POMS PRODUCTION AND OPERATIONS MANAGEMENT

PRODUCTION AND OPERATIONS MANAGEMENT POMS

Vol. 23, No. 12, December 2014, pp. 2209–2223
ISSN 1059-1478 | EISSN 1937-5956 | 14 | 2312 | 2209
DOI 10.1111/poms.12200

Panel Size and Overbooking Decisions for Appointment-Based Services under Patient No-Shows

Nan Liu
Department of Health Policy and Management, Mailman School of Public Health, Columbia University, New York, New York 10032, USA, nl2320@columbia.edu

Serhan Ziya
Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, North Carolina 27599, USA, ziya@unc.edu

Many service systems that work with appointments, particularly those in healthcare, suffer from high no-show rates. While there are many reasons why patients become no-shows, empirical studies found that the probability of a patient being a no-show typically increases with the patient's appointment delay, i.e., the time between the call for the appointment and the appointment date. This paper investigates how demand and capacity control decisions should be made while taking this relationship into account. We use stylized single server queueing models to model the appointments scheduled for a provider, and consider two different problems. In the first problem, the service capacity is fixed and the decision variable is the panel size; in the second problem, both the panel size and the service capacity (i.e., overbooking level) are decision variables. The objective in both cases is to maximize some net reward function, which reduces to system throughput for the first problem. We give partial or complete characterizations for the optimal decisions, and use these characterizations to provide insights into how optimal decisions depend on patient's no-show behavior in regards to their appointment delay. These insights especially provide guidance to service providers who are already engaged in or considering interventions such as sending reminders in order to decrease no-show probabilities. We find that in addition to the magnitudes of patient show-up probabilities, patients' sensitivity to incremental delays is an important determinant of how demand and capacity decisions should be adjusted in response to anticipated changes in patients' no-show behavior.

Key words: service operations; health care management; queueing theory; appointment scheduling
History: Received: August 2012; Accepted: November 2013 by Sergei Savin, after 3 revisions.

1. Introduction

Patient non-attendance (commonly known as "no-shows") at scheduled medical appointments is a serious problem faced by many outpatient clinics (Ulmer and Troxler 2004). Patient no-shows not only cause administrative difficulties for clinics, but can also lead to disruption of the patient-provider relationship and as a result reduced quality of care (Jones and Hedley 1988, Pesata et al. 1999). The financial loss due to patient no-shows can also be substantial (Moore et al. 2001). Studies have identified a variety of factors that correlate with patients' no-show behavior. These include patient characteristics such as age, sex, ethnicity, marital status, and socioeconomic status, but also provider-related factors such as the physician scheduled to be seen and the patient's *appointment delay*, i.e., the time between the patient's call for an appointment and the day the appointment is scheduled (Daggy et al. 2010, Gupta and Wang 2012, Kopach et al. 2007, Norris et al. 2012). In particular, a strong relationship between appointment delays and patients' no-show behavior has been identified in many settings including primary care clinics (Grunebaum et al. 1996), outpatient clinics in academic medical centers (Liu et al. 2010), mental health clinics (Gallucci et al. 2005), outpatient OB/GYN clinics (Dreiher et al. 2008), and health care referral services (Bean and Talaga 1995). The main goal of this article is to investigate the optimal demand and capacity control decisions for a clinic which is cognizant of such a relationship.

There are mainly two leverages that can be used strategically by clinics to control or at least influence appointment delays and thereby reduce the inefficiencies caused by no-shows. One is the size of the population (panel size) the physician (or the clinic) is committed to provide services for (Green and Savin 2008); the other is the number of patients to be seen on each day via perhaps choosing to overbook (LaGanga and Lawrence 2007, Shonick and Klein 1977). These two decisions can be seen as mechanisms that control no-shows indirectly. Many clinics also engage in practices that directly aim to reduce no-shows.

These include sending reminders one or two days before patient appointments, providing financial incentives such as transport vouchers, and charging fines to no-show patients. While such interventions do not eliminate the no-show problem altogether, they typically have a positive effect (see review articles such as Macharia et al. [1992] and Guy et al. [2012]). Because patient population and baseline no-show rates are different, the effects of these interventions may vary significantly (Geraghty et al. 2007, Hashim et al. 2001).

This paper mainly has two objectives. First, to provide insights into the optimal panel size and capacity/overbooking decisions. Second, to investigate how these decisions should be revised in response to changes in patients' no-show behavior, which might be a result of a newly implemented no-show reduction intervention such as those mentioned above. To that end, we adopt the framework used by Green and Savin (2008) and use a stylized representation of a clinic's appointment backlog, which views the scheduled appointments as a single-server queue. Our objective is not to develop a decision support tool that can readily be used to make actual panel size and overbooking decisions in practice but rather to inform such decisions by investigating how the two decisions "should" depend on each other, system characteristics, and patients' no-show behavior.

Specifically, we consider two different scenarios. First, we assume that the daily service capacity is fixed and the clinic does not have the option of overbooking patients. In this scenario, the only decision variable is the arrival rate of the patients, which can equivalently be interpreted as the panel size, and the objective is to maximize throughput, i.e., the long-run average number of patients served per day. For the second scenario, we assume that the clinic's service capacity is somewhat flexible and thus is a decision variable together with the arrival rate. This capacity decision can be seen as the clinic's overbooking decision. We assume that the clinic has a regular daily capacity, but at extra cost it can make additional number of appointment slots available beyond this capacity on a daily basis. A nominal reward of one is accrued for each patient served. The objective of the clinic is to maximize the long-run average *net* reward. For both models, we provide characterizations of the optimal decisions and investigate how the optimal decisions change with changes in patients' show-up probabilities, which might be predicted in response to one of the newly adopted interventions such as sending reminders to patients. One key finding of our analysis is that when making panel size and overbooking decisions patients' sensitivity to incremental delays (i.e., how no-show probabilities change with additional delays) may play a more important role than the magnitude of the no-show probabilities.

One simplifying assumption we make in our mathematical formulation is that patients neither cancel their appointments nor balk. (A patient is said to balk if s/he chooses *not* to book an appointment when offered a long appointment delay.) Patient cancellation and balking are commonly observed in practice and there is evidence to suggest that patients are more likely to cancel or balk when their appointment delays are longer (Diwas and Osadchiy 2012, Liu et al. 2010). It is thus natural to suspect that incorporating such effects could have changed some of the insights that come out of our analysis. However, our simulation study, which we carried out to investigate this question among others, suggested that the key insights generated by our mathematical analysis continue to hold even when patients may cancel their appointments or balk without making any appointments.

Our work is closely related to the operations literature on appointment systems; see Cayirli and Veral (2003) and Gupta and Denton (2008) for in-depth reviews. One way of classifying earlier work is according to the type of waiting modeled. Gupta and Denton (2008) define *direct waiting* for a patient as the time between the patient's arrival to the clinic on the day of her appointment and the time the doctor sees her, and *indirect waiting* as the time between the patient's request for an appointment and the time of her scheduled appointment. Majority of the work in the appointment scheduling literature deals with direct waiting times mostly focusing on the trade-off between patients' waiting time on the day of their appointment and physician utilization.

Since we study the design of systems in which patients exhibit appointment delay-dependent no-show behavior, we use a formulation that captures patients' indirect waiting times. As Gupta and Denton (2008) discuss, very few articles in the literature deal with indirect waiting times. Among the few, Patrick et al. (2008), Gupta and Wang (2008), Liu et al. (2010), Wang and Gupta (2011), and Schütz and Kolisch (2012) all deal with developing effective dynamic scheduling policies to determine whether or not to admit or when to schedule incoming appointment requests given the record of scheduled appointments. Among this group of work, the most relevant one to ours is Green and Savin (2008), from which we adopt the single server queue framework. However, our research questions and the nature of our contribution differ significantly from theirs. Green and Savin (2008) focus on the panel size decisions for a clinic that uses Open Access. In contrast, we are interested in both panel size and overbooking decisions that optimize some system-level objective such as throughput or long-run average net reward. While

Green and Savin (2008) develop a model for estimating the largest panel size that an Open Access clinic can handle, we develop analytically tractable models that lead to useful insights on panel size and overbooking level decisions.

The remainder of this article is organized as follows. In section 2, we introduce our basic formulation and investigate optimal panel size decisions for a clinic that does not overbook. Section 3 builds on the model of section 2 to incorporate overbooking decisions. In section 4, we report the results of our numerical study. Section 5 provides our concluding remarks. The proofs of all the analytical results can be found in the Online Appendix.

2. Panel Size Decisions without an Overbooking Option

We consider an appointment-based service system (e.g., a primary care clinic) where the service provider can control the appointment demand arrival rate. In this section, we assume that the service provider does not have the option of overbooking appointments and thus the service capacity is fixed. Because our objective is to provide insights on general design questions, following Green and Savin (2008), we assume a macroscopic view of the appointment system and model the scheduled appointments as a single server queue. In the rest of the paper, we use the words *patient* and *customer* interchangeably.

2.1. Model Description

Suppose that new appointment requests arrive according to a Poisson process with rate λ, and they are scheduled for the earliest available time. We assume that customers do not cancel their appointments, and therefore the new appointment requests join the appointment queue from the very end. To better interpret how our model approximates what happens in practice, suppose for now that the length of each appointment slot is deterministic with length $1/\mu$. Note that the actual service time of customers may have some variability, but the server is assumed to be able to finish the service within $1/\mu$ units of time. Therefore, when a new patient arrives, the service provider can tell the patient precisely when her appointment is. Our queue is a "virtual" queue for the appointments, a list of scheduled customers. It does not empty out at the end of each day. During the times when the clinic is closed, there will not be any activity in this queue. No one will join and no one will leave. Therefore, we can ignore those "dead" periods, merge the time periods during which the appointment queue is active, and carry out a steady-state analysis with the understanding that time is measured in terms of work days and work hours. Note that in our model, customer waiting time is not determined by waiting in the clinic (called direct waiting) but by waiting elsewhere for the day and time of the appointment to come (called indirect waiting).

When the time for the patient's appointment arrives, the patient may not show up. However, if she shows up, she shows up on time. We assume that whether or not a customer shows up for her appointment depends on the number of customers ahead of her, i.e., the appointment queue length, upon the arrival of her request for appointment.[1] Consider a customer who finds $j \in \mathbb{Z}$ scheduled appointments in the queue (including the customer in service), where $\mathbb{Z}$ denotes the set of non-negative integers. We use $p_j \in [0,1]$ to denote the probability that this customer will show up for her appointment. It is possible that some of these j customers ahead of her may not show up for their appointments, but this does not change the fact that this new customer will have to "wait" for j appointment slots to pass because she will not show up at the clinic until her scheduled appointment time (if she shows up at all). (It might be helpful for the reader to view the server of this queue as "serving" appointment slots as opposed to patients. The server does not idle when the appointment is a no-show, it "serves" the no-show appointment slot. Serving the appointment slot takes the same amount of time regardless of whether the holder of that appointment slot showed up or not and therefore the waiting time of a new patient is determined by the number of currently scheduled appointment slots.)

Motivated by empirical studies (see, e.g., Grunebaum et al. 1996) which find that the length of a patient's appointment delay is positively correlated with her no-show probability, we assume that $p_j \geq p_{j+1}$ for $j \in \mathbb{Z}$ and let $p_\infty = \lim_{j\to\infty} p_j$ (see Corollary 2.11 in Browder, 1996). To avoid a trivial scenario, we also assume that there exists $0 < j < k$ such that $p_j > p_k$, which also implies that $p_0 > 0$. We assume that for every scheduled customer who shows up, the system accrues one nominal unit of reward. If a scheduled patient does not show up or there are no scheduled patients in the queue, the provider might be able to fill in the slot by a walk-in patient who also leaves a reward of one. If neither a scheduled nor a walk-in patient appears in an appointment slot, the provider collects zero reward. We assume that the probability of successfully filling in an empty slot by a walk-in patient is ξ independently of the system state. This assumption is a better fit in cases where the clinic has dedicated providers to serve walk-ins. For example, in two large community health centers, each of which serves more than 26,000 patients annually in New York City, walk-in patients are seen by providers who exclusively see walk-ins and may be diverted to physicians who see scheduled patients only when some of these scheduled ones do not show up (D. Rosenthal, Personal communication, E. Fleck, Personal communication). To avoid an unrealistic and trivial scenario, we assume that $\xi \in [0,1)$.

Defining q_j as the probability that the appointment slot assigned to a patient who sees j appointments ahead will not be wasted (i.e., used by either that particular patient or a walk-in), we have

$$q_j = p_j + (1 - p_j)\xi. \tag{1}$$

We call $\{q_j,\ j = 0,1,\ldots\}$ *fill-in probabilities*. Then, $q_{j+1} = \xi + (1-\xi)p_{j+1} \le \xi + (1-\xi)p_j = q_j$, which in turn implies that the limit of q_j as $j \to \infty$ exists. We use q_∞ to denote this limit.

Define $\Pi_j(\lambda,\mu)$ to be the steady-state probability that there are j appointments in the queue including the ongoing service (which may actually be a "no-show service"). Let $T(\lambda,\mu)$ denote the long-run average reward that the system will collect and $\rho = \lambda/\mu$ be the traffic intensity in the system. Then, we can write

$$T(\lambda,\mu) = \lambda \sum_{j=0}^{\infty} \Pi_j(\lambda,\mu) q_j + \mu(1-\rho)\xi. \tag{2}$$

The first term on the right-hand side of Equation (2) is the reward obtained from patients who show up for their scheduled appointments and walk-in patients who are served in place of no-show patients. From PASTA (e.g., Kulkarni, 1995), $\Pi_j(\lambda,\mu)$ is the probability that there are j scheduled appointments at the arrival time of a new appointment in steady state and with probability q_j this appointment will be filled in either by the patient who makes the appointment or a walk-in patient in case of a no-show. The second term is the reward obtained from walk-in patients when there are no scheduled patients in the queue. To see that, we note that the steady-state probability that the server has no scheduled customers waiting is $1 - \rho$. That is, in the long run, $\mu(1-\rho)$ slots per day have no scheduled customers in them. Since each of these slots will be filled by a walk-in with probability ξ, the long-run average reward rate accrued from these slots is $\mu(1-\rho)\xi$.

Because appointment arrivals occur according to a Poisson process and time spent on each appointment is deterministic, the appointment queue can be modeled as an M/D/1 queue. One can numerically compute the steady-state distribution for this queue, i.e., $\{\Pi_j(\lambda,\mu)\}_{j=0}^{\infty}$ but we do not have a closed-form expression and as a result it is very difficult if not impossible to carry out mathematical analysis and establish structural properties. To overcome this problem, we approximate the steady-state probabilities assuming that service times are exponentially distributed, i.e., the appointment queue is an M/M/1 queue.

For the M/M/1 queue, it is well-known that (e.g., Kulkarni, 1995)

$$\Pi_j(\lambda,\mu) = (1-\rho)\rho^j, \quad \forall j \in \mathbb{Z}, \tag{3}$$

if $\rho = \lambda/\mu < 1$. Then, using Equations (1) and (3), one can write Equation (2) as

$$T(\lambda,\mu) = (1-\xi)\lambda \sum_{j=0}^{\infty} (1-\rho)\rho^j p_j + \mu\xi. \tag{4}$$

Let W denote the waiting time (appointment delay) for a random customer before her service in steady state (regardless of whether the customer shows up or not). It is well known that in an $M/M/1$ queue with an arrival rate of λ and service rate of μ such that $\lambda < \mu$,

$$\mathbf{E}(W) = \frac{\lambda}{\mu(\mu-\lambda)}. \tag{5}$$

The service provider's objective is to maximize the long-run average reward collected, by choosing the appointment demand rate λ for a fixed service capacity μ while making sure that the expected delay (time until appointment for a newly arriving patient) does not exceed a prespecified level κ. Thus, under the M/M/1 approximation, the optimal λ can be found by solving the following optimization problem:

$$\begin{aligned} &\max_{0 \le \lambda \le \mu} \quad T(\lambda,\mu) \\ &\text{s.t.} \quad \mathbf{E}(W) \le \kappa. \end{aligned} \tag{P1}$$

with $T(\mu,\mu)$ defined as $T(\mu,\mu) = \lim_{\lambda \to \mu} T(\lambda,\mu) = \mu[\xi + (1-\xi)p_\infty] = \mu q_\infty$ (see Lemma 1 in the online Appendix) and $\mathbf{E}(W) = \infty$ for $\lambda = \mu$. Note that the long-run average reward $T(\lambda,\mu) = T(\mu,\mu)$ for any $\lambda > \mu$ and therefore one can restrict attention to $\lambda \in [0,\mu]$ in problem (P1). That is, the optimal panel size will never lead to an overloaded system, where the arrival rate exceeds the service rate even when $\kappa = \infty$.

In the following, we provide characterizations of the optimal arrival rate for a fixed value of service capacity with and without a service level constraint on the expected appointment delay, and investigate how these optimal arrival rates change with customers' show-up probabilities. As in Green and Savin (2008), if one assumes that each individual in the panel calls to make an appointment with an exponential rate λ_0, choosing λ is equivalent to choosing the panel size $N = \lfloor \lambda/\lambda_0 \rfloor$ where $\lfloor x \rfloor$ is the integer part of x. Thus, our results for the optimal arrival rate have direct interpretations in the context of optimal panel size decisions.

Table 1 summarizes our notation some of which will be introduced in section 3.

2.2. Characterization of the Optimal Panel Size

We first investigate how the reward function $T(\lambda, \mu)$ changes with the arrival rate λ and give a characterization of the unique optimal arrival rate for problem (P1) for a given μ.

PROPOSITION 1. *For $\lambda \in [0,\mu]$, the long-run average reward $T(\lambda,\mu)$ is a strictly concave function of λ and hence $T(\lambda,\mu)$ has a unique maximizer denoted by $\bar{\lambda}_1$. In addition, if there exists $\tau \in (0,1)$ such that*

$$p_0 + \sum_{j=1}^{\infty}(j+1)\tau^j(p_j - p_{j-1}) = 0, \qquad (6)$$

then $\bar{\lambda}_1 = \mu\tau$; otherwise, $T(\lambda,\mu)$ is strictly increasing in $\lambda \in [0,\mu]$ and $\bar{\lambda}_1 = \mu$. Thus, when $\kappa = \infty$, the unique solution to (P1), λ_1^* is given by $\bar{\lambda}_1$; otherwise it is given by $\min\{\lambda_b, \bar{\lambda}_1\}$ where $\lambda_b = \kappa\mu^2/(\kappa\mu+1)$ is the arrival rate for which the constraint on the expected waiting time is satisfied as an equality.

Let $\rho_1^* = \lambda_1^*/\mu$ denote the optimal traffic load for problem (P1) for fixed μ or equivalently the optimal utilization, i.e., the fraction of time the physician is scheduled to see patients. One important observation we can make from Proposition 1 is that the walk-in probability ξ has no effect on the optimal panel size.

If there is no restriction on the expected delay, i.e., $\kappa = \infty$, the optimal traffic load is independent of the service capacity. In this case, when the appointment delays do not have a significant impact on customers' show-up probabilities, we have $\rho_1^* = 1$. If, however, the no-show rate drops fast as the appointment delay increases, then there exists an optimal arrival rate, which is strictly less than the service rate, i.e., $\rho_1^* < 1$. Thus, even when there is no restriction on the expected waiting time, the service provider does not prefer demand to be as high as possible since high demand would lead to long waiting times, which in turn would result in low show-up rates diminishing the system reward rate. Low demand rates would lead to high show-up rates, but clearly, the service provider would not want to set it so low as to cause the server idle frequently. Thus, there is an ideal value for the arrival rate (an ideal panel size for a healthcare clinic) that helps the system hit the "right" balance.

2.3. Effects of Introducing Policies to Improve Show-Up Probabilities

In this section, we investigate how the panel size should be adjusted in response to the adoption of a new policy, which is expected to change customers' show-up rate. As we discussed in section 1, such policies include making reminder phone calls, sending text messages or email reminders, providing financial incentives, and charging no-show fees. Specifically, we investigate how the optimal panel size changes with the show-up probabilities $\mathbf{p} = \{p_j\}_{j=0}^{\infty}$.

Consider the service system described in section 2.1 with show-up probabilities denoted by $\{p_j\}_{j=0}^{\infty}$. Suppose that once the new policy is adopted, the only change will be in customer show-up probabilities, which we will denote by $\{\hat{p}_j\}_{j=0}^{\infty}$. Also, suppose that once the new policy is adopted, customers are more

Table 1 Notation Used in the Paper

Symbol	Description
λ_0	Individual patient demand rate
N	Panel size
λ	Total patient demand rate, $\lambda = N\lambda_0$
μ	Provider service rate
ρ	Traffic intensity, $\rho = \lambda/\mu$
p_j	Show-up probability when a patient sees j patients ahead of her upon her arrival
ξ	Probability of filling a no-show slot by a walk-in
q_j	Probability that an appointment slot booked by an arriving patient who sees j patients in the system upon her arrival is not wasted, $q_j = p_j + (1-p_j)\xi$
$\Pi_j(\lambda,\mu)$	The steady-state probability that an arrival sees j appointments in the queue
$T(\lambda,\mu)$	The long-run average throughput rate
W	The appointment delay for a random customer before her service in steady state
κ	Maximum allowed value for the expected appointment delay
λ_1^*	The optimal demand rate when overbooking is not an option
ρ_1^*	The optimal traffic intensity when overbooking is not an option, $\rho_1^* = \lambda_1^*/\mu$
$\omega(\mu)$	Daily cost function when the daily service rate of the clinic is set to μ
M	Regular daily capacity of the service provider
$R(\lambda,\mu)$	The expected daily net reward for the service provider in steady-state
$\Lambda(\rho)$	Effective server utilization when the traffic intensity is ρ
λ_2^*	The optimal demand rate when overbooking is an option
μ_2^*	The optimal overbooking level
ρ_2^*	The optimal traffic intensity when overbooking is an option, $\rho_2^* = \lambda_2^*/\mu_2^*$

likely to show-up, i.e., $\hat{p}_j \geq p_j$ for all $j \in \mathbb{Z}$. Now, when is the optimal panel size larger, before the new policy takes effect or after? More precisely, letting $\hat{\lambda}_1^*$ denote the optimal arrival rate when show up probabilities are given by $\{\hat{p}_j\}_{j=0}^{\infty}$, which one is larger, λ_1^* or $\hat{\lambda}_1^*$?

There are two different ways of coming up with an answer to this question based on intuition. First, if patients are more likely to show up under the new policy, i.e., the probability of showing up is higher for any given queue length, the provider might tend to believe that the clinic can handle more patients effectively (after all there is less loss of efficiency due to no-shows) and choose to increase its panel size. Alternatively, one might argue that because patients are more likely to show up, the expected load per patient on the system is higher and thus there is less incentive to admit more patients. Consequently, the optimal panel size should be lower. As it turns out, both of these arguments are flawed. The answer is a little more subtle. First, consider the following simple example:

Example 1. Suppose that $\mu = 20$, $\xi = 0$, and $\kappa = \infty$. Let $p_j = (0.9)^{j+1}$ for $j \in \mathbb{Z}$; $\hat{p}_0 = 1$, $\hat{p}_1 = 0.9$, and $\hat{p}_j = (0.9)^{j+1}$ for $j \in \{2,3,\ldots\}$. Thus, $\hat{p}_j \geq p_j$ for all $j \in \mathbb{Z}$. But, one can show that $\lambda_1^* = 15.19$ while $\hat{\lambda}_1^* = 14.95$. (In the M/D/1 setting, λ_1^* and $\hat{\lambda}_1^*$ are 16.24 and 15.97, respectively.) That is, the optimal panel size is smaller when customers are more likely to show up.

In Example 1, the optimal reward rate increases from 10.39 to 11.01 (from 11.51 to 12.17 in the M/D/1 setting) when $\mathbf{p}$ increases to $\hat{\mathbf{p}}$. In fact, more generally, one can prove that for any fixed λ the reward rate under $\hat{\mathbf{p}}$ is always larger than that under $\mathbf{p}$ if $\hat{p}_j \geq p_j$ for all j. However, when patient show-up probabilities increase, increasing the panel size in response may actually result in lower reward rate. This shows that our first intuitive reasoning, which we discussed above, is incorrect. What really matters when determining the optimal load on the system is the *marginal sensitivity* of customers' show-up probabilities to incremental changes in appointment delays. It is possible that even though customers are more likely to show up, they might have become relatively more sensitive to incremental changes in their delays and this might cause the service provider to try to keep the queue lengths shorter than they used to be.

Now, consider the following condition:

Condition 1. $\hat{p}_{j+1} p_j \geq p_{j+1} \hat{p}_j$ for all $j \in \mathbb{Z}$.

When $p_j > 0$ and $\hat{p}_j > 0$ for all j, the condition above is equivalent to $\hat{p}_{j+1}/\hat{p}_j \geq p_{j+1}/p_j$, which essentially says that show-up probabilities under the new system are less sensitive to additional delays since the percentage drop for additional waiting is always less under the new system. It turns out that Condition 1 is sufficient to ensure that the optimal panel size is larger under the new system.

Proposition 2. *Under Condition* 1, $\hat{\lambda}_1^* \geq \lambda_1^*$. *In other words, the optimal panel size is larger when customer show-up probabilities are less sensitive to additional appointment delays.*

Proposition 2 makes it clear that what matters for the panel size decision is the customers' sensitivity to delays. In Example 1, Condition 1 holds in the opposite direction because $p_{j+1}/p_j = 0.9$, but $\hat{p}_{j+1}/\hat{p}_j = 0.9$ for $j = 0,2,3,\ldots$ and $\hat{p}_2/\hat{p}_1 = 0.81$. Therefore, it is not surprising for the optimal panel size to drop under the new show-up probabilities. Proposition 2 also implies that the intuitive argument that *the optimal panel size should decrease when show-up probabilities increase* is incorrect because one can easily come up with examples in which the show-up probabilities satisfy Condition (1) and $\hat{p}_j \geq p_j$ for all j.

In short, our analysis in this section suggests that with a new intervention that is strongly expected to improve patient show-up rates, providers would realize higher patient throughput if they do not change their panel size. However, one should be careful when choosing a new panel size in order to further benefit from changes in show-up probabilities since changes based on one's intuition alone might be counterproductive. It appears that, it is particularly important for the service provider to get a good sense of how the customers' sensitivities to additional delays will change with the new intervention. If the intervention helps reduce customer sensitivity to additional delays, then our results suggest that there is room for further improvement in throughput by increasing the panel size.

3. Joint Panel Size and Overbooking Level Decisions

One approach clinics use in order to improve the utilization of the appointment slots is to book more appointments than the clinic's regular daily capacity typically allows. In this section, we assume that in addition to the panel size, the service provider can also choose the number of appointments scheduled per day. We model this in a stylized manner by making service rate (i.e., number of appointments scheduled per day) another decision variable in addition to the arrival rate.

3.1. Description of the Model

The assumptions regarding the arrival of the appointment requests, service, and customer no-show behavior are the same as those for the model described in section 2.1. In order to integrate overbooking and panel size decisions, we adopt a reward/cost formulation that is similar to the one used in Liu et al. (2010). Specifically, we assume that for every filled appointment slot, the service provider collects a nominal reward. The daily cost incurred to the clinic is a function of the service rate μ it sets, i.e., the number of appointments scheduled per day. We use $\omega(\mu)$ to represent this cost function. We assume that there is a fixed cost of operating the clinic independently of the service rate chosen by the clinic and we assume that this cost is zero without loss of generality. As for the variable cost, we let $M \geq 0$ be the regular daily capacity of the clinic and thus $\max\{0, \mu - M\}$ can be thought of as the overbooking level. We assume that there is a cost if the clinic chooses to go above this capacity. This cost can be seen as the direct financial cost (e.g., overtime cost for the staff) and/or the indirect cost of patient dissatisfaction as a result of long waits on the day of the appointment and less time devoted to the care of each patient. Intuitively, the more the clinic overbooks, the higher this cost would be; in addition, it seems reasonable to assume that this cost increases faster at a higher overbooking level. Thus, we assume that $\omega(\mu) = 0$ if $\mu \leq M$, $\omega(\cdot)$ is continuous on $[0, \infty)$, strictly increasing and strictly convex on $[M, \infty)$, and twice differentiable on (M, ∞).

Let $R(\lambda, \mu)$ denote the expected daily net reward for the service provider. Then,

$$\begin{aligned} R(\lambda, \mu) &= T(\lambda, \mu) - \omega(\mu) \\ &= (1-\xi)\lambda \sum_{j=0}^{\infty} (1-\rho)\rho^j p_j + \mu\xi - \omega(\mu) \end{aligned} \quad (7)$$

where $T(\lambda, \mu)$ is given by Equation (4). The objective of the service provider is to choose the arrival and service rates which maximize $R(\lambda,\mu)$ while enforcing the expected appointment delay to remain below a certain level κ. Then, our problem (P2) can be written as

$$\begin{aligned} &\max_{\lambda,\mu:0 \leq \lambda \leq \mu} \quad R(\lambda, \mu) \\ &\text{s.t.} \quad \mathbf{E}(W) \leq \kappa \end{aligned} \quad \text{(P2)}$$

with $R(\mu, \mu)$ defined as $R(\mu, \mu) = \lim_{\lambda \to \mu} R(\lambda,\mu) = \mu q_\infty - \omega(\mu)$ and $\lim_{\lambda \to \mu} \mathbf{E}(W) = \infty$.

3.2. Characterization of the Optimal Solution

In this section, we establish some structural properties of the optimal solution to problem (P2). We first study the model without the service level constraint, i.e., setting $\kappa = \infty$. We know from Proposition 1 that for a fixed μ, there exists a unique value of λ that maximizes the reward $T(\lambda, \mu)$. We denote this optimal value by $\lambda(\mu)$. Then, maximizing $R(\lambda, \mu)$ with respect to λ and μ is equivalent to maximizing $R(\lambda(\mu), \mu)$ with respect to μ only.

Let $\bar{\lambda}_2$ and $\bar{\mu}_2$ denote the optimal values for λ and μ in problem (P2) without the waiting time constraint. From Lemma 2, which is provided in the online Appendix, we know that for $0 \leq \mu \leq M$, $R(\lambda(\mu), \mu)$ is a linear and strictly increasing function of μ, which immediately implies that the optimal service rate is no less than the regular daily capacity, i.e., $\bar{\mu}_2 \geq M$. This is not surprising since there is no incentive for the service provider not to use the capacity that is already available with zero additional cost. In order to derive a complete characterization of $\bar{\lambda}_2$ and $\bar{\mu}_2$, we rewrite the reward function $T(\lambda, \mu)$ as follows

$$T(\lambda, \mu) = \mu\Lambda(\rho),$$

where

$$\Lambda(\rho) = (1-\xi)\sum_{j=0}^{\infty}(1-\rho)\rho^{j+1}p_j + \xi. \quad (8)$$

Hence $\Lambda(\rho)$ can be regarded as the "effective" server utilization (proportion of time the server is busy with serving patients, either scheduled ones who actually show up or walk-ins) when the traffic intensity, λ/μ equals ρ.

Let $\omega^+(\mu)$ denote the right derivative of $\omega(\mu)$. Then, $\omega^+(\mu)$ is a strictly increasing function for $\mu \in [M,\infty)$ and it has an inverse, denoted by $(\omega^+)^{-1}(\cdot)$, which is also strictly increasing in its domain. Let $\bar{\rho}_1$ denote the optimal traffic intensity for problem (P1) when $\kappa = \infty$. Recall that $\bar{\rho}_1$ does not depend on μ. Hence, $\Lambda(\bar{\rho}_1)$ is the effective server utilization when system throughput (i.e., long-run average rate at which patients are served) is maximized when there is no restriction on the expected waiting time. Then, we can prove the following proposition.

Proposition 3. *Suppose that $\kappa = \infty$, i.e., there is no restriction on the expected appointment delay. Then, given the show-up probability vector $\mathbf{p} = \{p_j\}_{j=0}^{\infty}$, the optimal service rate $\bar{\mu}_2$ and arrival rate $\bar{\lambda}_2$ for problem (P2) take the following form:*

$$\bar{\mu}_2 = \begin{cases} \infty & \text{if } \omega^+(\mu) \leq \Lambda(\bar{\rho}_1), \quad \forall \mu \geq M, \\ M & \text{if } \omega^+(M) \geq \Lambda(\bar{\rho}_1), \\ (\omega^+)^{-1}(\Lambda(\bar{\rho}_1)) & \text{otherwise,} \end{cases}$$

and $\bar{\lambda}_2 = \bar{\rho}_1\bar{\mu}_2$. Furthermore, $(\bar{\lambda}_2, \bar{\mu}_2)$ is the unique optimal solution to problem (P2).

The expression for $\bar{\mu}_2$ provided in Proposition 3 may seem technical but in fact it has a straightforward interpretation. Notice that $\omega^+(\mu)$ is the marginal cost of additional unit capacity when the service capacity is μ. The service provider would be willing to increase the service capacity (and the arrival rate along with it) up to the point where marginal cost equals the rate with which the system generates revenue, which is equal to the effective server utilization. This corresponds to the third case in the statement for $\bar{\mu}_2$ in Proposition 3. However, if the marginal cost is below this revenue generation rate no matter what the service capacity is (which is unlikely in practice), then there is no point in restricting the number of people to be seen on a given day and thus $\bar{\mu}_2 = \infty$. If the marginal cost is higher even at the regular capacity, then there is no incentive to overbook and thus $\bar{\mu}_2 = M$.

Next, we consider problem (P2) with a non-trivial service level constraint, i.e., $\kappa < \infty$.

COROLLARY 1. *If $\kappa < \infty$ and there exists a finite μ such that $\omega^+(\mu) > 1$, then there exists a finite optimal value for μ.*

The condition given in Corollary 1 essentially implies that as one adds more appointments for a given day there is a certain level beyond which the incremental benefit of having one more appointment is outweighed by its incremental cost. Suppose that this realistic condition holds. It is not possible to obtain closed-form expressions for optimal arrival and service rates. However, we can show that optimal rates possess some convenient structural properties, which can be helpful in devising simple solution methods. Let (λ_2^*, μ_2^*) denote an optimal arrival and service rate pair when there is a constraint on the expected waiting time. Then, we can show the following.

PROPOSITION 4. *Let γ be defined as $\gamma = \begin{cases} \frac{\bar{\rho}_1}{\kappa(1-\bar{\rho}_1)} & \text{if } \bar{\rho}_1 < 1, \\ \infty & \text{if } \bar{\rho}_1 = 1. \end{cases}$ Then, for a fixed show-up probability vector **p**=$\{p_j\}$, if $\gamma < \bar{\mu}_2$, then $\mu_2^* = \bar{\mu}_2$ and $\lambda_2^* = \bar{\lambda}_2$. Otherwise, $\mu_2^* \le \gamma$ and the service level constraint is binding at optimality, i.e., $\lambda_2^*/\mu_2^* = 1 - 1/(\kappa\mu_2^* + 1)$.*

Proposition 4 suggests a relatively easy way to obtain an optimal solution. If $\gamma < \bar{\mu}_2$, then the optimal solution is given by the optimal solution to (P2) with no service level constraints, which is directly available from Proposition 3. If $\gamma \ge \bar{\mu}_2$, then the problem reduces to an optimization problem with a single decision variable since in this case λ_2^* can be expressed explicitly in terms of μ_2^*. More specifically, an optimal service rate can be obtained by solving the following optimization problem:

$$\max_{\mu \ge 0} R_b(\mu) \tag{9}$$

where

$$R_b(\mu) = (1-\xi)\mu\left(1 - \frac{1}{\kappa\mu+1}\right)\sum_{j=0}^{\infty}\left(\frac{1}{\kappa\mu+1}\right)\left(1-\frac{1}{\kappa\mu+1}\right)^j p_j + \mu\xi - \omega(\mu). \tag{10}$$

Consider the non-trivial case where $\bar{\rho}_1 < 1$. When $\kappa = \infty$ meaning that there is no restriction on the expected delay, $\gamma = 0$. Consequently, $\mu_2^* = \bar{\mu}_2$ and $\lambda_2^* = \bar{\lambda}_2$, as expected.

Since $R_b(\mu)$ defined in Equation (10) is not necessarily unimodal in μ, multiple optimal solutions may exist. Therefore, in the following, we shall refer to μ_2^* as the smallest optimal service rate, i.e., $\mu_2^* = \inf\{\mu^o : R_b(\mu^o) \ge R_b(\mu) \text{ for all } \mu \ge 0\}$. Then, we know from Proposition 4 that the corresponding optimal arrival rate λ_2^* and the optimal traffic load defined as $\rho_2^* = \lambda_2^*/\mu_2^*$ are also the smallest choices for these two variables.

3.3. Effects of Introducing Policies for Improving Show-Up Probabilities and Changing the Service Level Requirement

In this section, we investigate the sensitivity of the optimal panel size and overbooking decisions (optimal arrival and service rates in our formulation) to customers' show-up probabilities and κ, the service level requirement on the expected waiting time. In section 2.3, we showed that when overbooking is not an option and there is a fixed daily capacity, the optimal panel size is larger when customers' show-up probabilities are less sensitive to additional delays. When overbooking level is also a decision variable together with the panel size, it is not clear how "improvements" in show-up probabilities would affect the optimal decisions. When show-up rates of appointment slots are less sensitive to additional delays, does that mean that the service provider has less incentive to overbook since there is less uncertainty regarding whether or not the scheduled appointments will actually be filled? As for the panel size, if the optimal overbooking level is higher, intuition suggests that the optimal panel size would be larger as well, but it is difficult to predict how it would change otherwise. In any case, we find that the optimal panel size and overbooking level might change in unpredictable ways.

We first investigate the sensitivity of the optimal decisions to show-up probabilities. As in section 2.3, suppose that as a result of a new policy that aims to improve show-up rates, customer show-up probabilities $\{p_j\}_{j=0}^{\infty}$ become $\{\hat{p}_j\}_{j=0}^{\infty}$ and let $\hat{\lambda}_2^*$ and $\hat{\mu}_2^*$, respectively, denote the optimal demand and service rates under this new policy (the smallest optimal values in the unlikely event that there are multiple optimal solutions). Then, we can prove the following proposition:

PROPOSITION 5. *If $\hat{p}_j \geq p_j$ for all $j \in \mathbb{Z}$, then $\hat{\mu}_2^* \geq \mu_2^*$. If $\hat{p}_j \geq p_j$ for all $j \in \mathbb{Z}$ and Condition 1 holds (i.e., $\hat{p}_{j+1}p_j \geq p_{j+1}\hat{p}_j$ for all $j \in \mathbb{Z}$), then $\hat{\lambda}_2^* \geq \lambda_2^*$, $\hat{\mu}_2^* \geq \mu_2^*$, and $\hat{\rho}_2^* \geq \rho_2^*$, where $\rho_2^* = \lambda_2^*/\mu_2^*$ and $\hat{\rho}_2^* = \hat{\lambda}_2^*/\hat{\mu}_2^*$.*

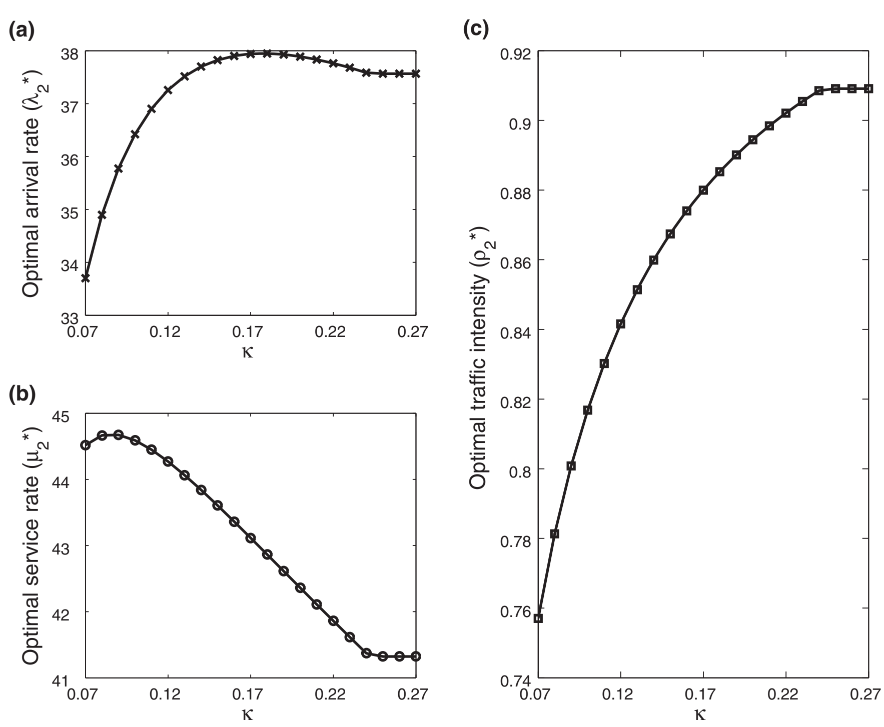
REMARK 1. In the second part of Proposition 5, it is sufficient to assume that $\hat{p}_0 \geq p_0$ (instead of $\hat{p}_j \geq p_j$ for all $j \in \mathbb{Z}$) together with Condition 1.

According to Proposition 5, if customers are more likely to show up under the new policy, the service provider chooses to increase the overbooking level. There is a cost of working with a higher overbooking level but the service provider knows that appointment slots are more likely to be filled in and thus the system is more likely to benefit from an increase in the capacity. However, it turns out that even if the service provider chooses to work at a higher overbooking level, it does not mean that she would choose to work with a larger panel as well (see Example 2 below). A larger panel size and a higher overbooking level are not guaranteed to be optimal when customers are more likely to show up. These show-up rates also need to be less delay sensitive for the service provider to choose both a higher overbooking level and a larger panel size.

EXAMPLE 2. Suppose that $p_0 = 0.4$ and $p_j = 0.38$ for all $j \geq 1$, and $\hat{p}_0 = 1$ and $\hat{p}_j = 0.4$ for all $j \geq 1$. Let $\xi = 0$, $\kappa = \infty$ and $\omega(\mu) = a\mu^2$ where a is a positive constant. Then, $T(\lambda,\mu) = \lambda[0.4(1-\rho) + 0.38\rho] = \mu(0.4\rho - 0.02\rho^2)$. For a fixed μ, $T(\lambda,\mu)$ is strictly increasing in $\rho \in [0,1]$. Hence $\bar{\rho}_1 = 1$ and $R(\lambda(\mu),\mu) = 0.38\mu - a\mu^2$. It then follows that $\lambda_2^* = \mu_2^* = \frac{19}{100a}$. Now, note that $\hat{T}(\lambda,\mu) = \lambda[(1-\rho) + 0.4\rho] = \mu(\rho - 0.6\rho^2)$. Hence $\hat{\lambda}_2^* = \frac{5}{6}\hat{\mu}_2^*$, and $\hat{R}(\lambda(\mu),\mu) = \frac{5}{12}\mu - a\mu^2$. It follows that $\hat{\mu}_2^* = \frac{5}{24a} > \mu_2^* = \frac{19}{100a}$ as expected since $\hat{p}_j > p_j$ for all j; however, $\hat{\lambda}_2^* = \frac{5}{6} \times \frac{5}{24a} = \frac{25}{144a} < \lambda_2^* = \frac{19}{100a}$ and $\hat{\rho}_2^* = \frac{5}{6} < 1 = \rho_2^*$.

Next, we investigate how changes in the service level requirement, which is determined by κ, affect the optimal decisions. For example, if the service provider commits herself to providing customers shorter waiting times, how should she adjust the panel size and the daily overbooking level? Intuition suggests that because the goal is to reduce the average waiting time, a reasonable adjustment would be to reduce the panel size and increase the overbooking level appropriately. Interestingly, however, that is *not* necessarily the case. As an example, suppose that $p_j = 0.99^j$ for $j \in \mathbb{Z}$ and $\omega(\mu)=0.01\mu^2$. In this case, Figure 1 shows how the optimal demand rate λ_2^*, the optimal service rate μ_2^*, and the optimal traffic load $\rho_2^* = \lambda_2^*/\mu_2^*$

Figure 1 The Optimal Decisions and Traffic Intensity vs. Service Level Parameter κ



change with the service level parameter κ which varies from 0.07 to 0.27. Recall that κ is the allowable maximum value set by the provider for the average waiting time of scheduled customers. One can observe that the optimal arrival rate and service rate are not monotone in κ. Interestingly, for small values of κ, the optimal decision is to increase the overbooking level even as the service level requirement gets less restrictive, i.e., as κ gets larger (see Figure 1b). However, for sufficiently large values of κ, the optimal panel size decreases with less restriction, i.e., with larger κ (see Figure 1a).

In Figure 1c, we observe that, unlike the optimal arrival and service rates, the optimal traffic load behaves as expected. It is monotonically increasing in κ. In fact, this behavior is not exclusive to this particular example and Proposition 6 proves that the optimal load ρ_2^* is an increasing function of κ.

Proposition 6. *The optimal traffic load ρ_2^* increases as the service level requirement becomes less restrictive, i.e., ρ_2^* increases with κ.*

According to Proposition 6, if the provider works with a less strict service level, the load on the system, and as a result expected length of the appointment queue will increase under the optimal policy. However, as our earlier example demonstrates, this does not mean that the clinic will actually be serving a larger panel of patients since from the clinic's point of view, it might be preferable to decrease both the panel size and overbooking level appropriately.

Finally, in this section we investigate how the optimal panel size and overbooking level change with walk-in probability ξ. When overbooking is not an option, we found that the walk-in probability has no effect on the optimal panel size. However, it turns out that when overbooking is an option, higher walk-in probabilities not only lead to higher overbooking levels but also larger panel sizes and traffic intensities.

Proposition 7. *When ξ increases, the optimal demand rate λ_2^*, the optimal overbooking level μ_2^* and the optimal traffic intensity ρ_2^* for problem (P2) increase.*

When the probability that an appointment slot is not wasted is higher, overbooking is more likely to benefit the clinic and thus the optimal overbooking level is higher. This result may seem counterintuitive at first. After all one reason clinics overbook is to alleviate the effect of no-shows on the system. However, one should note that the increase in the overbooking level is mainly due to the fact that the panel size is also a decision variable. When the overbooking level is higher, the clinic can handle more patients on a daily basis and therefore the optimal panel size is also larger. In fact, it turns out that the fractional change in the optimal panel size is larger than the fractional change in the optimal overbooking level, leading to a larger traffic load. When the no-show slots are more likely to be filled in by walk-ins, the clinic can handle a larger load efficiently.

In summary, lower no-show rates will always help if one is willing to keep the same panel size and overbooking level. There is room for further improvement if the provider is willing to make some changes. However, our analysis in this section suggests that one needs to be careful when choosing a new panel size and overbooking level as it might have some unexpected consequences. For example, even though a lower no-show rate encourages the provider to see more patients in a day, it does not mean that the provider should increase the panel size. Only when patient sensitivity to additional delays also decreases, a larger panel size would necessarily be more beneficial. On a separate note, our results also point to interesting relationships among expected appointment delay (the service level requirement), optimal panel size, and overbooking level. As it turns out, requiring the expected appointment delay to be lower may lead to a larger optimal panel size or a lower overbooking level.

Before we move on to the description and discussion of our numerical study, it might be helpful to provide a summary of our key analytical findings established in sections 2 and 3. In particular, Table 2 sorts out the conditions needed for the optimal panel size and overbooking level to be larger depending on whether only the former or both can be freely determined by the service provider.

4. Numerical Study

In this section, we report the findings of our extensive numerical study conducted to investigate whether the insights obtained using our analytical model depend on some of the key modeling assumptions (sections 4.1 and 4.2) and also how the panel sizes that are optimal according to our formulation compared with

Table 2 Summary of the Key Insights

Scenarios	N^*	(N^*,μ^*)
No-show rate ↓	?	(?, ↑)
Sensitivity to delay ↓	↑	(?,?)
No-show rate ↓ & Sensitivity to delay ↓	↑	(↑,↑)

The second column indicates the changing direction of the optimal panel size N^* (when only the panel size can be chosen) in different scenarios. The third column shows the changing directions of both the optimal panel size and overbooking level (N^*,μ^*) when both can be chosen. "↑", "↓" and "?" respectively mean "increases," "decreases," and "no definite answer."

those that are determined to be ideal for implementing Open Access in the literature (section 4.3).

4.1. Comparison of the M/M/1 and M/D/1 Models

So far in this paper, mainly for analytical tractability, we assumed that the appointment queue can be modeled as an M/M/1 queue. In some respects, however, the M/D/1 queue can seem like a more fitting choice. This is mainly because in our formulation the server is essentially "serving" appointment slots whose lengths are deterministic. It is thus of interest to investigate whether the results would change significantly if the appointment queue were assumed to operate as an M/D/1 queue. First, it is important to note that as we have already indicated, Example 1 demonstrates that under both the M/M/1 and the M/D/1 setup, having patients who are more likely to show-up does not mean that the optimal panel size is also larger. Thus, our numerical experiments concentrated on investigating whether having less delay sensitive patients, as defined in Condition 1, would lead to a larger optimal panel size even under the M/D/1 setup.

To populate our numerical experiments, we adopted the following parametric form for patient show-up probabilities:

$$p_j = \frac{1}{1 + e^{\alpha+\beta j}}, \tag{11}$$

where α and β are parameters and $\beta > 0$ so that p_j decreases with j. One reason we chose this parametric form was that it naturally arises if one uses logistic regression to estimate show-up probabilities as a function of patients' appointment delay. The form is also compatible with Condition 1 in the sense that if we let $\hat{p}_j = 1/(1 + e^{\hat{\alpha}+\hat{\beta}j})$, then Condition 1 holds when $\hat{\alpha} \le \alpha$ and $\hat{\beta} \le \beta$ (see Lemma 6 in the Appendix). One implication of this in light of Propositions 2 and 5 is that under our M/M/1 formulation any collective increase in the estimated show-up probabilities would lead to a larger optimal panel size and a higher optimal overbooking level regardless of which one of the two parameters the change is captured by.

In the numerical experiments, we assumed that the regular daily capacity $M = 20$ and the daily cost function $\omega(\mu) = 0.2\times[(\mu-M)^+]^2$. We varied the show-up probability parameters α and β, the delay threshold κ and the walk-in rate ξ. Specifically, we considered 36 different combinations of these four parameters by choosing them so that $\alpha \in \{-5,-3,-1\}$, $\beta \in \{0.01,0.03,0.05\}$, $\kappa \in \{0.5,1\}$, and $\xi \in \{0,0.5\}$. For each combination, we calculated the optimal panel size and overbooking level under the assumption that the appointment queue operates as an M/M/1 queue as well as the assumption that it operates as an M/D/1 queue. Table 3 provides the results for a selected subset of the 36 combinations.

While not all the results are reported here, we find that across all 36 combinations we tested, the average absolute percentage difference between the optimal panel size under the M/M/1 setup and that under the M/D/1 setup is only 3.3%. That for the overbooking level and the average net reward are only 0.23% and 2.98%, respectively. Furthermore, all the key structural results obtained under the M/M/1 setup continue to hold under M/D/1 setup. In particular, for any fixed combination of κ and ξ, the optimal panel size and overbooking level in both cases are decreasing in α for fixed β and decreasing in β for fixed α.

4.2. Impact of Customer Cancellation and Balking

One simplifying assumption we made in our mathematical analysis was that patients neither cancel their appointments nor balk, i.e., choose not to join the appointment queue when they find the appointment delay long. In this section, we investigate the effect of this simplification on our key findings via simulation.

In our simulation model, we assumed that a patient who finds j scheduled appointments in the queue, independently of the others, would choose to join the appointment queue with probability $e^{-\eta j}$, where $\eta > 0$ is the parameter for balking intensity. This probability model captures the fact that patients are more likely not to schedule an appointment when the appointment queue/delay is longer. Note also that a larger η indicates a higher likelihood to balk for any given queue length j.

To model cancellations, we assumed that each patient who joined the appointment queue, independently of the others and the system state, would choose to cancel her appointment after some random amount of time that has exponential distribution. However, cancelation only occurs if the patient has not received service until then. In line with what mostly happens in practice, when an appointment is canceled, appointments that follow the canceled appointment are not rescheduled to fill in the newly

Table 3 Selected Comparison Results between the M/M/1 and M/D/1 Systems

Parameters	N_2^*(M/M/1)	μ_2^*(M/M/1)	N_2^*(M/D/1)	μ_2^*(M/D/1)
$(\alpha,\beta) = (-5, 0.05)$	2642	22.3	2709	22.4
$(\alpha,\beta) = (-3, 0.05)$	2546	22.1	2630	22.2
$(\alpha,\beta) = (-1, 0.05)$	2342	21.4	2449	21.5
$(\alpha,\beta) = (-1, 0.03)$	2428	21.5	2516	21.6
$(\alpha,\beta) = (-1, 0.01)$	2551	21.6	2608	21.7

The optimal panel size is denoted by N_2^*, and μ_2^* represents the optimal overbooking level. The results are derived by assuming $\kappa = 1$ and $\xi = 0$.

vacated slot. Thus, the canceled appointment slot leaves a "hole" in the queue leaving other scheduled appointments intact. When a new appointment request arrives, rather than scheduling it at the end of the queue, one of these holes is filled if there are any. Specifically, the patient is scheduled for the slot that is closest to service. (If the first available slot happens to be the first one in the queue, then the second hole in the queue is filled because the "service" for the first slot has already started. This is to capture the fact that very late cancelations cannot be filled.) If there are no holes in the queue, the appointment is scheduled at the end of the queue. As in the previous section, patients who do not cancel their appointments, show up with probabilities that follow the parametric form Equation (11).

We considered three different balking intensities $\eta \in \{0, 0.001, 0.003\}$ and three different cancelation rates $\theta \in \{0.1, 0.2, 0.5\}$. For each combination of η and θ, we varied the show-up probability parameters. In particular, we chose $\alpha \in \{-5, -1\}$ and $\beta \in \{0.01, 0.05\}$. To get a better sense of what exactly these choices mean, consider a provider who sees 20 patients per day. Then, when $\eta = 0.001$, for a patient who finds an appointment queue length of 2 or 5 days, and 10 days at the time of her arrival, the balking probability is respectively 4%, 10%, and 18%. When $\theta = 0.1$, her cancelation probability before appointment is, respectively, 18%, 39%, and 63%. Finally, when $\alpha = -5$ and $\beta = 0.01$ her no-show probability is, respectively, 1%, 2%, and 5%.

We assumed that the regular daily capacity $M = 20$, the daily cost function $\omega(\mu) = 0.2 \times [(\mu - M)^+]^2$ and service times were deterministic. We also assumed there were no walk-in patients and there is no restriction on the expected delay, i.e., $\xi = 0$ and $\kappa = \infty$. Thus, in total, we considered 36 combinations for the choice of $(\eta, \theta, \alpha, \beta)$. For each combination, we varied the panel size N with a step size of 10 and the daily service capacity level μ with a step size of 0.1, and ran simulation for each pair of (N, μ). We used the batch-means method with a batch length of 3000 days and 11 batches, where the first batch was used as the warm-up period. We computed the average reward for each pair of (N, μ) and named the pair that gave the largest average reward the optimal solution. We summarize our results in Table 4, where Cases 1 through 4 correspond to four different combinations of the show-up probability parameters $(\alpha, \beta) = (-5, 0.01)$, $(-5, 0.05)$, $(-1, 0.01)$, and $(-1, 0.05)$, respectively.

One can observe from Table 4 that the higher the balking intensity, the larger the optimal panel size. This is not surprising because balking makes the appointment queue lengths (and hence delays) shorter on average, which in turn means that patients are more likely to show-up giving the clinic an incentive to increase its panel size. The effect of balking on the overbooking level is more salient. We observe that the optimal overbooking level does not appear to have a monotone relationship with the balking intensity, most likely due to the fact that the clinic has the flexibility to choose the panel size as well. The clinic can manage different degrees of balking behavior by playing with the panel size appropriately but keeping the overbooking level more or less constant.

As in the case of balking, a higher cancellation rate leads to a larger optimal panel size (see Tables 3 and 4). However, the reader should note that in our simulation model, patients who cancel their appointments do *not* reschedule, and thus cancellation has the effect of reducing the load on the clinic and thereby enabling it to handle a larger panel of patients. Thus, while our results suggest that with no or little rescheduling, the optimal panel size increases with cancellation rate, it would be natural to expect that the same insight might no longer hold when a high percentage of canceled appointments are rescheduled.

Cases 1 through 4 can be seen as being ordered from being the least sensitive to incremental delay to the most. From Lemma 6, we know that in Case 2 patients are more sensitive than patients in Case 1 and patients in Case 4 are more sensitive than patients in Case 3. Cases 2 and 3 can actually not be ordered (Condition 1 does not hold either way) but we can numerically verify that Case 2 patients are less sensi-

Table 4 Optimal Panel Sizes and Overbooking Levels under Cancellation

(N^*, μ^*)		Case 1	Case 2	Case 3	Case 4
$\eta = 0$	$\theta = 0.1$	(3170, 22.5)	(2990, 22.4)	(2820, 21.8)	(2550, 21.5)
	$\theta = 0.2$	(3400, 22.6)	(3230, 22.5)	(2910, 21.5)	(2660, 21.5)
	$\theta = 0.5$	(3590, 22.4)	(3590, 22.4)	(3290, 21.7)	(2890, 21.6)
$\eta = 0.001$	$\theta = 0.1$	(3310, 22.6)	(3080, 22.5)	(2860, 21.8)	(2580, 21.6)
	$\theta = 0.2$	(3510, 22.5)	(3290, 22.5)	(2970, 21.6)	(2700, 21.5)
	$\theta = 0.5$	(3820, 22.5)	(3600, 22.5)	(3300, 21.7)	(2900, 21.5)
$\eta = 0.003$	$\theta = 0.1$	(3390, 22.5)	(3170, 22.5)	(2940, 21.8)	(2620, 21.4)
	$\theta = 0.2$	(3650, 22.5)	(3360, 22.5)	(3020, 21.9)	(2710, 21.4)
	$\theta = 0.5$	(3870, 22.5)	(3680, 22.5)	(3430, 21.8)	(2920, 21.3)

N^* denotes the optimal panel size, and μ^* represents the optimal overbooking level. The results are derived by assuming $\kappa = \infty$ and $\xi = 0$.

tive than Case 3 patients at least when delays are relatively short, specifically when there are fewer than 50 appointments in the queue, which is the case a large proportion of the time under the optimal decisions. Looking at Table 4, we can observe that the optimal panel size and the optimal overbooking level are larger when the patients are less sensitive to appointment delays. Thus, this study suggests that our analytical results continue to hold even with patient cancellation and balking.

4.3. Estimating the "Optimal" Panel Size and Comparison with Open Access

Our mathematical model is a stylized representation of an appointment queue and is not meant to be used primarily as decision support tool to make precise decisions on the panel size and overbooking level. Nevertheless, it is still of interest to investigate what the model would suggest as the optimal panel size and how it would compare with panel sizes that are recommended for Open Access implementation using data from an actual clinic.

To that end, in this section, we estimate the "throughput maximizing" panel size using our model and compare our numbers with those suggested by Green and Savin (2008) for Open Access implementation. To make a proper comparison, we use exactly the same data and the same no-show estimates used in Green and Savin (2008). Specifically, we let the individual appointment demand rate λ_0 to be 0.008 per day and the service rate to be 20 customers per day, i.e., $\mu = 20$. We use the following parametrical model for show-up probabilities, $p_j, j \in Z$:

$$p_j = 1 - (\gamma_{\max} - (\gamma_{\max} - \gamma_0)e^{-\lfloor j/\mu \rfloor / C})$$

where γ_0 is the minimum observed no-show rate, $\gamma_{\max}$ is the maximum observed no-show rate, and C is the no-show backlog sensitivity parameter. Using data from a magnetic resonance imaging (MRI) facility, Green and Savin (2008) estimated γ_0, $\gamma_{\max}$, and C to be 0.01, 0.31, and 50, respectively. Using these estimates, we find that the optimal panel size under the M/M/1 setup is 2459, while that under the M/D/1 setup is 2471. The M/M/1 model estimate simply uses $\bar{\lambda}_1$ obtained through Proposition 2; the M/D/1 model estimate is obtained by numerically maximizing $T(\lambda, 20)$ as defined in Equation (2) under the assumptions of Poisson arrivals and deterministic service times with the queue length truncated at 1000.

As we discussed before, the goal of Green and Savin (2008) is to estimate the ideal panel size for Open Access, i.e., the panel size that will keep the clinic "in balance" for Open Access implementation. Using four different models (M/D/1/K, M/M/1/K, and two simulation models) and four different desired values for the same-day appointment probability, they obtained 16 different panel size estimates ranging from 2205 to 2368. Our throughput maximizing panel size estimates are larger than what Green and Savin (2008) suggest for Open Access. This is not surprising mainly due to two reasons. First, having to provide same-day service forces the service provider to keep the panel sizes smaller. Second, our model assumes that no-show customers do not reschedule new appointments, while Green and Savin (2008) assumed that all no-show customers immediately scheduled a new appointment. Since rescheduling means more work load per customer, it leads to a smaller optimal panel size.

One question of interest is the secondary effects of using a throughput maximizing panel size. The clinic might be maximizing throughput, but how about the effect on the delays the patients experience? Do they wait for a long time, at least significantly longer than they would under the same-day scheduling policy? Table 5 clearly demonstrates the trade-off among system throughput, customer delays, and the fraction of customers who can book appointments within a day or two. To be able to make a more meaningful comparison, following Green and Savin (2008), we truncate the queue length at 400 here, i.e., we assume an M/M/1/K queue with $K = 400$ instead of an M/M/1 queue. This does not affect the results in the first four rows in the table significantly; however, it is necessary for the last row since otherwise, the queue would have been unstable.

In Table 5, the maximum throughput is 19.194 with a panel size of 2460. In this case, the average appointment delay is 3 days (since the daily service rate is 20 patients) and approximately 28% of the patients can be seen on the day they call for an appointment. The clinic can improve the waiting times and same-day access probabilities by reducing the panel size. For example, if the panel size is 2220, approximately 90% of the customers can get same-day appointments, but the throughput drops to about 17.572. This is more

Table 5 Computed Values for the Key Performance Measures under the *M/M/*1 Model with No Rescheduling

Panel size	Throughput	$E(Q)$	$E(W)$	P_S	P_T
2220	17.572	7.929	0.396	0.907	0.991
2300	18.191	11.500	0.575	0.811	0.964
2380	18.783	19.833	0.992	0.626	0.860
2460	19.194	60.877	3.043	0.276	0.476
2540	18.134	338.191	16.860	0.000	0.002

$E(Q)$ is the average appointment queue length. $E(W)$ is the average patient appointment delay. P_S is the probability that an arriving appointment request is accommodated in the same day. P_T is the probability that an arriving appointment request is accommodated within the next 2 days.

than an 8% decrease, which suggests that committing to same-day appointments may not be optimal from a system efficiency point of view. If the clinic is interested in providing a high service level and throughput is a secondary concern, then same-day appointment scheduling could work well. If efficiency is more important, it might pay off to be more flexible. One can still stick with Open Access but perhaps implement it a little less strictly by promising customers appointments within 2 or 3 days as opposed to the same day.

5. Conclusion

This paper uses stylized models to investigate the relationships between patients' no-show behavior and the optimal panel size and overbooking decisions. Our results provide insights particularly for clinics interested in reducing patient no-shows by behavioral interventions such as sending reminders for appointments. These interventions typically improve patient attendance but cannot eliminate no-shows completely. In general, clinics prefer higher show-up probabilities, which not only mean less time wasted and more patients served, but also help clinics make better operational decisions because of the reduced uncertainty. What is far less clear is how clinics should alter their decisions in response to changes in patient show-up probabilities. Our findings suggest that responses based on one's intuition might not work as expected. For example, having patients who are more likely to show-up does not necessarily imply that the optimal panel size should be larger or smaller. What appears to be more important is whether patients are more or less sensitive to additional delays. Past empirical studies on the effectiveness of intervention policies have largely focused on changes in no-show rates (Guy et al. 2012, Macharia et al. 1992), but did not investigate the changes in the sensitivity of no-show probabilities to incremental changes in appointment delays. This is an important avenue for future research.

The generic nature of our formulation allows us to generate insights without restricting ourselves to any specific appointment scheduling policy, and our main finding, which says that panel size decisions should be informed by the sensitivity of the show-up probabilities to incremental changes in appointment delays is likely to hold under more general conditions and various appointment scheduling schemes. Nevertheless, our formulation is stylized and more research is needed to provide support for this claim. It is also important to note that how exactly one defines "sensitivity to incremental delays" may need to be reconsidered if one uses a more detailed formulation of an appointment system. Our Condition 1 in this paper, however, can help in identifying similar conditions in other formulations. Thus, one avenue for future work is to model clinics that use standard appointment scheduling policies in more detail (e.g., by explicitly considering day and time of the appointment, patient preferences, etc.) and investigate the connection between optimal panel sizes and show-up probabilities. Proving analytical results may not be possible, but investigation via a simulation study would likely be fruitful.

Acknowledgments

The authors are grateful to the departmental editor, the senior editor, and the anonymous referees, whose comments and suggestions have helped improve this work. The work of the second author was supported by the National Science Foundation (Grant no CMMI-0927607 and CMMI-1234212).

Note

[1]The queue length at the appointment time serves as a proxy for the delay that the patient will experience. Note that there is no one-to-one relationship between the queue length at the time of an appointment request arrival and the expected delay until the appointment. In particular, the expected delay for any two patients who see the same number of appointments upon their request may be different in practice because of the times (e.g., nights and weekends) during which the clinic will be closed. However, for a clinic which sees the same number of patients everyday, the difference is guaranteed to be less than one workday under the assumption that the clinic is open all workdays.

References

Bean, A. G., J. Talaga. 1995. Predicting appointment breaking. *J. Health Care Market.* **15**(1): 29–34.

Browder, A. 1996. *Mathematical Analysis: An Introduction*. Springer Verlag, New York, NY.

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

Daggy, J., M. Lawley, D. Willis, D. Thayer, C. Suelzer, P. C. DeLaurentis, A. Turkcan, S. Chakraborty, L. Sands. 2010. Using no-show modeling to improve clinic performance. *Health Inform. J.* **16**(4): 246–259.

Diwas, K. C., N. Osadchiy. 2012. Matching supply with demand in an outpatient clinic. Working paper, Emory University, Goizueta Business School, Atlanta, GA.

Dreiher, J., M. Froimovici, Y. Bibi, D. A. Vardy, A. Cicurel, A. D. Cohen. 2008. Non-attendance in obstetrics and gynecology patients. *Gynecol. Obstet. Invest.* **66**(1): 40–43.

Gallucci, G., W. Swartz, F. Hackerman. 2005. Brief reports: Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatr. Serv.* **56**(3): 344–346.

Fleck, E. 2012. Director, Internal Medicine, NewYork-Presbyterian Hospital/Ambulatory Care Network. Personal communication. *July 19, 2012*.

Geraghty, M., F. Glynn, M. Amin, J. Kinsella. 2007. Patient mobile telephone text reminder: A novel way to reduce non-attendance at the ENT out-patient clinic. *J. Laryngol. Otol.* **122**(3): 296–298.

Green, L. V., S. Savin. 2008. Reducing delays for medical appointments: A queueing approach. *Oper. Res.* **56**(6): 1526–1538.

Grunebaum, M., P. Luber, M. Callahan, A. C. Leon, M. Olfson, L. Portera. 1996. Predictors of missed appointments for psychiatric consultations in a primary care clinic. *Psychiatr. Serv.* **47**(8): 848–852.

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

Gupta, D., L. Wang. 2008. Revenue management for a primary-care clinic in the presence of patient choice. *Oper. Res.* **56**(3): 576–592.

Gupta, D., W. Y. Wang. 2012. Patient appointments in ambulatory care. R. Hall, ed. *Handbook of Healthcare System Scheduling, International Series in Operations Research and Management Science*, vol. 168. Springer, New York, NY, 65–104.

Guy, R., J. Hocking, H. Wand, S. Stott, H. Ali, J. Kaldor. 2012. How effective are short message service reminders at increasing clinic attendance? A meta-analysis and systematic review. *Health Serv. Res.* **47**(2): 614–632.

Hashim, M. J., P. Franks, K. Fiscella. 2001. Effectiveness of telephone reminders in improving rate of appointments kept at an outpatient clinic: A randomized controlled trial. *J. Am. Board Fam. Pract.* **14**(3): 193–196.

Jones, R. B., A. J. Hedley. 1988. Reducing non-attendance in an outpatient clinic. *Public Health* **102**(4): 385–391.

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

Kulkarni, V. G. 1995. *Modeling and Analysis of Stochastic Systems.* Chapman & Hall/CRC, Boca Raton, FL.

LaGanga, L. R., S. R. Lawrence. 2007. Clinic overbooking to improve patient access and increase provider productivity. *Decis. Sci.* **38**(2): 251–276.

Liu, N., S. Ziya, V. G. Kulkarni. 2010. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manuf. Serv. Oper. Manag.* **12**(2): 347–364.

Macharia, W. M., G. Leon, B. H. Rowe, B. J. Stephenson, R. B. Haynes. 1992. An overview of interventions to improve compliance with appointment keeping for medical services. *J. Am. Med. Assoc.* **267**(13): 1813–1817.

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

Norris, J. B., C. Kumar, S. Chand, H. Moskowitz, S. A. Shade, D. R. Willis. 2012. An empirical investigation into factors affecting patient cancellations and no-shows at outpatient clinics. *Decis. Support Syst.* **57**(1): 428–443.

Patrick, J., M. L. Puterman, M. Queyranne. 2008. Dynamic multi-priority patient scheduling for a diagnostic resource. *Oper. Res.* **56**(6): 1507–1525.

Pesata, V., G. Pallija, A. A. Webb. 1999. A descriptive study of missed appointments: Families' perceptions of barriers to care. *J. Pediatr. Health Care* **13**(4): 178–182.

Schütz, H.-J., R. Kolisch. 2012. Approximate dynamic programming for capacity allocation in the service industry. *Eur. J. Oper. Res.* **218**(1): 239–250.

Shonick, W., B. W. Klein. 1977. An approach to reducing the adverse effects of broken appointments in primary care systems: Development of a decision rule based on estimated conditional probabilities. *Med. Care* **15**(5): 419–429.

Ulmer, T., C. Troxler. 2004. The Economic Cost of Missed Appointments and the Open Access System. *Community Health Scholars*. University of Florida, Gainsville, FL.

Wang, W. Y., D. Gupta. 2011. Adaptive appointment systems with patient preferences. *Manuf. Serv. Oper. Manag.* **13**(3): 373–389.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix: Proofs of the Results.