = 1$ throughout the paper. Based on the $c\mu$ rule, we know that it is optimal to serve scheduled patients, if any, before walk-ins. [In contrast, interested readers may refer to Koeleman and Koole (2012) for a different model setting in which emergency patients arrive randomly (as walk-ins in our model) but need to be served before scheduled patients.]

Next, we evaluate patient wait time, starting with the scheduled patients who have priority. Let s_t be the number of scheduled patients waiting at the end of slot t. We can write s_t recursively as $s_t = (s_{t-1} + x_t - 1)^+$ for $t = 2, \ldots, T$ with $s_1 = (x_1 - 1)^+$. It follows that the wait time of scheduled patients $\Gamma_S(\boldsymbol{x})$ can be calculated as,

$$\Gamma_S(\boldsymbol{x}) = \sum_{t=1}^{T} s_t + \sum_{j=1}^{s_T-1} j. \tag{9}$$

Let $\Gamma_T(\boldsymbol{x})$ be the expected total wait time of *all* patients given a schedule $\boldsymbol{x}$. Using (7), we have

$$\Gamma_T(\boldsymbol{x}) = \sum_{t=1}^{T}\sum_{k=1}^{\overline{N_t}} k\Pi_t(k) + \sum_{k=1}^{\overline{N_T}} \Big(\sum_{j=1}^{k-1} j\Big)\Pi_T(k).$$

Noting that Γ_W is the difference between Γ_T and Γ_S, we obtain

$$\Gamma_W(\boldsymbol{x}) = \Gamma_T(\boldsymbol{x}) - \Gamma_S(\boldsymbol{x}). \tag{10}$$

Finally, our optimization problem in this section can be represented as,

$$\min_{\boldsymbol{x}\in\mathbb{Z}_+^T} \Gamma_S(\boldsymbol{x}) + C_W\Gamma_W(\boldsymbol{x}) + C_D\Gamma_D(\boldsymbol{x}) - C_I\sum_{t=1}^{T} x_t \tag{P1}$$

$\Gamma_S(\boldsymbol{x}), \Gamma_W(\boldsymbol{x}), \Gamma_D(\boldsymbol{x})$ are defined in(8), (9), (10), respectively, where $\mathbb{Z}_+^T$ represents the set of all T-dimensional nonnegative integer vectors.[3]

4.1. Multimodularity of the Objective Function

The problem (P1) is a combinatorial optimization problem that is difficult to solve. In the following sections, we explore the properties of (P1) and develop efficient solution algorithms. To facilitate our discussion, we first introduce the concept *multimodularity*.

Definition 1 (Hajek 1985). Define the vectors in $\mathbb{Z}^T$ by

$$\left\{\begin{array}{c} v_0 \\ v_1 \\ v_2 \\ \vdots \\ v_{T-1} \\ v_T \end{array}\right\} = \left\{\begin{array}{c} (-1,0,0,\ldots,0,0) \\ (1,-1,0,\ldots,0,0) \\ (0,1,-1,\ldots,0,0) \\ \vdots \\ (0,0,0,\ldots,1,-1) \\ (0,0,0,\ldots,0,1) \end{array}\right\}, \tag{11}$$

and let $\boldsymbol{V}^\diamond = \{v_0, v_1, \ldots, v_T\}$. We say that a function g on $\mathbb{Z}_+^T$ is *multimodular* if for all $\boldsymbol{x}$ in $\mathbb{Z}_+^T$,

$$g(\boldsymbol{x}+v_i) - g(\boldsymbol{x}) \geq g(\boldsymbol{x}+v_j+v_i) - g(\boldsymbol{x}+v_j),$$

whenever $v_i, v_j \in \boldsymbol{V}^\diamond$, $\boldsymbol{x}+v_i \in \mathbb{Z}_+^T$, $\boldsymbol{x}+v_j \in \mathbb{Z}_+^T$ and $v_i \neq v_j$.

Multimodularity can be interpreted as follows: The marginal difference in the function value from perturbing a solution $\boldsymbol{x}$ by v_i is greater than or equal to that from perturbing a solution $\boldsymbol{x}+v_j$ by v_i. One perhaps most useful property of a multimodular function is stated in the lemma below.

Lemma 1 (Murota 2005). *If a function $g(\boldsymbol{x})$ is multimodular, then a local minimum on its domain is a global minimum.*

Prior literature has shown that the objective function in appointment-scheduling problems can be multimodular in certain settings (Kaandorp and Koole 2007, Zacharias and Pinedo 2017); see Section 2 for a detailed discussion on this literature. We extend this literature by showing that this elegant property of the objective function still holds with exogenous, random walk-ins.

Proposition 1. *Define $f(\boldsymbol{x})$: $\mathbb{Z}_+^T \to \mathbb{R}$, the objective function of* (P1), *by $f(\boldsymbol{x}) := \Gamma_S(\boldsymbol{x}) + C_W\Gamma_W(\boldsymbol{x}) + C_D\Gamma_D(\boldsymbol{x}) - C_I\sum_{t=1}^{T} x_t$. Then, $f(\boldsymbol{x})$: $\mathbb{Z}_+^T \to \mathbb{R}$ is multimodular on its domain $\mathbb{Z}_+^T$.*

Proposition 1 and Lemma 1 suggest that a local optimal solution of (P1) is also globally optimal. To illustrate, we first define the neighbor of a solution in our modeling context.

Definition 2 (Neighbor of $\boldsymbol{x}$). We say $\boldsymbol{x}'$ is a feasible neighbor of $\boldsymbol{x}$ if $\boldsymbol{x} \in \mathbb{Z}_+^T$ and $\boldsymbol{x}' = \boldsymbol{x} + \sum_{v\in \boldsymbol{V}} v$ for some $\boldsymbol{V} \subsetneq \boldsymbol{V}^\diamond$ and $\boldsymbol{V} \neq \emptyset$, where $\boldsymbol{V}^\diamond$ is defined in (11).

Note that $\boldsymbol{V}$ is a nonempty strict subset of $\boldsymbol{V}^\diamond$. For instance, $\boldsymbol{x}+v_1$ is a neighbor of $\boldsymbol{x}$; the former moves a patient from slot 2 to slot 1, while keeping the positions of other patients unchanged. Then, we obtain the following criteria to determine whether a solution is optimal or not, based on a result in Altman et al. (2000).

Corollary 1. *If $\boldsymbol{x} \in \mathbb{Z}_+^T$ and $f(\boldsymbol{x}) \leq f(\boldsymbol{x}')$ for any feasible neighbor $\boldsymbol{x}'$ of $\boldsymbol{x}$, then $\boldsymbol{x}$ is a global optimal solution for* (P1).

Corollary 1 guarantees that we can arrive at the optimal schedule via a local search (i.e., starting from any feasible solution, moving to a feasible neighbor solution, if any, that improves the current solution, and continuing in this fashion until no better solutions can be found in the neighborhood). However, a feasible solution can have at most $2^{(T+1)} - 1$ neighbors, which may make local search ineffective in solving large-scale problems. We next explore the structural properties of the optimal schedule to gain additional insights and to further simplify the solution process.

4.2. Structural Properties of the Optimal Schedule

When a manager can choose the total number of patients to schedule and if all scheduled patients show up, then there seem to be no incentives for the manager to overbook—that is, schedule multiple patients into a single appointment slot. Consider a schedule that does overbook; then, one can improve it by avoiding overbooking in one of the two following ways. If there are empty slots after the overbooked slot, then moving the overbooked patient to the next closest empty slot only decreases the total waiting cost of scheduled patients and does not affect other costs. If, however, all slots after the overbooked slot are booked, then removing additional patients from the overbooked slot altogether reduces the total waiting cost of both scheduled and walk-in patients, as well as the overtime cost. Following this rationale, we have the following structural result for the optimal schedule.

Proposition 2. *For* (P1), *there exists an optimal schedule that does not overbook—that is, $x_t \leq 1$ for all $t = 1, 2, \ldots, T$.*

Proposition 2 indicates that in order to find an optimal schedule, we only need to examine at most 2^T possible ones that do not overbook. Recall that Corollary 1 suggests only local search in the neighborhood of the current solution is needed. Thus, in order to find an optimal schedule, one only needs to consider those non-overbooking schedules in local search. Leveraging *both* the structural properties of the objective function *and* those of the optimal schedule can drastically reduce the search space for the optimal schedule.

5. Model Incorporating No-Show Behavior

In this section, we discuss the model to optimize the appointment schedule when both random walk-in and customer no-show behaviors are present. To be consistent with our earlier developments, let $\boldsymbol{x}$ be the schedule, our decision vector, and $\boldsymbol{\beta}$ represent the vector of random walk-ins. Let $\boldsymbol{\alpha}(\boldsymbol{x}) = (\alpha_1(x_1),$

$\alpha_2(x_2),\ldots,\alpha_T(x_T))$ denote the number of show-up patients among those scheduled. That is, $\alpha_t(x_t)$ is the number of show-ups at t given that x_t patients are scheduled at t. We assume that each scheduled patient independently shows up (or not). We start by considering the homogeneous case. Specifically, let q^s be the show-up probability for all scheduled patients, then $\alpha_t(x_t)$ follows the binomial distribution with its probability mass function described as follows:

$$\mathbf{Pr}(x_t = k) = q_t(k, x_t) = \binom{x_t}{k}(q^s)^k(1-q^s)^{(x_t-k)}, \qquad k = 0, 1, \ldots, x_t.$$

Later in Section 5.4, we will relax this assumption and discuss how to handle heterogeneous and time-dependent patient no-shows.

Same as in Section 4, the objective of our optimization model here is to minimize the expected total weighted cost—that is, $\mathit{\Gamma}_S + C_W\mathit{\Gamma}_W + C_D\mathit{\Gamma}_D - C_I\mathbb{E}\big[\sum_{t=1}^{T}\alpha_t(x_t)\big]$. For convenience, we still use $\Pi_t(k)$ to denote the probability of k patients waiting at the end of t. Given $\boldsymbol{x}$, we have

$$\Pi_t(k) = \sum_{i=0}^{x_t}\sum_{j=0}^{k-i+1}\Pi_{t-1}(j)q_t(i,x_t)p_t(k-i-j+1) + \begin{cases}\Pi_{t-1}(0)q_t(0,x_t)p_t(0) & \text{if } k = 0,\\ 0 & \text{if } k > 0,\end{cases} \tag{12}$$

for $k = 0,..,\overline{N_t}$ and $t = 1,\ldots,T$ with $\Pi_0(0) = 1$. The first term on the RHS of (12) is the probability of j patients waiting at the end of $t-1$, i out of x_t scheduled patients showing up at t, and $k-i-j+1$ patients walking in at t. For $k = 0$, there is one more term, which is the probability that the system is empty at $t-1$ and no scheduled patients or walk-ins arrive at t.

Similar to our earlier derivation of (8), the expected duration $\mathit{\Gamma}_D$ here is T plus the expected number of patients at $T+1$. Thus, we have

$$\mathit{\Gamma}_D(\boldsymbol{x}) = T + \sum_{k=1}^{\overline{N_T}} k\Pi_T(k). \tag{13}$$

Recall that scheduled patients are given priority over walk-ins.[4] Let $\Psi_t(k)$ be the probability of k scheduled patients waiting at the end of t. Then, we can write $\Psi_t(k)$, $t = 1, 2, \ldots, T$, recursively as

$$\Psi_t(k) = \sum_{j=0}^{k+1}\Psi_{t-1}(j)q_t(k-j+1,x_t) + \begin{cases}\Psi_{t-1}(0)q_t(0,x_t) & \text{if } k = 0,\\ 0 & \text{if } k = 1,2,\ldots,n,\end{cases} \tag{14}$$

where $n = \sum_{t=1}^{T} x_t$ and $\Psi_0(0) = 1$. The first term of the RHS is the probability of j scheduled patients waiting at the end of $t-1$, one of them served, and $k-j+1$ scheduled patients showing up at t. For $k = 0$, there is one more term, which is the probability that the system is empty at $t-1$ and no scheduled patients show up at t. It follows that the expected total wait time for *scheduled* patients, $\mathit{\Gamma}_S$, can be calculated by summing up the expected number of scheduled patients waiting at the end of each appointment slot. More precisely, we have

$$\mathit{\Gamma}_S(\boldsymbol{x}) = \sum_{t=1}^{T}\sum_{k=1}^{\overline{N_t}} k\Psi_t(k) + \sum_{k=1}^{\overline{N_T}}\Big(\sum_{j=1}^{k-1} j\Big)\Psi_T(k). \tag{15}$$

Similarly, the expected total wait time of all patients $\mathit{\Gamma}_T$ can be calculated by summing up the expected number of all patients waiting at each slot—that is,

$$\mathit{\Gamma}_T(\boldsymbol{x}) = \sum_{t=1}^{T}\sum_{k=1}^{\overline{N_t}} k\Pi_t(k) + \sum_{k=1}^{\overline{N_T}}\Big(\sum_{j=1}^{k-1} j\Big)\Pi_T(k).$$

And, the expected wait time of walk-ins $\mathit{\Gamma}_W$ is the difference between $\mathit{\Gamma}_T$ and $\mathit{\Gamma}_S$—that is,

$$\mathit{\Gamma}_W(\boldsymbol{x}) = \mathit{\Gamma}_T(\boldsymbol{x}) - \mathit{\Gamma}_S(\boldsymbol{x}). \tag{16}$$

Finally, the optimization model when both random walk-in and patient no-show behaviors are present can be formulated as follows,

$$\min_{\boldsymbol{x}\in\mathbb{Z}_+^T} \mathit{\Gamma}_S(\boldsymbol{x}) + C_W\mathit{\Gamma}_W(\boldsymbol{x}) + C_D\mathit{\Gamma}_D(\boldsymbol{x}) - C_I\mathbb{E}\Big[\sum_{t=1}^{T}\alpha_t(\boldsymbol{x}_t)\Big] \tag{P2}$$

$$\mathit{\Gamma}_S(\boldsymbol{x}), \mathit{\Gamma}_W(\boldsymbol{x}), \mathit{\Gamma}_D(\boldsymbol{x}) \text{ are defined in (13), (15), and (16), respectively.}$$

We note that the objective function of (P2) remains multimodular, with both patient no-shows and exogenous random walk-ins considered. This result extends Proposition 1 and is formalized below.

Proposition 3. *Define $h(\boldsymbol{x})$: $\mathbb{Z}_+^T \to \mathbb{R}$, the objective function of (P2), by $h(\boldsymbol{x}) := \mathit{\Gamma}_S(\boldsymbol{x}) + C_W\mathit{\Gamma}_W(\boldsymbol{x}) + C_D\mathit{\Gamma}_D(\boldsymbol{x}) - C_I\mathbb{E}\big[\sum_{t=1}^{T}\alpha_t(\boldsymbol{x}_t)\big]$. Then, $h(\boldsymbol{x})$: $\mathbb{Z}_+^T \to \mathbb{R}$ is multimodular on its domain $\mathbb{Z}_+^T$.*

Note that Proposition 3 still holds under a general walk-in process (e.g., when β_t's, the number of walk-ins in different time slots, are correlated), because the proof does *not* require the independence of walk-in counts at different times. Following Proposition 3, we have an equivalent result of Corollary 1 below.

Corollary 2. *If $\boldsymbol{x} \in \mathbb{Z}_+^T$ and $h(\boldsymbol{x}) \le h(\boldsymbol{x}')$ for any feasible neighbor $\boldsymbol{x}'$ of $\boldsymbol{x}$ (in the sense of Definition 2), then $\boldsymbol{x}$ is an optimal solution for (P2).*

Although a local search can lead to an optimal schedule for (P2), we observe that (P2) is much harder than (P1). First of all, when patient no-shows present, it is more difficult and takes much more time to evaluate $\Gamma_S(\boldsymbol{x})$, $\Gamma_W(\boldsymbol{x})$ and $\Gamma_D(\boldsymbol{x})$ for a given $\boldsymbol{x}$. More importantly, there seems to be no clear structures for the optimal schedule when both walk-ins and no-shows are present. In an optimal schedule, some slots may be overbooked (this is different from Proposition 2), and some may be purposefully left open (this is different from the No Hole property identified in Robinson and Chen 2010). In short, overbooking and "holes" may coexist in an optimal schedule, without a straightforward pattern; see Section C.5 in the online appendix for some concrete examples of complex optimal schedules with walk-ins. Such a lack of clear structures for the optimal schedule prohibits one from ruling out nonoptimal schedules easily by checking the pattern of the schedule.

5.1. Two-Stage Programming Model

To solve (P2) more efficiently, we propose a two-stage programming approach. As demonstrated later, this two-stage optimization model is quite novel—it provides a much quicker way to evaluate the objective function, and in addition, the multimodularity result obtained in Proposition 3 earlier, if it holds, can be used to *guide* the solution search in this two-stage programming model.

To facilitate our discussion, we use a sample path representation of the problem. We use $\Omega^o(\boldsymbol{x})$ to denote the set of all possible scenarios given a schedule $\boldsymbol{x}$. Let $\omega^o \in \Omega^o(\boldsymbol{x})$ be an arbitrary scenario, and $\boldsymbol{\alpha}(\omega^o, \boldsymbol{x})$ and $\boldsymbol{\beta}(\omega^o)$ be the vector of show-up patients and the vector of walk-ins associated with scenario ω^o, respectively. Let $\overline{T}$ be a number such that the probability of any patient waiting after $\overline{T}$ is sufficiently small; one natural choice of $\overline{T}$ is $T + \overline{N_T} - 1$. Let y_t be the total number of patients waiting at the end of slot t and $\boldsymbol{y} = \{y_1, y_2, \ldots, y_{\overline{T}}\}$. It follows that

$$y_t = \begin{cases} (y_{t-1} + \alpha_t(x_t, \omega^o) + \beta_t(\omega^o) - 1)^+ & \text{for } 1 \le t \le T \\ & \quad \text{with } y_0 = 0, \\ (y_{t-1} - 1)^+ & \text{for } T < t \le \overline{T}. \end{cases} \tag{17}$$

Let y_t^s be the number of *scheduled* patients waiting at the end of slot t and $\boldsymbol{y}^s = \{y_1^s, y_2^s, \ldots, y_{\overline{T}}^s\}$. We have

$$y_t^s = \begin{cases} (y_{t-1}^s + \alpha_t(x_t, \omega^o) - 1)^+ & \text{for } 1 \le t \le T \\ & \quad \text{with } y_0^s = 0, \\ (y_{t-1}^s - 1)^+ & \text{for } T < t \le \overline{T}. \end{cases} \tag{18}$$

Note that the number of walk-ins waiting at the end of slot t is $y_t - y_t^s$. Also, $\Gamma_D = T + y_T$, and thus $C_D T$ is a constant that can omitted from the objective function of (P2). We can rewrite (P2) as follows:

$$\min_{\boldsymbol{x} \in \mathbb{Z}_+^T} \mathbb{E}_{\omega^o}\left[\Upsilon(\boldsymbol{x}, \omega^o) - C_I \sum_{t=1}^{T} \alpha_t(x_t, \omega^0)\right], \tag{T1}$$

where

$$\Upsilon(\boldsymbol{x}, \omega^o) = \left\{\sum_{t=1}^{\overline{T}} y_t^s + C_W \sum_{t=1}^{\overline{T}} (y_t - y_t^s) + C_D y_T \,\middle|\, (17), (18)\right\}.$$

The main difficulty in solving (T1) is that it is neither a two-stage linear nor integer programming model. The complicating term is $\alpha_t(x_t, \omega^o)$, which for a given scenario ω^o may not be represented as a linear function of x_t. That is, $\alpha_t(x_t, \omega^o)$ cannot be represented as $f_1(x_t) \times f_2(\omega^o)$ for some $f_1(\cdot)$ and $f_2(\cdot)$. In the next section, we introduce a simple and yet innovative reformulation that transforms (T1) into a stochastic integer programming model.

5.2. Problem Reformulation and Its Matrix Form

We define a new set of decision variables $z_{t,i}$, $t = 1, 2, \ldots, T$ and $i = 1, 2, \ldots, N_S$, such that if patient i is scheduled at t, then $z_{t,i} = 1$, otherwise $z_{t,i} = 0$. We choose N_S to be a sufficiently large number so that at optimality no more than N_S patients would be scheduled. (Lemma 8 in the online appendix shows how to obtain such an N_S.) Let $\boldsymbol{z} = (z_{1,1}, \cdots, z_{1,N_S}, z_{2,1}, \cdots, z_{2,N_S}, \cdots, z_{T,1}, \cdots, z_{T,N_S})' \in \{0,1\}^{T \cdot N_S}$, where the superscript ′ of a vector or a matrix represents the transpose operator. Noting that $x_t = \sum_{i=1}^{N_S} z_{t,i}$, $\forall t = 1, 2, \ldots, T$, we obtain an equivalent two-stage stochastic integer programming model of (T1), described in the following proposition. Let $\Omega(\boldsymbol{z})$ be the set of all possible scenarios given $\boldsymbol{z}$. For a scenario $\omega \in \Omega(\boldsymbol{z})$, $\boldsymbol{\gamma}(\omega) = (\gamma_{1,1}(\omega), \cdots, \gamma_{T,N_S}(\omega))'$ where $\gamma_{t,i}(\omega)$ is the indicator for patient i's show-up status at t (1 means show-up and 0 otherwise), and $\boldsymbol{\beta}(\omega) = (\beta_1(\omega), \cdots, \beta_T(\omega))$ where $\beta_t(\omega)$ is the realized number of walk-ins in t.

Proposition 4. *Problem* (T1) *is equivalent to the following formulation*:

$$\begin{aligned} &\min_{\boldsymbol{z} \in \{0,1\}^{T \cdot N_S}} \mathbb{E}_{\omega}\left[\Upsilon(\boldsymbol{z}, \omega) - C_I \sum_{t=1}^{T} \sum_{i=1}^{N_S} \gamma_{t,i}(\omega) z_{t,i}\right] \\ &\qquad \sum_{t=1}^{T} z_{t,i} \le 1 \textit{ for } 1 \le i \le N_S, \end{aligned} \tag{T1-R}$$

where

$$\Upsilon(\boldsymbol{z}, \omega) = \left\{\begin{aligned} &\min_{\boldsymbol{y}, \boldsymbol{y}^s \in \mathbb{Z}_+^{\overline{T}}} \sum_{t=1}^{\overline{T}} y_t^s + C_W \sum_{t=1}^{\overline{T}} (y_t - y_t^s) + C_D y_T \\ &\qquad y_t \ge y_{t-1} + \sum_{i=1}^{N_S} \gamma_{t,i}(\omega) z_{t,i} + \beta_t(\omega) - 1 \\ &\qquad\qquad \textit{for } 1 \le t \le T \textit{ with } y_0 = 0, \\ &\qquad y_t \ge y_{t-1} - 1 \quad \textit{for } T < t \le \overline{T}, \\ &\qquad y_t^s \ge y_{t-1}^s + \sum_{i=1}^{N_S} \gamma_{t,i}(\omega) z_{t,i} - 1 \\ &\qquad\qquad \textit{for } 1 \le t \le T \textit{ with } y_0^s = 0, \\ &\qquad y_t^s \ge y_{t-1}^s - 1 \quad \textit{for } T < t \le \overline{T}. \end{aligned}\right\}.$$

To economize on notation, we introduce the matrix form of (T1-R) below. Let $\boldsymbol{e}$ be an N_S dimensional unit vector. Let T identity matrices make up $\boldsymbol{W}$—that is, $\boldsymbol{W} = [\boldsymbol{I}\,\boldsymbol{I} \cdots \boldsymbol{I}]$ where $\boldsymbol{I}$ is the N_S dimensional identity matrix. Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_{\overline{T}}, y_1^s, y_2^s, \ldots, y_{\overline{T}}^s)'$. Let $\boldsymbol{c} = (C_W, \ldots, C_W, C_W + C_D, C_W, \ldots, C_W, 1 - C_W, \ldots., 1 - C_W)'$ be a $2\overline{T}$-dimensional vector where all the first $\overline{T}$ elements are C_W except for Tth element and the last $\overline{T}$ elements are $1 - C_W$. Let $\boldsymbol{M}(\omega)$ be a $2\overline{T}$ by $N_S \times T$ matrix, where element $M_{t,(t-1)\times N_S + i}(\omega)$ and $M_{t+\overline{T},(t-1)\times N_S+i}(\omega)$ equal to $\gamma_{t,i}(\omega)$ for all $t \le T, i \le N_S$, and all other elements are 0. Let $\boldsymbol{d}(\omega)$ be a $2\overline{T}$-dimensional vector where the first T elements are $\beta_t(\omega) - 1$ and other elements are –1. Let $\boldsymbol{U}$ be a $2\overline{T}$-dimensional square matrix such that

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}^0 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{U}^0 \end{bmatrix}, \quad \text{where} \quad \boldsymbol{U}^0 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}.$$

Noting that the matrix $\boldsymbol{U}$ is totally unimodular,[5] we conclude that (T1-R) can be simplified into a stochastic integer programming problem where the first stage is a 0-1 integer program and the second stage is a *pure linear program* without integer constraints on $\boldsymbol{y}$. This result is formalized in the following theorem.

Theorem 1. *Problem* (T1) *can be reformulated as follows*:

$$\min_{z \in \{0,1\}^{T \cdot N_S}} \mathbb{E}_\omega\big[\Upsilon(\boldsymbol{z}, \omega) - C_I \gamma(\omega) \boldsymbol{z}\big] \qquad \text{(T2)}$$
$$\boldsymbol{W}\boldsymbol{z} \le \boldsymbol{e},$$

where

$$\Upsilon(\boldsymbol{z}, \omega) = \Big\{ \min_{y \ge \boldsymbol{0}} \boldsymbol{c}'\boldsymbol{y} \,|\, \boldsymbol{U}\boldsymbol{y} \ge \boldsymbol{M}(\omega)\boldsymbol{z} + \boldsymbol{d}(\omega) \Big\}. \qquad \text{(Prim)}$$

5.3. Solution Approaches

5.3.1. Sample Average Approximation.

One common approach to solve a two-stage stochastic programming problem is via sample average approximation (SAA)—that is, randomly generating a sufficient number of sample scenarios and then minimizing the average cost of these samples. With a slight abuse of the notations, we let Ω be the set of all samples randomly generated, and $\omega \in \Omega$ represent one sample in the set. Let $|\Omega|$ denote the number of samples. Then, we can (approximately) solve (T2) by solving the following integer programming problem:

$$\min_{\boldsymbol{y}(\omega) \ge 0, z \in \{0,1\}^{T \cdot N_S}} \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \big[\boldsymbol{c}'\boldsymbol{y}(\omega) - C_I \gamma(\omega) \boldsymbol{z}\big] \qquad \text{(T2-SAA)}$$
$$\boldsymbol{W}\boldsymbol{z} \le \boldsymbol{e},$$
$$\boldsymbol{U}\boldsymbol{y}(\omega) - \boldsymbol{M}(\omega)\boldsymbol{z} \ge \boldsymbol{d}(\omega), \forall \omega \in \Omega.$$

By reformulating the original problem (T1) into a mixed-integer linear program (T2-SAA), we make a challenging problem amenable by many off-the-shelf optimization software packages such as Gurobi.[6] Directly solving (T2-SAA) via optimization software is clearly one solution approach, but this method does not take full advantage of the multimodularity result established in Proposition 3. To leverage this important property of the objective function, we can find a potentially optimal schedule via local search in the first stage and evaluate this solution via solving the second-stage problem. Using the fact that the second-stage problem is a pure linear program (LP), we can further speed up the search procedure in the first stage. Extensive numerical experiments in Section 6 show that our proposed approach, which exploits *both* the structural properties of the objective function *and* the linear reformulation, is much faster than all known methods that can solve the present problem. Details of our proposed approach are illustrated in the section below.

5.3.2. Constraint Generation Algorithm.

Motivated by Wollmer (1980), we can write the dual of the second-stage problem (Prim) for given $\boldsymbol{z}$ and scenario ω as follows:

$$\max_{v \ge 0} \boldsymbol{v}'(\boldsymbol{M}(\omega)\boldsymbol{z} + \boldsymbol{d}(\omega)) \qquad \text{(Dual)}$$
$$\boldsymbol{U}'\boldsymbol{v} \le \boldsymbol{c}.$$

Recall that the primal problem (Prim) is to calculate the cost under scenario ω and decision $\boldsymbol{z}$, so it is always feasible and bounded. Thus, the dual problem (Dual) is also feasible and bounded. Let $\boldsymbol{v}(\boldsymbol{z}, \omega)$ be the optimal solution of (Dual) given $\boldsymbol{z}$ and ω. Denote the set $\{\boldsymbol{z} | \boldsymbol{W}\boldsymbol{z} \le \boldsymbol{e}, \boldsymbol{z} \in \{0,1\}^{T \cdot N_S}\}$ as $\mathscr{Z}$. Let

$$\boldsymbol{a}(\boldsymbol{z}) = \mathbb{E}_\omega[\boldsymbol{v}(\boldsymbol{z}, \omega)'\boldsymbol{M}(\omega)], b(\boldsymbol{z}) = \mathbb{E}_\omega[\boldsymbol{v}(\boldsymbol{z}, \omega)'\boldsymbol{d}(\omega)],$$
$$h(\boldsymbol{z}) = \mathbb{E}_\omega\big[C_I \gamma(\omega)\boldsymbol{z}\big].$$

Proposition 5. *Problem* (T2) *is equivalent to the following Problem* (T2-D):

$$\min_{z \in \mathscr{Z}, \boldsymbol{u}} u \qquad \text{(T2-D)}$$
$$\boldsymbol{a}(\boldsymbol{z}')\boldsymbol{z} + b(\boldsymbol{z}') - h(\boldsymbol{z}) \le u \text{ for all } \boldsymbol{z}' \in \mathscr{Z}. \qquad (19)$$

Proposition 5 is essential in the development of our algorithm. Its proof can be found in the online appendix, and here we provide an intuitive explanation. By strong duality, we know that for each z and ω, the objective value of (Dual) is an upper bound to that of (Prim), and this bound is tight. It can be shown that this relationship also holds when taking expectation with respect to ω. Specifically, let $\Upsilon(\boldsymbol{z}) = \mathbb{E}_\omega[\Upsilon(\boldsymbol{z}, \omega)]$, where $\Upsilon(\boldsymbol{z})$ is the objective value for (Prim). Then, for any given $\boldsymbol{z} \in \mathscr{Z}$, we have

$$\boldsymbol{a}(\boldsymbol{z}')\boldsymbol{z} + b(\boldsymbol{z}') - h(\boldsymbol{z}) \le \Upsilon(\boldsymbol{z}) - h(\boldsymbol{z}), \forall \boldsymbol{z}' \in \mathscr{Z}. \qquad (20)$$

In particular, when $z' = z$, we have

$$\boldsymbol{a}(z)z + b(z) - h(z) = \Upsilon(z) - h(z). \tag{21}$$

Now, let us fix a $z \in \mathscr{Z}$. There is a smallest u, denoted as $u(z)$, which satisfies the subset of the constraints in (19) for that fixed z. That is, $u(z) = \min\{u : \boldsymbol{a}(z')z + b(z') - h(z) \le u, \forall z' \in \mathscr{Z}\}$. By (20) and (21), we know that $u(z) = \Upsilon(z) - h(z)$. Solving (T2-D) leads to the smallest $u(z)$ for $z \in \mathscr{Z}$, equivalent to minimizing $\Upsilon(z) - h(z)$, which is exactly the objective of our original problem (Prim). The algorithm below specifies how to solve (T2-D).

Algorithm 1 Constraint Generation Algorithm (CGA)

```
1: initialize a schedule $z^*$, $\bar{u} \leftarrow \Upsilon(z^*) - h(z^*)$, $\boldsymbol{A} \leftarrow \boldsymbol{a}(z^*)$,
   $\boldsymbol{b} \leftarrow b(z^*)$, $\boldsymbol{e} \leftarrow 1$
2: while $indicator = 1$ do
3:    $indicator \leftarrow 0$
4:    for all neighbors of $z^*$ (in the sense of Definition 2) do
5:       $z^0$ denotes the current neighbor
6:       if $\boldsymbol{A}z^0 + \boldsymbol{b} - h(z^0)\boldsymbol{e} < \bar{u}\boldsymbol{e}$ then
7:          $\boldsymbol{A} \leftarrow (\boldsymbol{A}; \boldsymbol{a}(z^0))$, $\boldsymbol{b} \leftarrow (\boldsymbol{b}; b(z^0))$, $\boldsymbol{e} \leftarrow (\boldsymbol{e}; 1)$
8:          if $\boldsymbol{a}(z^0)z^0 + b(z^0) - h(z^0) < \bar{u}$ then
9:             $\bar{u} \leftarrow \boldsymbol{a}(z^0)z^0 + b(z^0) - h(z^0)$, $z^* \leftarrow z^0$,
               $indicator \leftarrow 1$
10:            break
11:         end if
12:      end if
13:   end for
14: end while
15: return $z^*$
```

In Algorithm 1, z^* represents the best solution found so far, and $\bar{u}$ is a known upper bound for the optimal objective value. Line 1 initializes a solution and its corresponding constraints described by (19). In the "while" loop, *indicator*=1 means that a better solution has been found. In the "for" loop, neighbors of z^* (in the sense of Definition 2) are checked one by one. For the neighbor currently being checked, say, z^0, if the condition $\boldsymbol{A}z^0 + \boldsymbol{b} - h(z^0)\boldsymbol{e} < \bar{u}\boldsymbol{e}$ in line 6 is satisfied, then it has potential to improve z^*, and $\boldsymbol{a}(z^0)$ and $b(z^0)$ are added into $\boldsymbol{A}$ and $\boldsymbol{b}$—that is, constraint generation. Line 8 checks whether z^0 is strictly better than z^*. If so, $\bar{u}$ and z^* are updated; this "for" loop is broken because a better solution has been found, and the algorithm goes back to line 2 and continues to check neighbors of the new z^*. If the condition in line 6 or line 8 is not satisfied, the algorithm goes back to line 2 to check another neighbor of z^* (not updated). If all neighbors of z^* are checked and none can improve z^*, then z^* is optimal (*indicator* becomes 0).

Theorem 2. *Algorithm 1 stops in a finite number of iterations, and its output is an optimal solution of problem (T2).*

5.4. Value of Linear Reformulation (T2)

In Section 5.2, we introduce a simple, and yet innovative variable expansion (from x_t to $z_{t,i}$) to transform the original problem (T1) into a stochastic two-stage linear program (T2) with binary constraints in the first stage. This reformulation makes the original problem much more amenable. Without the reformulation, the original problem (T1) can only be solved by local search (see Proposition 3 and Corollary 2). This approach, although better than complete enumeration, still requires evaluating the cost for each potential solution based on recursive Equations (12) and (14), and can take a long time. In the reformulation (T2), the second stage is a pure LP problem, which can be solved directly to obtain the cost for each potential solution; this is much more efficient than recursive calculations. And, we can further speed up the local search procedure by leveraging the dual (see Algorithm 1). Our extensive numerical study in Section 6 shows that our proposed approach is the most efficient one among those known methods and can solve large-scale problems.

In addition to the above computational benefits, one unique and critical strength of the reformulation (T2) lies in its capability to deal with two important uncertainties in the system with very general forms, which, to the best of our knowledge, cannot be handled by existing approaches.

- **General Walk-ins:** With a general walk-in process (e.g., when walk-in counts in different time slots are correlated), one cannot use recursive Equations (12) and (14) to evaluate the objective function. However, our two-stage programming model and reformulation approach still work. To implement SAA or CGA, one only needs to generate walk-in samples [i.e., $\beta_t(\omega)$'s] based on the joint distribution of walk-ins, which can be any general distribution. The increase in computational times, if any, is only due to random sample generation of walk-ins.
- **General No-shows:** When patient no-show probability depends on time, or become different among patients, the multimodularity result (i.e., Proposition 3) fails. In this case, building an mixed-integer linear program (MILP) model with our linear reformulation is the only known method to get an exact optimal schedule, without complete enumeration. To use our reformulation (T2), one only needs to draw the show-up status variable $\gamma_{t,i}(\omega)$ based on both time slot t and individual patient i, accordingly.

6. Numerical Study

Our numerical study has several purposes. First, we compare existing solution approaches for our model and demonstrate that the Constraint Generation Algorithm developed in this paper is by far the most efficient one. Second, we investigate how changes in the practice environment (e.g., walk-in pattern/volume and no-show rate changes) influence the optimal appointment schedule. This analysis gives rise to important insights on how to manage an appointment-based service in the presence of walk-ins. Third, we develop a simple heuristic policy, which can be used by practitioners as a "rule of thumb" in making their scheduling decisions. We then use real data collected from our collaborating organization to carry out case studies and evaluate the efficiency gains that may result from adopting our proposed approaches (i.e., the scheduling optimization model as well as the heuristic policy) to replace current practice. Full benefit of these proposed approaches can be realized with sufficient demand. Finally, in case patient demand is insufficient to fill up all scheduled appointment slots to optimum, we propose simple "online" scheduling methods (based on our optimization model) to assign patients one by one to appointment slots as their appointment requests arrive; we numerically demonstrate that these online methods perform quite well compared with their offline benchmarks.

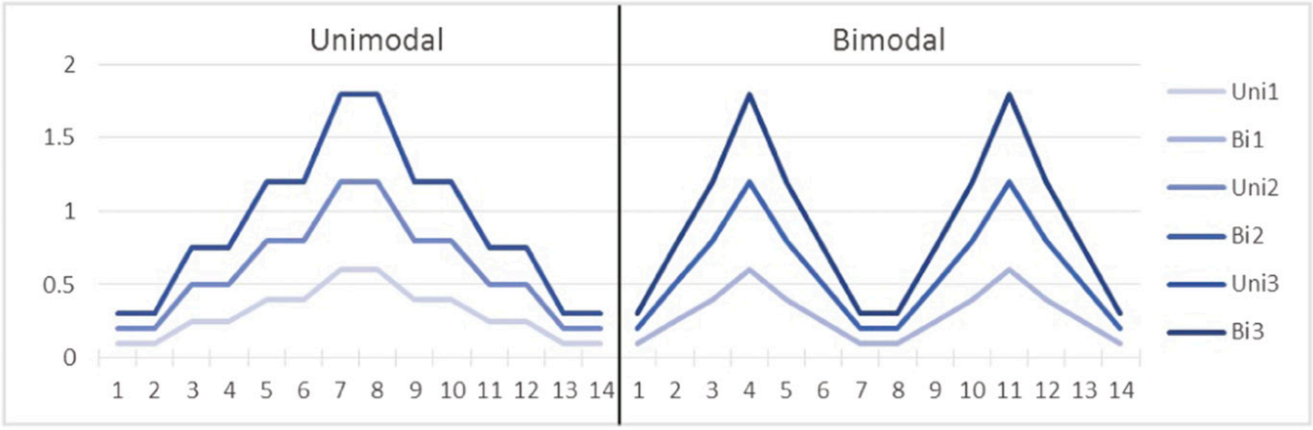
We use a variety of model parameters in our numerical study to capture various settings. For ease of exposition, we focus on $T = 10$ and $T = 14$ as the length of the session.[7] Motivated by our empirical findings in Section 3, we consider two different "shapes" of walk-in pattern: unimodal and bimodal. We vary the walk-in volumes so that the average expected number of walk-ins per slot is 0.3, 0.6, and 0.9, respectively. Figure 2 shows the expected number of walk-ins in each time slot for different scenarios we consider when $T = 14$; in each slot, the number of walk-ins follows a Poisson distribution with the corresponding mean, and the numbers of walk-ins in different slots are independent random variables unless otherwise specified. (A similar figure for $T = 10$ can be found in Online Appendix E.) We consider two levels of no-show probability: 0.5 and 0.1. Previous literature suggests that the provider unit overtime cost is around 15 times of the patient unit waiting cost, and the provider unit idling cost is around 10 times of that (Robinson and Chen 2010, LaGanga and Lawrence 2012, Zacharias and Pinedo 2014). Recall that we normalize the waiting cost for scheduled patients to be 1. We thus set C_D (sum of unit overtime cost and unit idling cost) to be 15 or 25, and set C_I (idling cost) to be 5 or 10. For walk-ins, we set their unit waiting cost C_W to be 0.5 or 0.9.

6.1. Performance Comparison of Different Solution Approaches

In this section, we use an extensive number of problem instances to evaluate and compare the computational performances of four different solution approaches specified below.

- **Local Search** starts from a feasible solution, moves to a neighbor solution defined by Definition 2, if any, that improves the current solution, and continues in this fashion until no better solutions can be found in the neighborhood. Given a schedule, its cost is calculated recursively via (12) and (14). By Proposition 3 and Corollary 2, such a procedure always stops at the optimal schedule. However, the number of neighbors of one schedule is $2^{(T+1)} - 1$, which increases exponentially with the size of the problem, and in addition, the recursive calculations may be computationally expensive.
- **Mixed-Integer Linear Program** approach is developed in Section 5.2 by reformulating the original problem (T1), which is very challenging, into an MILP (T2-SAA), which is amenable by many off-the-shelf optimization software packages. In our numerical experiments, we use Gurobi, perhaps one of the fastest solvers, to solve (T2-SAA) directly. However, this approach disregards the multimodularity property of the objective function.

Figure 2. (Color online) Expected Number of Walk-ins for $T = 14$



Notes. The expected number of walk-ins in each time slot for each scenario is specified as follows. Uni1: (0.1, 0.1, 0.25, 0.25, 0.4, 0.4, 0.6, 0.6, 0.4, 0.4, 0.25, 0.25, 0.1, 0.1), Uni2 doubles Uni1, and Uni3 triples Uni1; Bi1: (0.1, 0.25, 0.4, 0.6, 0.4, 0.25, 0.1, 0.1, 0.25, 0.4, 0.6, 0.4, 0.25, 0.1), Bi2 doubles Bi1, and Bi3 triples Bi1.

- **Local Search + Linear Reformulation (LS + LR)** follows the same procedure as Local Search, the first approach above, except that this approach solves the second-stage linear program in (T2) to get the cost for a given schedule, instead of using recursive Equations (12) and (14). This approach takes advantage of both the multimodularity result and the linear reformulation. Solving a linear program is likely to be faster than using recursive equations.
- **Constraint Generation Algorithm** is proposed in Section 5.3.2. Similar to the third approach (LS + LR), this approach leverages the multimodularity result and linear reformulation. However, it can eliminate a nonoptimal schedule without knowing its cost, and thus it is expected to further speed up the search procedure.

Compared with "Local Search," "MILP" is in general faster, but can be slow in some cases. This may be due to the settings of the solver. As expected, "LS + LR" is much faster than the pure "Local Search" in all cases, suggesting that directly solving the second-stage problem as an LP indeed takes much less time than using recursive equations. In "CGA," the solution search path is identical to those of Local Search and LS + LR, but the objective function can be evaluated much faster by using the dual of the linear reformulation. Thus, CGA ought to be much faster than Local Search and LS + LR. Indeed, CGA performs extremely well and is the best among all four approaches. Specifically, CGA can be 2–5 times faster than LS + LR and 2–10 times faster than the pure Local Search. Details can be found in Table E10 of the online appendix.

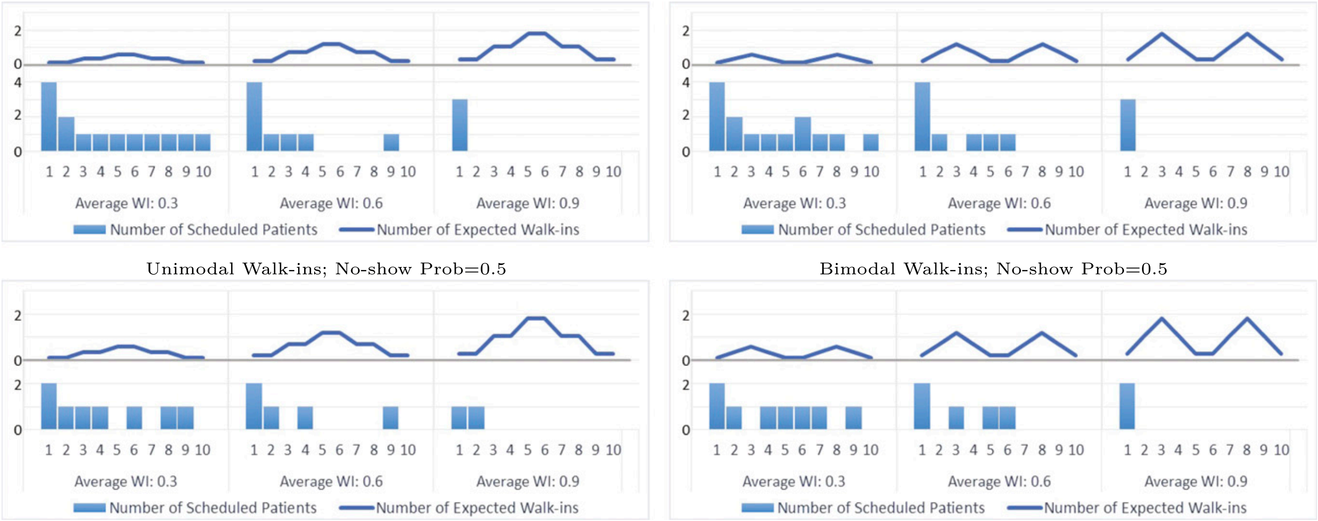
We also test the performance of our solution approaches for problem instances with correlated walk-ins and heterogeneous no-shows. We use CGA to solve problems with correlated walk-ins only. For each of such problem instances, we randomly generate a correlation matrix and then follow Cario and Nelson (1997) to generate multivariate Poisson data with this correlation structure, which are then used as walk-in samples. When heterogeneous no-shows present, MILP is the only known method that can solve the problem for optimality without complete enumeration. We consider two levels of no-show probabilities: 0.5 and 0.1, respectively. Our solution approaches can achieve optimality in these general instances within reasonable amounts of time (see Table E11 in the online appendix).

6.2. Analysis of the Optimal Schedule Pattern

Using the model parameter setting above, we conduct an extensive sensitivity analysis to investigate the impact of walk-in pattern/volume, no-show rate, and unit costs on the optimal schedule. Specifically, we look into the "pattern" and the expected total cost of the optimal schedule, as well as the optimal number of patients to be scheduled.

Figures 3 and 4 depict, respectively, the shape of the optimal schedule under various settings. For ease of discussion, we focus on two cost-parameter settings that represent two extremes of our parameter spectrum. In Figure 3, the unit idling cost is large ($C_I = 10$), the unit overtime cost is small ($C_D - C_I = 5$), and walk-ins are less important ($C_W = 0.5$). In contrast, Figure 4 shows the results when the unit idling cost is small ($C_I = 5$), overtime cost is large ($C_D - C_I = 20$), and walk-ins are more important ($C_W = 0.9$). Each figure contains four panels, and each panel consists of three subfigures. Panels on the left have a unimodal walk-in pattern, whereas panels on the right see a bimodal walk-in pattern. The two panels on the top have higher no-show probabilities than the panels below. Within each panel, the average expected number of walk-ins per slot

Figure 3. (Color online) Comparison of the Optimal Schedules ($T = 10$, $C_I = 10$, $C_D = 15$, $C_W = 0.5$)



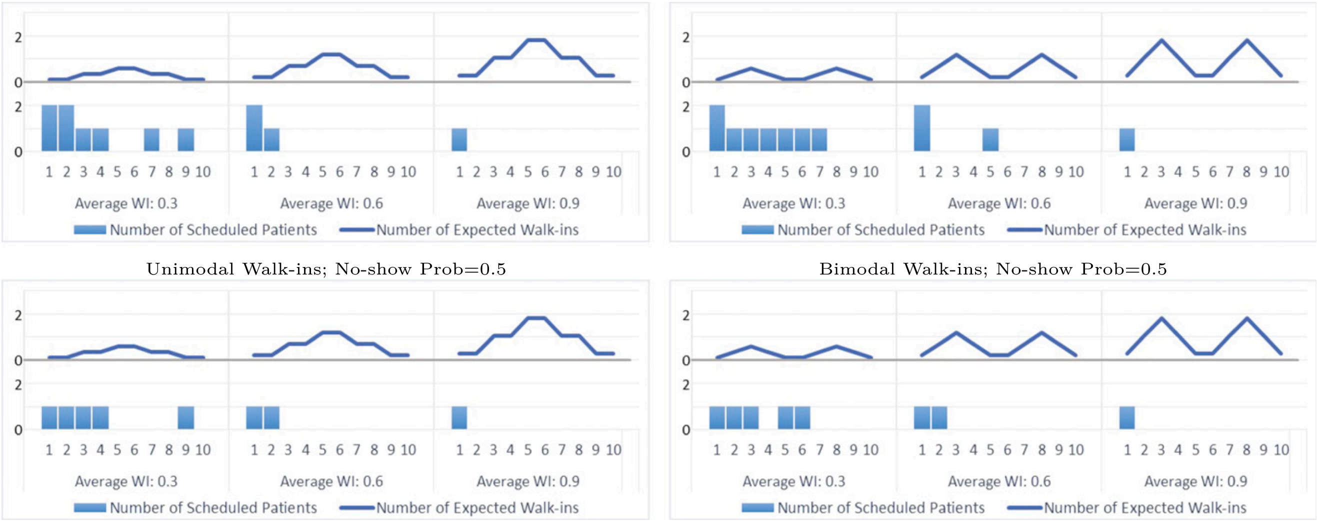
Unimodal Walk-ins; No-show Prob=0.5

Bimodal Walk-ins; No-show Prob=0.5

Unimodal Walk-ins; No-show Prob=0.1

Bimodal Walk-ins; No-show Prob=0.1

Figure 4. (Color online) Comparison of the Optimal Schedules ($T = 10$, $C_I = 5$, $C_D = 25$, $C_W = 0.9$)



Unimodal Walk-ins; No-show Prob=0.5

Bimodal Walk-ins; No-show Prob=0.5

Unimodal Walk-ins; No-show Prob=0.1

Bimodal Walk-ins; No-show Prob=0.1

increases from 0.3 to 0.9 at an increment size of 0.3, from the leftmost subfigure to the rightmost one. In each subfigure, the height of each bar represents the optimal number of patients scheduled in each time slot; the curve on the top shows the expected number of walk-ins arriving in each time slot.

Intuition suggests us to reserve holes (i.e., to purposefully leave some slots empty) in the appointment schedule in anticipation for walk-ins—and—to overbook (i.e., to schedule multiple patients in one slot) to compensate for potential no-shows. These two countervailing forces make it very difficult, if not impossible, to conceive the exact optimal schedule without resorting to an optimization approach. If we think of a consecutive period without scheduled patients as a single hole, we see that the optimal number of holes does *not* have to be the same as the number of modes in the distribution of walk-ins. In addition, overbooking and holes may coexist in an optimal schedule, suggesting that walk-ins cannot fully offset the impact of no-shows and vice versa.

Although the exact optimal schedule depends on a variety of model parameters, we can make a few meaningful observations on the pattern. First, patients tend to be scheduled in slots with low walk-in rates, and holes that are reserved for anticipated walk-ins often follow the peaks of walk-in arrivals (due to queueing effects). When walk-in volume increases, more holes are reserved around peaks of walk-in arrivals. Second, when no-show rate is high and walk-in volume is small, patients tend to be overbooked in early slots. This "front-loading" pattern is consistent with that reported in earlier literature when walk-ins are not considered [see, e.g., Hassin and Mendel (2008)]. However, we can expect that front-loading would disappear if walk-in rate is high in early appointment slots. Third, the optimal schedule tends to (over)book more when the unit idling cost is higher, the overtime cost is lower, and the wait time cost of walk-ins is lower (Figure 3); if these cost parameters change to the opposite direction, the optimal schedule reserves more holes (Figure 4).

Next, we investigate how walk-in patterns, no-shows and cost parameters influence the optimal cost (C^*) and the optimal number of scheduled patients (n^*). Detailed results for $T = 14$ are reported in Table 2, where we also show the cases without walk-ins as a benchmark (results for $T = 10$ can be found in Table E13 of the online appendix). As expected, a larger no-show rate or walk-in volume leads to larger variability in the system, and thus a higher C^*. At the same time, a higher level of no-show rate results in a larger n^*, whereas a larger walk-in volume makes n^* smaller. When the walk-in volume is large, we observe that a bimodal arrival pattern gives rise to a lower C^* compared with a unimodal arrival pattern.[8] This is likely because the variability in the arrival process is smaller for a bimodal walk-in arrival pattern than a unimodal one (given that the average per-slot walk-in rate is fixed). Finally, one noteworthy second-order effect is that when the unit idling cost is smaller and overtime cost is larger, the increase of the overall cost due to the increase in walk-in volumes is much more significant (than the case when the idling cost is larger and overtime cost is smaller). A higher volume of walk-ins means less idling, but more overtime, which may not be easily contained by scheduling decisions alone. Thus, the overall cost is more sensitive to walk-in volumes when unit overtime cost is large (and unit idling cost is small).

From these numerical results, we can glean quite a few important high-level managerial insights. First of all, both walk-in and no-show behaviors create

Table 2. Optimal Cost and Optimal Number of Scheduled Patients ($T = 14$)

No-show	Cost structure			No walk-in		Uni1		Bi1		Uni2		Bi2		Uni3		Bi3	
Prob	C_I	C_D	C_W	C^*	n^*	C^*	n^*	C^*	n^*	C^*	n^*	C^*	n^*	C^*	n^*	C^*	n^*
0.5	5	15	0.5	21.59	23	24.93	14	25.06	14	28.79	6	29.53	6	47.64	3	43.93	2
0.1	5	15	0.5	7.00	14	19.37	9	19.82	9	26.44	4	27.36	4	45.99	2	42.94	1
0.5	10	15	0.5	30.15	26	32.88	19	33.14	19	36.84	10	36.91	10	45.35	5	44.03	4
0.1	10	15	0.5	11.21	15	24.05	11	24.12	11	32.03	6	32.33	6	42.38	3	42.10	2
0.5	5	25	0.5	23.05	22	27.51	13	27.58	12	33.36	5	35.46	5	68.44	2	62.97	1
0.1	5	25	0.5	7.00	14	22.19	8	22.46	8	31.08	3	33.46	3	67.33	2	61.83	1
0.5	10	25	0.5	35.40	24	40.43	16	40.58	16	45.94	7	47.10	8	68.20	4	64.99	3
0.1	10	25	0.5	13.27	15	31.32	10	31.52	10	41.96	5	42.72	5	65.33	2	63.06	2
0.5	5	15	0.9	21.59	23	29.04	13	29.17	13	34.42	4	34.61	5	60.01	2	53.72	1
0.1	5	15	0.9	7.00	14	23.14	8	23.43	8	31.80	3	32.48	3	58.76	2	52.61	1
0.5	10	15	0.9	30.15	26	40.30	17	40.56	17	45.86	9	45.92	8	59.70	4	55.94	3
0.1	10	15	0.9	11.21	15	29.80	10	30.36	10	40.29	5	41.17	5	56.76	2	53.80	2
0.5	5	25	0.9	23.05	22	31.04	11	31.27	11	38.38	4	39.93	4	80.54	2	72.40	1
0.1	5	25	0.9	7.00	14	25.47	8	25.80	8	35.98	3	37.93	3	79.55	1	71.51	1
0.5	10	25	0.9	35.40	24	46.51	15	46.75	15	53.38	6	54.19	7	81.38	3	75.59	2
0.1	10	25	0.9	13.27	15	35.82	9	36.63	9	48.47	4	49.84	4	78.11	2	74.01	1

Note. C^* is the cost of the optimal schedule; n^* represents the optimal number of scheduled patients; walk-in patterns Uni1, Bi1, Uni2, Bi2, Uni3, and Bi3 are specified in Figure 2.

variability (in the arrival process) to the system. Although proper scheduling can counter some of their negative impact, scheduling is not a panacea (because the overall cost still increases as no-shows and walk-ins increase). Practitioners and researchers still need to explore effective ways to reduce no-shows and control walk-ins. Second, walk-in pattern significantly influences the optimal schedule and the system cost. A bimodal walk-in pattern, which "smooths" out walk-in arrivals over time, tends to result in a smaller cost compared with a unimodal walk-in arrival process. Thus, adjusting the walk-in arrival pattern to reduce variability in arrivals, if possible, can be quite useful for practitioners to improve clinic patient flow. Third, if the unit idling cost is small and overtime cost is large—say, in a practice environment where all providers are salaried and no one has strong incentives to overwork—allowing intensive walk-ins may be quite undesirable (because in this case, raising the volume of walk-ins can increase the overall cost significantly, even if one can adjust the appointment schedule properly).

6.3. Heuristic Scheduling Rule

In this section, we design a simple heuristic scheduling policy that can serve as a rule of thumb for practitioners to use. We demonstrate that this simple heuristic performs fairly well, and its optimality gap is on average 10% in our numerical tests.

This heuristic policy has two easy steps. The first step is to *determine n*, the number of scheduled patients. To make it simple, we choose to ignore the waiting cost of patients and only take into account the idling cost and overtime cost of the provider here. Let $\kappa(n)$ be a random variable that represents the total number of patients arriving for services. To determine n, we solve the following simple newsvendor-like optimization problem:

$$\min_{n \in \mathbb{Z}_+} C_I \mathbb{E}(T - \kappa(n))^+ + C_O \mathbb{E}(T - \kappa(n))^-. \tag{22}$$

Let n^h be the solution to (22). Then, the second step is to schedule these n^h patients into T slots. Recall that two major insights obtained in Section 6.2 are (i) slots with high walk-in rates are often kept empty; and (ii) we tend to front-load patients in early slots to counter the negative impact of no-shows. Inspired by these insights, we propose the following simple *allocation rule*.

- First, we try to match "supply" and "demand" in each slot by calibrating the expected number of patients who arrive for service is each slot to be 1, adjusting for their waiting costs. Note that if there are too many walk-ins in a slot or their waiting cost rate C_W is high, we reserve holes.
- If we cannot exhaust allocating all n^h patients in the first phase, we consider front-loading.

The detailed procedure of our allocation rule can be found in Online Appendix D.

We test this simple heuristic using the same parameter settings as in Section 6.2. For $T = 10$ and $T = 14$, the average percentage optimality gaps of this heuristic across all scenarios we tested are 9% and 12%, respectively. Details can be found in Table E14 in the online appendix. Although this heuristic obviously is not as good as solving the optimization model, it performs reasonably well in general. Given the simplicity and easiness to implement, it can be quite useful for practitioners with limited analytical capabilities.

However, it should be cautioned that the performance of this simple rule may not be very robust; in some cases, the optimality gap can be 30% or larger.

6.4. Case Studies

In this section, we examine the potential performance improvement by adopting the optimal and heuristic appointment schedule suggested by our research to current practice. To parameterize our case study, we use the same data set as in Section 3. We select Providers KNI and GAR as cases due to their representativeness: These two providers have quite different walk-in patterns and no-show rates as discussed below. For each provider, we sample a number of days during which he/she works through the whole clinic session (sometimes the providers may leave early).

In May 2011, KNI worked from 9 a.m. to 4 p.m. every Friday, and from 9 a.m. to 1 p.m. every Saturday. We choose to analyze the morning session on Fridays, because KNI takes a lunch break at 1 p.m. As this health center uses half-hour slots, we have eight slots for each morning session. The average patient no-show rate for KNI is estimated to be 0.36 based on the data. As for walk-ins, we use the empirical result in Section 3.3. Recall that the walk-in pattern of KNI in the morning is a time-varying zero-inflated Poisson process with a peak at 10 a.m.

For GAR, we use data of all Fridays from July 1, 2011, to August 5, 2011. During that period, GAR worked from 9 a.m. to 3 p.m. and did not take a lunch break. Thus, we reconstruct 6 original schedules, each with 12 slots. The average no-show rate faced by GAR is estimated to be 0.16, and the walk-ins follow a Poisson process with increasing arrival rates over time.

For each clinic session reconstructed above, we evaluate the expected wait times of scheduled patients and walk-ins, and provider idle time and overtime under the observed schedule, those under the schedule suggested by our optimization model, and those under the heuristic schedule proposed in Section 6.3, based on the provider-specific data. We then calculate changes in these different time components of the objective function if the optimal/heuristic schedule replaces the observed one. A positive change means that the optimal/heuristic schedule reduces the corresponding time component in the observed schedule. Table 3 shows the results for provider GAR if the optimal schedule were adopted (see Tables E15 and E16 of the online appendix for additional results).

Compared with the observed schedule, the optimal schedule adjusts different time components in the objective function according to the cost parameters. In general, if one particular cost parameter becomes larger, the optimal schedule leads to more reduction of the corresponding time component, possibly at the price of a (slight) increase in other time components. For instance, if the idling cost rate C_I increases, the optimal schedule seeks to reduce provider idle time, but may increase other competing time components in the objective function such as provider overtime. Although cost parameters (such as patient waiting cost rate) may not be straightforward to estimate, cost components in the objective function (such as total expected patient wait times) are much more tangible. Thus, information such as that presented in Table 3 can guide the manager to choose a schedule based on her preferred trade-off among these different time components—cost parameters become only a tool to arrive at such schedules.

Table 3. Changes in Objective Function Components by Adopting the Optimal Schedule for Provider GAR (Change Unit: Slots)

$\Delta\Gamma_S$ $\Delta\Gamma_W$	Cost parameters: (C_I, C_D, C_W)															
$\Delta\Gamma_I$ $\Delta\Gamma_O$	(5,15,0.5)		(5,15,0.9)		(5,25,0.5)		(5,25,0.9)		(10,15,0.5)		(10,15,0.9)		(10,25,0.5)		(10,25,0.9)	
7/1/2011	0.00	4.58	0.00	4.85	0.00	6.61	0.00	6.87	0.00	–1.60	0.00	–0.38	0.00	0.95	0.00	1.83
	–0.39	1.30	–0.42	1.27	–1.08	1.45	–1.10	1.43	0.77	0.77	0.66	0.66	0.29	1.13	0.25	1.09
7/8/2011	1.98	0.66	1.98	0.93	1.98	2.70	1.98	2.95	1.98	–5.52	1.98	–4.29	1.98	–2.97	1.98	–2.09
	0.78	1.62	0.75	1.59	0.08	1.77	0.07	1.76	1.93	1.09	1.83	0.98	1.45	1.45	1.41	1.41
7/15/2011	0.71	–4.13	0.71	–3.86	0.71	–2.10	0.71	–1.84	0.71	–10.31	0.71	–9.09	0.71	–7.77	0.71	–6.89
	1.81	0.12	1.78	0.09	1.11	0.27	1.10	0.25	2.96	–0.41	2.86	–0.52	2.48	–0.05	2.44	–0.09
7/22/2011	0.73	6.69	0.73	6.96	0.73	8.73	0.73	8.98	0.73	0.51	0.73	1.74	0.73	3.06	0.73	3.94
	–1.23	1.30	–1.26	1.27	–1.93	1.45	–1.94	1.43	-0.08	0.77	–0.18	0.66	–0.56	1.13	–0.60	1.09
7/29/2011	3.08	12.50	3.08	12.77	3.08	14.53	3.08	14.78	3.08	6.32	3.08	7.54	3.08	8.86	3.08	9.74
	–0.75	2.62	–0.78	2.60	–1.44	2.77	–1.46	2.76	0.41	2.09	0.30	1.99	–0.07	2.46	–0.11	2.42
8/5/2011	3.33	–1.12	3.33	–0.85	3.33	0.91	3.33	1.16	3.33	–7.30	3.33	–6.08	3.33	–4.76	3.33	–3.88
	1.41	2.26	1.39	2.23	0.72	2.41	0.71	2.39	2.57	1.73	2.47	1.62	2.09	2.09	2.05	2.05

Notes. (1) $T = 12$ and the no-show rate is 0.16. (2) Rows represent clinic sessions in different days and columns for different cost parameter settings. (3) Each combination of clinic session and cost parameter setting corresponds to four numbers, the upper left being the reduction of scheduled patients' wait time, the upper right being that of walk-ins' wait time, the lower left being that of provider idle time, and the lower right being that of provider overtime. (4) The measurement unit is appointment slot.

Using different cost parameter settings, we also evaluate the percentage reduction in expected total daily cost if our optimal/heuristic schedules were adopted, for each clinic session reconstructed above. Such a percentage can be viewed as an overall metric to measure the improvement that may result from adopting our proposed scheduling approaches to current practice. We note that, if the optimal schedule was adopted, the potential daily cost savings for provider KNI ranges from 21% to 93%, and from 10% to 67% for provider GAR. On average, KNI sees a 73% cost reduction and GAR 42%. As for the heuristic schedule, the potential daily cost savings for provider KNI range from −14% to 92%, and from −10% to 67% for provider GAR. On average, KNI sees a 64% cost reduction and GAR 39%. This confirms our earlier findings on our heuristic rule: It can be a quite useful tool given its simplicity and good performance overall, but its performance may not be very robust. (Detailed results can be found in Tables E17 and E18 in the online appendix.)

6.5. Dealing with Insufficient Demand

When patient demand is sufficient, one can fill up the daily appointment template with n^* scheduled patients, where n^* is the optimal number of scheduled patients given by our model. In this case, full benefit of our model is realized. However, if patient demand is uncertain and insufficient, then the daily appointment template may not be filled up to optimum. In this section, we discuss how our model may be applied in such a situation and its performances.

Given n^* prescribed by our model, we propose two simple, *online* scheduling policies, which assign patients "on the fly" as their requests for appointments arrive. We will demonstrate that these two *online* policies have very good performances compared with their *offline* benchmarks. The first policy is called horizontal scheduling, which assigns patients from slot 1 throughout T one at a time (if the corresponding slot has at least one scheduled patient in the optimal schedule) and repeats if necessary until all patients have been assigned. The second policy is called vertical scheduling, which assigns up to x_t^* patients to slot t in the order of $t = 1, 2, \ldots, T$ until all patients have been assigned (recall that x_t^* is the optimal number of scheduled patients in slot t). For example, suppose that $T = 3$ and the optimal schedule is (2,0,1); if the realized demand is 2, then we will end up with schedule (1,0,1) by the horizontal policy and (2,0,0) by the vertical policy.

In our numerical experiments, we set $T = 14$ and consider six different walk-in patterns (see Figure 2), eight different cost parameter combinations, and two different no-show probabilities (similar to those considered in Table 2). For each of these parameter settings, we obtain n^* (see Table 2), and then we consider the scenarios in which 1, 2, … or $n^* - 1$ patient requests arrive. For each scenario, we calculate the offline optimal cost—that is, the optimal expected total daily cost if we had known the number of patient requests in advance—by adding to the optimization model such a linear constraint on the total number of patients to schedule. Such an offline optimum represents the best performance of any scheduling policy. We then calculate the percentage gaps between the offline optimum and the system costs under two online policies (horizontal and vertical) described above, respectively. These percentage gaps can be viewed as the optimality gaps of our models applied to situations where demand is insufficient.

Table 4 summarizes, for each walk-in pattern, the average, maximum, and median optimality gaps among all scenarios tested. Both online policies have comparable performances. In particular, the average and median optimality gaps of both online policies are lower than 3%; the maximum optimality gap for the horizontal policy is no more than 10%. All these results suggest that our (offline) optimization model is quite useful, even in an online setting with insufficient demand.

7. Conclusion

In this paper, we study how to schedule patients in a clinic session during which a random number of walk-ins may arrive for services. Scheduled patients, however, may not show up. The objective is to minimize the expected total cost of patient waiting, provider idling,

Table 4. Optimality Gap of the Appointment-Scheduling Optimization Model Under Insufficient Demand

Walk-in pattern	Online policy	No. of scenarios	Walk-in rate = 0.3			No. of scenarios	Walk-in rate = 0.6			No. of scenarios	Walk-in rate = 0.9		
			AVG	MAX	MED		AVG	MAX	MED		AVG	MAX	MED
Unimodal	Horizontal	177	0.7%	7.4%	0.2%	73	1.00%	10.1%	0.0%	22	0.1%	1.7%	0.0%
	Vertical		2.6%	14.8%	1.6%		0.1%	2.5%	0.0%		0.1%	0.5%	0.0%
Bimodal	Horizontal	175	1.3%	6.8%	0.86%	71	0.9%	5.8%	0.4%	10	0.2%	0.6%	0.0%
	Vertical		2.6%	10.8%	2.3%		0.5%	3.7%	0.0%		0.0%	0.0%	0.0%

Notes. (1) $T = 14$. (2) The optimality gap is calculated by $\frac{\text{Online Policy Cost} - \text{Offline Optimal Cost}}{\text{Offline Optimal Cost}}$.

and overtime. We formulate the problem as a two-stage stochastic optimization model and develop effective solution approaches for various settings. Methodologically speaking, our model is the first known approach that can jointly handle general walk-in processes and heterogeneous, time-dependent patient no-show behaviors. Because of this flexibility, our approach can incorporate almost any finding on these uncertain patient behaviors based on empirical data, thus presenting great utility for practitioners.

Our research reveals several important managerial insights. First, with walk-ins, the optimal schedule has a completely different structure from those identified in the previous literature, which often does not consider walk-ins. Intuitively speaking, in anticipation for walk-ins, some appointment slots need to be purposely left empty. However, due to the complex nature of the problem, an optimal schedule is impossible to conceive without resorting to the methods developed in this paper.

Second, walk-ins introduce a new source of uncertainties to the system and cannot be viewed as a simple solution to compensate for patient no-shows. Scheduling, however, is an effective tool to counter some of the negative impact due to uncertain patient behaviors (walk-ins and no-shows). We demonstrate important practical values of our scheduling approaches, by using data from practice to show a significant cost reduction if the providers were to switch from their current schedules (which tend to overlook walk-ins in planning) to the schedules suggested by us. Full benefit of our model can be realized with sufficient demand; when demand is insufficient, our model can still be applied in an online fashion and deliver excellent performances compared with the offline benchmarks.

Last, in addition to optimizing appointment schedules alone, it may be useful to explore means to influence and control uncertain patient behaviors (and thus to mitigate their potential negative impact). For instance, a less variable walk-in process may lead to reduction in overall system cost.

Although our work is motivated by healthcare applications, our optimization models, numerical results, and managerial insights can be applied to general appointment-based services facing random walk-ins. There are several ways to extend our research. First, instead of using discrete slots, one may consider a different modeling approach and decide the scheduled arrival time for each patient. In addition, one may consider adding a service-level constraint in the formulation to limit patient waits, instead of charging a waiting cost in the objective. Second, some previous literature (e.g., Kaandorp and Koole 2007) has considered exponentially distributed service times in their scheduling models. Our model can deal with such random service times, and most results still hold (see Online Appendix F). But it would be meaningful to incorporate generally distributed random service times in the scheduling model. Finally, it may also be of interest to study a decision model that explicitly considers the endogeneity of walk-ins to the provider's work schedule, as discussed earlier. Interesting research questions include, for instance, how to reduce the variability in walk-in pattern and whether additional control policies (like restricting walk-ins during certain hours) could be beneficial. The model in this paper can be a tool, combined with behavioral experiments, to address some of these questions in future research.

Acknowledgments

The authors thank the department editor, the associate editor, and all the referees for constructive comments throughout the review process and Dr. David Rosenthal for connecting us with our collaborating healthcare organization, which motivated this study and provided us with valuable data.

Endnotes

[1] In reality, patients may arrive anytime within a slot. Our assumption above at most misjudges the wait time of a walk-in by half a slot—that is, 10 minutes or so—and thus will not misinterpret individual patient's experience too much.

[2] One can easily show that scheduling patients into overtime slots can never be strictly better than not allowing that. Thus, it suffices to only consider scheduling patients in slot 1 through T but not beyond, as we have done here.

[3] Constraints on $\sum_{t=1}^{T} x_t$, such as an upper bound for it, can be added into the optimization without influencing the main results in this paper.

[4] In Section 4, we note that this priority order is optimal without no-shows if $C_W \leq 1$. The same result still holds when no-shows are present assuming $C_W \leq 1$.

[5] $\boldsymbol{U}$ is totally unimodular because any element in $\boldsymbol{U}$ is 0, 1, or −1, and every row in $\boldsymbol{U}$ has at most two nonzero elements.

[6] Accessed January 17, 2018, http://www.gurobi.com/products/features-benefits.

[7] Our solution approach can solve large-scale problem instances (e.g., $T = 30$) to optimality within a reasonable amount of time; see Table E12 in the online appendix.

[8] When the walk-in volume is relatively low, we do not observe this ordering result, possibly due to the fact that in such a case, the variability in the walk-in process is not significant enough to make a huge difference between the two.

References

Alexandrov A, Lariviere MA (2012) Are reservations recommended? *Manufacturing Service Oper. Management* 14(2):218–230.

Altman E, Gaujal B, Hordijk A (2000) Multimodularity, convexity, and optimization properties. *Math. Oper. Res.* 25(2):324–347.

Berman E (2016) Chief of staff, community healthcare network data provided by personal communication with the authors, April 1.

Bertsimas D, Shioda R (2003) Restaurant revenue management. *Oper. Res.* 51(3):472–486.

Bitran GR, Gilbert SM (1996) Managing hotel reservations with uncertain arrivals. *Oper. Res.* 44(1):35–49.

Cario MC, Nelson BL (1997) Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical Report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL.

Cayirli T, Gunes ED (2014) Outpatient appointment scheduling in presence of seasonal walk-ins. *J. Oper. Res. Soc.* 65(4):512–531.

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

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

Chen RR, Robinson LW (2014) Sequencing and scheduling appointments with potential call-in patients. *Production Oper. Management* 23(9):1522–1538.

Denton B, Gupta D (2003) A sequential bounding approach for optimal appointment scheduling. *IIE Trans.* 35(11):1003–1016.

Djuretic T, Catchpole M, Bingham JS, Robinson A, Hughes G, Kinghorn G (2001) Genitourinary medicine services in the United Kingdom are failing to meet current demand. *Internat. J. STD AIDS* 12(9):571–572.

Gans N, Savin S (2007) Pricing and capacity rationing for rentals with uncertain durations. *Management Sci.* 53(3):390–407.

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

Hajek B (1985) Extremal splittings of point processes. *Math. Oper. Res.* 10(4):543–556.

Hassin R, Mendel S (2008) Scheduling arrivals to queues: A single-server model with no-shows. *Management Sci.* 54(3):565–572.

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

Kaandorp GC, Koole G (2007) Optimal outpatient appointment scheduling. *Health Care Management Sci.* 10(3):217–229.

Kim S-H, Whitt W (2014) Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing Service Oper. Management* 16(3):464–480.

Koeleman PM, Koole GM (2012) Optimal outpatient appointment scheduling with emergency arrivals and general service times. *IIE Trans. Healthcare Systems Engrg.* 2(1):14–30.

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

Kopach R, DeLaurentis P-C, Lawley M, Muthuraman K, Ozsen L, Rardin R, Wan H, Intrevado P, Qu X, Willis D (2007) Effects of clinical characteristics on successful open access scheduling. *Health Care Management Sci.* 10(2):111–124.

LaGanga LR, Lawrence SR (2012) Appointment overbooking in health care clinics to improve patient service and clinic performance. *Production Oper. Management* 21(5):874–888.

Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34(1):1–14.

Liu N (2016) Optimal choice for appointment scheduling window under patient no-show behavior. *Production Oper. Management* 25(1):128–142.

Moore CG, Wilson-Witherspoon P, Probst JC (2001) Time and money: Effects of no-shows at a family practice residency clinic. *Fam. Med.* 33(7):522–527.

Murota K (2005) Note on multimodularity and l-convexity. *Math. Oper. Res.* 30(3):658–661.

Robinson LW, Chen RR (2010) A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing Service Oper. Management* 12(2):330–346.

Vuong QH (1989) Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57(2):307–333.

Wollmer RD (1980) Two stage linear programming under uncertainty with 0–1 integer first stage variables. *Math. Programming* 19(1): 279–288.

Zacharias C, Pinedo M (2014) Appointment scheduling with no-shows and overbooking. *Production Oper. Management* 23(5): 788–801.

Zacharias C, Pinedo M (2017) Managing customer arrivals in service systems with multiple identical servers. *Manufacturing Service Oper. Management* 19(4):639–656.

Zacharias C, Yunes T (2019) Multimodularity in the stochastic appointment scheduling problem with discrete arrival epochs. *Management Sci*. Forthcoming.