M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 12, No. 2, Spring 2010, pp. 330–346
ISSN 1523-4614 | EISSN 1526-5498 | 10 | 1202 | 0330

informs®
DOI 10.1287/msom.1090.0270
© 2010 INFORMS

A Comparison of Traditional and Open-Access Policies for Appointment Scheduling

Lawrence W. Robinson
Johnson Graduate School of Management, Cornell University, Ithaca, New York 14853, lwr2@cornell.edu

Rachel R. Chen
Graduate School of Management, University of California at Davis, Davis, California 95616, rachen@ucdavis.edu

This paper compares two types of appointment-scheduling policies for single providers: traditional and open-access. Under traditional scheduling, each of a specified number of patients per day is booked well in advance, but may not show up for his or her appointment. Under open-access scheduling, a random number of patients call in the morning to make an appointment for that same day. Thus the number of patient arrivals will be random, for different reasons, under both policies. We find that the open-access schedule will significantly outperform the traditional schedule—in terms of a weighted average of patients' waiting time, the doctor's idle time, and the doctor's overtime—except when patient waiting time is held in little regard or when the probability of no-shows is quite small.

Key words: service operations; health-care management
History: Received: June 10, 2008; accepted: June 4, 2009. Published online in *Articles in Advance* October 2, 2009.

1. Introduction

Facing rising costs and an aging population in need of more services, the health-care system is under increasing pressure to make effective use of its existing capacity. It has long been standard practice for a patient to make a routine appointment months ahead of time. Although this type of appointment policy does allow the doctor's workday to be fully booked, there is an attendant downside: in many cases there is a significant probability that a patient will not actually show up for his or her appointment. To mitigate the effects of these no-shows, the doctor may resort to overbooking appointment time slots, which in turn will increase patient waiting. The so-called "open-access" policy is a recently introduced alternative to this traditional appointment-scheduling policy that seeks to avoid this trade-off between doctor idle time and patient waiting time altogether. Each patient calls into the doctor's office in the morning to make an appointment for that same day. This virtually eliminates no-shows, albeit at the expense of introducing considerably more variability (and overtime) in the day-to-day workload.

In this paper we compare the performance of these two appointment policies, which can be divided into two components. The first is the time that patients spend waiting for the doctor, and the second is the length of the doctor's day, including an overtime surcharge (e.g., 50%) whenever it exceeds the standard length. We consider here a single-provider service system under which each patient arrives (or not) for a previously scheduled appointment. The paradigm for this system would be a general practitioner's office, and for the sake of concreteness, we will refer to the service provider as the "doctor" and his or her customers as "patients" throughout this paper. However, this same scheduling system would be encountered by other service professionals, including attorneys, tax accountants, personal trainers, and barbers.

We consider first a "traditional" scheduling policy, under which patients are scheduled well in advance to fill each day. One of the most costly sources of uncertainty is whether the patient will actually arrive for his or her appointment. When the appointment is scheduled far ahead of time, the patient may simply forget about it or may have had a conflict arise in the meantime. Attempts to reduce no-shows by reminding patients of their appointments—through postcards, telephone calls, or e-mail—have reduced but not eliminated the problem. For example, O'Brien and Lazebnik (1998) show that reminder telephone calls within 24 hours of a patient's appointment at the Adolescent Clinic of Cleveland reduced the no-show rate from 56.6% to 34.8%—a sizeable improvement but by no means a complete solution. Similar to overbooking an airplane, the doctor may decide to double-book (or even triple-book) some time slots in anticipation of patients not showing up. Although this approach will certainly reduce the doctor's idle time, it can be quite costly in terms of waiting time for the patients when they do in fact show up.

We then consider the "open-access" scheduling policy (also known as a "same-day" or "advanced-access" policy), under which patients call in for an appointment on the morning of the day that they want to be seen. The first widely publicized implementation of open-access scheduling was led by Mark Murray and Catherine Tantau in the early 1990s at Kaiser Permanente in northern Sacramento Valley, California (Murray and Tantau 2000).

Both the traditional and the open-access scheduling policies encounter substantial variability in the number of patients seen per day: traditional because of no-shows within the fixed number of appointments for the day, and open-access because of the varying number of patients who call in to be seen each day. The goals of this paper are to compare the effects of these two types of variability on the operational cost of the doctor's office and to identify conditions under which each policy will be preferred. In order to keep the focus on this comparison, as well as for reasons of tractability, our model consciously does not include other sources of variability (e.g., variability in service times, patient tardiness, or unscheduled emergency interruptions) that would still be present. In that regard, our research is similar to the existing literature, which has also largely focused on a single source of variability (service times) while—with very few exceptions—omitting these other types of variability (no-shows, etc.). Of the literature that does incorporate patient no-shows, most assumes that the service times are exponentially distributed. In this paper we choose to focus on the variability in the number of patients arriving per day, and consider service times to be deterministic. It is unclear if or how variability in service times—or any other source of variability—would differentially affect the two policies. Investigating the effects of incorporating additional kinds of variability is beyond the scope of this current paper.

To our knowledge, this is the first paper that directly compares the performances of traditional and open-access scheduling policies under their respective sources of variabilities. It is also the first to find the optimal overbooking policy for the traditional scheduling system with no-shows and deterministic service times. For open-access scheduling, we examine two types of policies, depending on whether all patients must be seen the day they call in (the same-day policy) or whether some patients are willing to wait to be seen the following day (the "same-or-next-day" policy). In practice, most physicians will use a combination of both same-day and traditional policies; we examine such "hybrid" policies in §5.

One of our major findings is that the open-access schedule outperforms the traditional schedule in the wide majority of cases. The traditional scheduling policy will be preferred only when the no-show probability is small (less than 5%) or the cost of patients' waiting is trivial relative to the cost of the doctor's time. The same-day policy will perform better for larger no-show probabilities, for larger workloads, and for smaller overtime surcharges. Also, the same-or-next-day scheduling policy is substantially less costly than the same-day policy only when the work day is approximately equal to the expected workload; otherwise, the option of deferring service makes little difference. In addition, we show that moving from a traditional schedule to an open-access schedule will allow the physician to increase the panel size (the number of patients on his or her books) by up to 30% when patient waiting is especially costly.

We further observe that whenever the optimal traditional schedule calls for any overbooking, the first time slot is one of the ones overbooked. In particular, we prove that if exactly one patient is overbooked, then it is optimal to double-book him or her into the first time slot. Finally, we examine the same-day policy when the length of the standard day keeps pace with the expected workload, and show that as the workload becomes infinite, the limiting cost per patient becomes inversely proportional to the square root of the number of patients.

The outline of the paper is as follows. Following a review of the relevant literature in §2, we introduce the notation and model formulations for both the traditional and open-access policies in §3. In §4, we present our comparative numerical results for a moderate-sized example. In §5 we extend our analysis to develop bounds on the performance of a hybrid scheduling policy that handles a combination of previously scheduled and same-day patients. Section 6 summarizes our general conclusions and identifies possible directions for future research.

2. Literature Review

In the delivery of health-care services, variability in the patient arrival and service processes can cause excessive patient waiting times and poor utilization of the facility's resources. The impact of variability on the performance of health-care organizations has been studied since the 1950s. Some analytical studies apply queueing theory to obtain the steady-state behavior of the queue (Mercer 1960, Jansson 1966, Soriano 1966), which is representative only of facilities that never close. For transient queues, which better represent facilities that have a fixed length of day, the problem of setting optimal appointment times that balance the patients' waiting time against the doctor's idle time is analytically thorny and largely unsolved. The majority of the literature focuses on the variability in the service time while assuming that patients are strictly punctual (Weiss 1990; Pegden and Rosenshine 1990; Wang 1993, 1997; Denton and Gupta 2003; Robinson and Chen 2003). Chen (2007) is among the few papers that analyze the effects of the variability of patients' arrivals on the optimal scheduling policy. All of these studies acknowledge—but do not model—the possibility of a patient not showing up for his or her appointment.

Several empirical studies demonstrate that patient no-shows are widely prevalent in health-care systems and can cause substantial financial loss to primary care facilities (Pesata et al. 1999, Moore et al. 2001). Additionally, the probability of missing an appointment has been found to depend strongly on the type of health care service offered, as well as on characteristics of the patient population (Hixon et al. 1999, Moore et al. 2001, Xakellis and Bennett 2001, Ulmer and Troxler 2006). For example, in a recent study of a primary health-care clinic in a New York metropolitan hospital, Cayirli et al. (2006) report that the average no-show probability is 38% but varies between 0% for colonoscopies to 67% for pediatric neurology. Similarly, in a survey of 200 pediatric resident continuity clinics, Rust et al. (1995) report no-show probabilities ranging between 3% and 80%, with an average of 30.9%. Galucci et al. (2005) report that the probability of no-shows increases with the appointment delays.

The large impact of no-shows on the appointment scheduling policy is affirmed in several simulation studies that allow for general complexities in appointment systems while investigating the effects of varying the service time mean and variability (Cayirli et al. 2006; Ho and Lau 1992, 1997). A few analytical papers incorporate no-shows into queueing models of health-care systems. In the earliest studies, Mercer (1960, 1973) considers a queueing system in which patients may arrive late or not at all. Green and Savin (2008) consider that the probability of a no-show is nondecreasing in the size of the appointment queue at the time when the patient makes an appointment and are the first to incorporate such a relationship in their study of a single-server queueing system. Kaandorp and Koole (2007) develop an algorithm to search for the optimal appointment times in the face of patient no-shows, assuming that the service times are exponential. They restrict appointment times to a set of discrete points. Zeng et al. (2009) extend this model to include heterogeneous no-show rates. A similar problem, in continuous time, is considered by Hassin and Mendel (2008). In contrast to the literature that considers static scheduling, under which the complete set of patients to be scheduled is known before scheduling decisions are made, Muthuraman and Lawley (2008) consider a policy that assigns an appointment time to each patient when he or she calls in.

Another stream of research considers policies that mitigate the effects of no-shows. Fetter and Thompson (1966) investigate how walk-ins can offset no-shows, although they do not consider overbooking. Vissers and Wijngaard (1979) and Vissers (1979) allow the mean and variance of service times to be adjusted to compensate for no-shows and walk-ins. LaGanga and Lawrence (2007b) use simulation to analyze how overbooking improves the doctor's utilization under no-shows and increases profits through scheduling more patients. Our model of the traditional appointment scheduling under no-shows is similar to that of LaGanga and Lawrence (2007a), who also assume deterministic service times. Whereas they develop a heuristic appointment schedule, we instead characterize properties of the optimal schedule that enable us to find it efficiently, through enumeration.

Ever since Murray proposed open-access scheduling in the early 1990s, its effectiveness has been the subject of considerable debate (Murray and Tantau 2000, Murray and Berwick 2003, Shuster 2003). However, these articles—whether in favor of or opposed to open-access—tend to be based on anecdotal evidence and qualitative arguments rather than on rigorous quantitative analysis. The paper by Dobson et al. (2006) is the first to analytically examine a variant of the open-access appointment schedule. Their "carve-out" scheduling approach reserves a portion of the daily capacity for urgent cases. They show that when capacity is balanced with demand, their open-access appointment policy results in minor delays for routine cases and avoids excessive outsourcing of urgent cases. Qu et al. (2007) present an analytical method to determine the percentage of time slots that should be left open in a carve-out system in order to maximize the expected number of patients seen, when the routine patients have a higher no-show probability than the urgent patients. Green et al. (2007) consider what constitutes an appropriate balance between capacity and demand in an open-access system and propose a simple method in estimating the optimal panel size.

3. Notation and Model Development

In this section we develop the two models that represent the extremes of appointment scheduling. The first is the traditional scheduling policy, under which a given number of patients N are scheduled for each day. However, not everyone scheduled will in fact arrive: there is some given probability p that each patient will not show up. The second is the open-access policy, where the number of patients who request an appointment that day varies randomly, although each patient will always show up for his or her appointment. We will extend the same-day appointment model to allow patients to sometimes be deferred to the following day.

To focus on the effects of the no-show probability, we assume in this paper that other sources of variability are small enough to be safely ignored or—failing that—are at least roughly equal in their effect on the two types of schedules. Thus we assume, for example, that service times are constant, patients arrive promptly at their appointment times, and the doctor's time is never preempted for emergency patients. These assumptions are consistent with the simulation results of Ho and Lau (1992) and Cayirli et al. (2006), who both found that patient no-shows had a substantially larger effect on system performance than service time variability.

3.1. Traditional Appointment Scheduling

We first consider the traditional method of appointment scheduling, where each patient makes an appointment well ahead of time. The length of time between appointments is equal to the deterministic service time, although to counter no-shows multiple patients may be scheduled for the same time slot. Because of the long lead times typically involved, we assume that the doctor can smooth out the variability in patient demand, and so is able to schedule a specified number of patients, N, to be seen every day. (This may not be true for physicians who operate below capacity, but it should hold approximately when there is a long backlog to see the doctor.) We assume that N has been exogenously determined by the doctor and reflects his or her utility for income versus spare time. A second (and less realistic) assumption of the traditional model is that there are no unscheduled walk-in patients. Although this assumption may hold perfectly only for nonurgent specializations (e.g., a photo studio or a dermatology practice), it allows us to analyze the diametric opposite of the open-access scheduling policy. We briefly discuss a more realistic (albeit intractable) intermediate hybrid policy in §5, which allows for both scheduled and same-day appointments. We define

$T =$ number of time slots for appointments within a standard (e.g., eight-hour) day;

$N =$ number of patients scheduled for each day (exogenously determined);

$p =$ probability that any given patient will not show up for his or her appointment;

$\bar{n} = (1-p)N =$ expected workload of patients, per day;

$x_t =$ number of patients scheduled for time slot t;

$t_{\max} = \max\{t \mid x_t \geq 1\}$ = time slot scheduled for the final patient N;

$b(k \mid z, \phi)$ = probability mass function (p.m.f.) for the binomial distribution, with extended domain

$$= \begin{cases} \binom{z}{k}\phi^k(1-\phi)^{z-k} & \text{if } 0 \leq k \leq z, \\ 0 & \text{otherwise;} \end{cases}$$

$\pi_t(k)$ = probability that there will be k patients in the office (system) during time slot t;

z_t = maximum possible number of patients in the office (system) during time slot t $= (z_{t-1}-1)^+ + x_t$, with $z_0 = 0$.

Note that in situations where a lunch break essentially decouples morning and afternoon appointments, we can consider T to be the number of appointments in either half of the day.

We can define the state probabilities recursively as

$$\pi_t(k) = b(k \mid x_t, 1-p)\pi_{t-1}(0) + \sum_{j=(k+1-z_{t-1})^+}^{\min\{x_t,\, k\}} b(j \mid x_t, 1-p)\pi_{t-1}(k+1-j) \quad (1)$$

for $k = 0, \dots, z_t$ and $t = 1, \dots, t_{\max}$, with $\pi_0(0) = 1$ and $\pi_0(k) = 1$ for $k \geq 1$. The first term is the probability that the system was empty in the previous period and k patients arrived this period. Otherwise, there was a patient who was served and departed in the previous period. To arrive at the limits of summation of the second term, note that the number of arrivals this period is limited as $0 \leq j \leq x_t$, while the number of patients in the system in the previous period is limited as $1 \leq k+1-j \leq z_{t-1}$, or $k+1-z_{t-1} \leq j \leq k$. The limits of summation are the intersection of these two ranges.

To describe the daily cost of a traditional scheduling policy, we define

α = the cost of patient waiting, as a fraction of the cost of the doctor's idle time;

β = the overtime premium for the doctor; $\beta = 50\%$ by default;

$\bar{D}$ = the doctor's expected day length, i.e., the doctor's expected completion time;

$\bar{I}$ = the doctor's expected idle time before the end of the day;

$\bar{O}$ = the doctor's expected overtime;

$\bar{W}$ = the expected total waiting time of the patients;

$C^T(\{x_t\})$ = the total cost of a traditional policy, given schedule $\{x_t\}$;

where the superscript "T" designates the traditional scheduling policy.

For any traditional schedule $\{x_t\}$, the total cost C^T is the sum of three components. The first is the expected length of the doctor's day. We assume that the doctor can and does leave after seeing all the patients who showed up. Note that he or she must remain through the end of time slot $t_{\max} - 1$ in any case, in order to observe whether the patient(s) scheduled for the final time slot $t_{\max}$ actually arrive. The additional time required to see all of the patients will be determined by the distribution of the number in the system for the final time slot $t_{\max}$, so that the expected length of the doctor's day will be

$$\bar{D}^T = t_{\max} - 1 + \sum_{k=0}^{z_{t_{\max}}} k \cdot \pi_{t_{\max}}(k).$$

Within the scheduling literature, the idle time is consistently defined to be the difference between the day length and the workload (Mondschein and Weintraub 2003). Because the expected workload $\bar{n}$ is a constant, we can subtract it from $\bar{D}^T$ to consider equivalently the doctor's expected idle time:

$$\bar{I}^T = \bar{D}^T - \bar{n} = t_{\max} - 1 + \sum_{k=0}^{z_{t_{\max}}} k \cdot \pi_{t_{\max}}(k) - \bar{n}. \quad (2)$$

The second component of the cost is a surcharge β on any overtime incurred. Overtime is calculated as the positive difference between the actual server completion time (D) and the day length (T). We follow the literature (see Cayirli et al. 2006) by assuming as the default that overtime costs are 150% of standard. This is equivalent to an incremental surcharge of $\beta = 50\%$, because the idle time already incurs a 100% penalty on overtime (the expected day length $\bar{D}^T$ includes both overtime and regular time, and $\bar{I}^T = \bar{D}^T - \bar{n}$). The expected overtime, $E_D(D-T)^+$, depends on both $t_{\max}$ and T as

$$\bar{O}^T = \begin{cases} 0 & \text{if } t_{\max} \leq T - z_{t_{\max}}, \\ \displaystyle\sum_{k=1}^{t_{\max}+z_{\max}-T-1} k \cdot \pi_{t_{\max}}(k+T+1-t_{\max}) & \\ & \text{if } T - z_{t_{\max}} + 1 \leq t_{\max} \leq T, \\ \bar{D}^T - T & \text{if } t_{\max} \geq T+1. \end{cases} \quad (3)$$

The final component reflects the goodwill cost of patient waiting. Without this component, the optimal policy would be to trivially schedule everyone for the start of the day (i.e., set $x_1 = N$), which would eliminate all idle time and minimize the doctor's overtime. Note that in the final period $t_{\max}$, each patient j in the system must wait an additional $(j-1)$ periods before being served. Thus the total patient waiting time will be

$$\begin{aligned}\overline{W}^T &= \sum_{t=1}^{t_{\max}-1} \sum_{k=1}^{z_t} (k-1)\cdot \pi_t(k) + \sum_{k=1}^{z_{t_{\max}}} \left(\sum_{j=1}^{k}(j-1)\right)\cdot \pi_{t_{\max}}(k)\\ &= \sum_{t=1}^{t_{\max}} \sum_{k=1}^{z_t} (k-1)\cdot \pi_t(k) + \tfrac{1}{2}\sum_{k=1}^{z_{t_{\max}}}(k-2)\\ &\quad \cdot (k-1)\cdot \pi_{t_{\max}}(k). \end{aligned} \tag{4}$$

We consider the cost of patients' waiting to be a given fraction α of the cost of the doctor's idle time. We expect, but do not require, $\alpha \le 1$. Thus the total cost for any given traditional schedule can be found from (2), (3), and (4) to be

$$C^T(\{x_t\}) \doteq \bar{I}^T + \alpha \overline{W}^T + \beta \bar{O}^T. \tag{5}$$

This objective function includes the cost of the doctor's idle time (standardized at 1), the patients' waiting time valued at α, and an overtime surcharge β on any time beyond the day length T. As reviewed in Mondschein and Weintraub (2003), the standard objective function for service-scheduling systems is a weighted average of expected total customer waiting time and server completion time (e.g., Pegden and Rosenshine 1990; Wang 1993, 1997). Because the server completion time is equal to the sum of customer service time (constant in expectation) and the doctor's total idle time, we can equivalently substitute idle time for the service completion time in this objective function (e.g., Jansson 1966; Vissers and Wijngaard 1979; Weiss 1990; Ho and Lau 1992, 1997; Hassin and Mendel 2008). Several studies (e.g., Denton and Gupta 2003, Kaandorp and Koole 2007, see Cayirli and Veral 2003 for others) also include an explicit surcharge of overtime. A few other studies (LaGanga and Lawrence 2007a, b; Muthuraman and Lawley 2008) maximize profits instead of minimizing costs.

We assume that the number of patients scheduled per day, N, has been exogenously determined, so revenues will be proportional to $\bar{n} = (1-p)N$, and profit maximization will be equivalent to the cost minimization used here and in the bulk of the service-scheduling literature. In §4.2 we will briefly discuss how our results fit into the context of profit maximization when the expected daily workload $\bar{n}$ (and so revenues) is endogenous to the problem.

As previously mentioned, we assume that the doctor can leave after seeing the final scheduled patient (or on learning that he or she has not shown up); the doctor may sometimes leave before the end of the day. Denton and Gupta (2003) define *earliness* to be the positive difference between the day length (T) and the actual completion time of the last patient (D). There is typically no additional cost or benefit from the expected earliness $E_D(T-D)^+$, but since it can be expressed in terms of the expected day length $\bar{D}$ and expected overtime $\bar{O}$ (and constants), we note in passing that our model can easily be extended to explicitly include earliness by means of parameter redefinition. In §4.3.6, we will consider an alternative "no-golf" scheduling policy, under which the doctor must stay until the end of the session (time slot T) even if no more patients are scheduled to arrive. That policy essentially considers earliness to be idle time.

The objective function (5) is to be minimized subject to the constraints

$$\sum_{t=1}^{N} x_t = N \tag{6}$$

$$\sum_{t=1}^{k} x_t \ge k \quad k = 1, \ldots, N-1, \quad \text{with} \tag{7}$$

$$x_t \ge 0 \quad \text{and integer.} \tag{8}$$

Constraint (7) requires at least k patients to be scheduled in the first k time slots, and ensures that no time slot is necessarily idle, which would be suboptimal. (We could always move the later patients forward by one time slot, keeping their waiting times the same while reducing the doctor's idle time and overtime, if any.)

3.2. Properties and Bounds for the Optimal Traditional Scheduling Problem

One useful property of the optimal policy is that it contains no "holes"; i.e., if it is optimal to schedule a patient for a given time slot, then it will be optimal to schedule patients for every earlier time slot. This is shown in the following proposition.

Proposition 1. *Consider the optimal scheduling policy $\{x_t^*\}$ where $x_k^* \geq 1$ for some time slot k. Then $x_t^* \geq 1$ for $t = 1, \dots, k$.*

The proof for this and all subsequent propositions can be found in the online appendix.

This proposition drastically reduces the number of scheduling policies that need to be considered. For example, when $N = 16$ the formulation given by (6)–(8) will have 35,357,670 feasible schedules. But under Proposition 1, we will always assign the first patient to the first time slot of the day; each subsequent patient can be assigned either to only two time slots, conditional on previous assignments, without introducing a "hole" in the schedule. In other words, each patient can be assigned to the same time slot as the previous patient or to the following time slot. Thus for $N = 16$ there are now only $2^{N-1} = 32{,}768$ different schedules that need to be considered. This reduction of 99.91% yields a set of possible schedules that is small enough that enumeration is a reasonable solution technique.

The cost of any specific policy will of course provide an upper bound on the cost of the optimal policy. Consider the single booking policy, which assigns each patient to his or her own time slot: $x_t = 1$ for $t \leq N$, with $t_{\max} = N$. Because this policy eliminates patient waiting, it will be optimal for sufficiently large values of α. Under this policy, the doctor must stay at least through the end of period $N-1$ to see if the final patient arrives; thus the expected length of the doctor's day is $N - p$. For each of the first $N-1$ time slots there is a probability of p that the scheduled patient does not show up, so that the doctor's expected idle time will be

$$\bar{I}^{T:\mathrm{UB}} = p(N-1).$$

The expected overtime depends on whether the number of time slots in a day T exceeds the $N-1$ periods, so that

$$\bar{O}^{T:\mathrm{UB}} = \begin{cases} N - p - T & \text{if } T \leq N-1, \\ 0 & \text{if } T \geq N; \end{cases}$$

and

$$C^{T:\mathrm{UB}} = I^{T:\mathrm{UB}} + \beta \bar{O}^{T:\mathrm{UB}} = N - p - \bar{n} + \beta(N-p-T)^{+}. \quad (9)$$

Finally, note that we can concisely express the range of α for which single booking is the optimal policy, depending on the relationship between T and N.

Proposition 2. *It is optimal to assign each patient to an individual time slot ($t_{\max}^* = N$) if and only if*

$$\alpha \geq \frac{p}{(1-p)^2} \cdot \begin{cases} 1+\beta & \textit{if } T \leq N-2, \\ 1 + (1-p)\beta & \textit{if } T = N-1, \\ 1 & \textit{if } T \geq N. \end{cases} \quad (10)$$

Note that the term to the right of the brace ranges between 1 and $(1+\beta)$, depending on the likelihood that overtime will be required.

Corollary 3 immediately follows from the proof of the sufficiency condition of Proposition 2 (see the online appendix for more details), and states that whenever it is optimal to overbook exactly one patient, then that patient should be overbooked in the first time slot. In our numerical results in §4.1, we observe that every optimal schedule that calls for some overbooking includes the first time slot as one of the ones overbooked.

Corollary 3. *If it is optimal to overbook one patient ($t_{\max}^* = N-1$), then $x_1^* = 2$ and $x_t^* = 1$ for $t = 2, \dots, N-1$.*

3.3. Open-Access Scheduling

In order to fairly compare open-access and traditional scheduling policies, we specify that the expected daily workload $\bar{n}$ be the same for both policies. Under open-access scheduling, a random number of patients s call each morning to make an appointment to be seen that day. We assume that the panel size is quite large and that each patient will independently call in for an appointment on any given day with the same small probability; thus the number of patients s requesting appointments will follow the binomial distribution. However, this setting (a large number of independent trials with a small probability of success) is precisely where the binomial distribution converges to the Poisson, which we will use in our analysis. Under Poisson arrivals, the variance of the daily workload will be equal to its mean, $\bar{n}$; this can be considerably larger than the variance of the daily workload under the traditional policy with binomial arrivals, $p(1-p)N = p\bar{n}$.

We assume that under open-access scheduling, each of these patients will show up for his or her appointment. In other words, the forgetfulness and/or schedule conflicts that resulted in no-shows under the traditional policy will no longer apply under an open-access policy when patients called in that morning. Steinbauer et al. (2006) report that most practices that have implemented open-access scheduling have reduced their no-show rates to near zero.

To facilitate our analysis, we assume here that the no-show rate is zero for open-access scheduling. The optimal patient-sequencing problem is then trivial: assign patients sequentially to time slots, starting with the earliest in the day. This eliminates both idle time and patient waiting time; α becomes irrelevant. Thus, the randomness in the demand arrivals will affect the system through overtime only.

We examine two types of open-access scheduling policies, depending on whether all patients must be seen the day that they call (the same-day policy) or whether some patients may be willing to wait to be seen the following day (the same-or-next-day policy). The incremental notation for open-access scheduling policies is summarized below.

$s =$ number of patients who call in the morning;

$p(s \mid \bar{n}) =$ p.m.f. of the Poisson distribution, with mean $\bar{n}$;

$P(s \mid \bar{n}) =$ cumulative distribution function (c.d.f.) of the Poisson distribution, with mean $\bar{n}$;

$\bar{P}(s \mid \bar{n}) = 1 - P(s \mid \bar{n})$;

$d =$ the maximum number of patients who can be deferred to the following day;

$q_i =$ probability of starting a day with i patients deferred from the previous day ($i = 0, \ldots, d$);

$C^{\mathrm{SD}} =$ the expected cost under same-day scheduling;

$C^{\mathrm{SND}}(d) =$ the expected cost under same-or-next-day scheduling if up to d patients can be deferred.

Please be aware that the notation for the Poisson p.m.f. $p(\cdot \mid \bar{n})$, is unavoidably similar to the probability of no-show, p.

3.3.1. Same-Day Scheduling.

Although there will be neither idle time nor patient waiting time under same-day scheduling, the doctor's overtime cost may be substantial whenever a large number of patients happen to arrive that day. We can express the doctor's expected overtime as

$$\begin{aligned}\bar{O}^{\mathrm{SD}} &= \sum_{s=T+1}^{\infty} (s-T)p(s \mid \bar{n}) \\ &= \bar{n} \cdot \bar{P}(T-1 \mid \bar{n}) - T \cdot \bar{P}(T \mid \bar{n}) \qquad (11)\end{aligned}$$

for the Poisson distribution, yielding an expected daily cost of

$$C^{\mathrm{SD}} \doteq \beta \cdot \bar{O}^{\mathrm{SD}}. \qquad (12)$$

3.3.2. Same-or-Next-Day Scheduling.

In some scenarios it might be permissible, when demand is unusually high, to defer some patients' appointments to the following day. For example, Murray and Berwick (2003) discuss open-access policies that reserve time slots for patients who need to be seen that day or the next. Let d represent the maximum number of patients who can be deferred to the following day; we need consider only $d \leq T$, because there is no benefit to deferring patients who would be seen during overtime on either day. We can represent this extension as a Markov chain; define q_i to be the probability of starting a day with i ($i = 0, \ldots, d$) patients deferred from the previous day. The transition equations can be written as

$$q_0 = \sum_{j=0}^{d} q_j \cdot P(T-j \mid \bar{n}), \qquad (13)$$

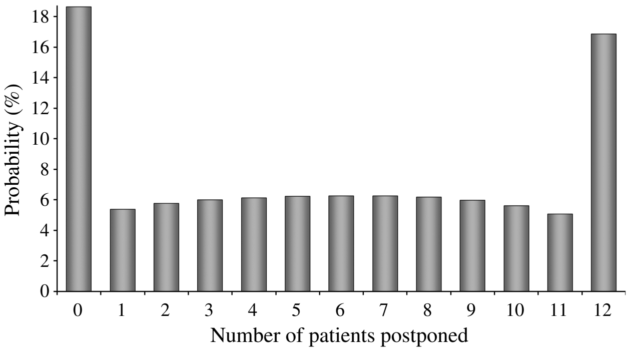
$$q_i = \sum_{j=0}^{d} q_j \cdot p(T+i-j \mid \bar{n}) \quad i = 1, \ldots, d-1, \qquad (14)$$

$$q_d = \sum_{j=0}^{d} q_j \cdot \bar{P}(T+d-j-1 \mid \bar{n}), \quad \text{with} \qquad (15)$$

$$\sum_{j=0}^{d} q_j = 1. \qquad (16)$$

Simultaneously solving the $d+1$ equations given by (14), (16), and either (13) or (15) yields the steady-state probabilities $\{q_i\}$. As can be seen in Equations (13)–(15), q_0 and q_d are both defined as the sum of the products of q_j and the Poisson c.d.f. (or its complement), whereas each interior probability q_i is the sum of the products of q_j with the Poisson p.m.f. only. Roughly speaking, there are more combinations

Figure 1 **Distribution of the Number of Patients Postponed Until the Following Day** ($d = \bar{n} = T = 12$)



of outcomes that lead to 0 or d patients being postponed to the next day, so that the end probabilities q_0 and q_d are significantly higher than the interior probabilities, as shown in Figure 1 for the example $d = \bar{n} = T = 12$.

We can now calculate the distribution for the number of patients seen each day; the probability that k patients are seen in a day is

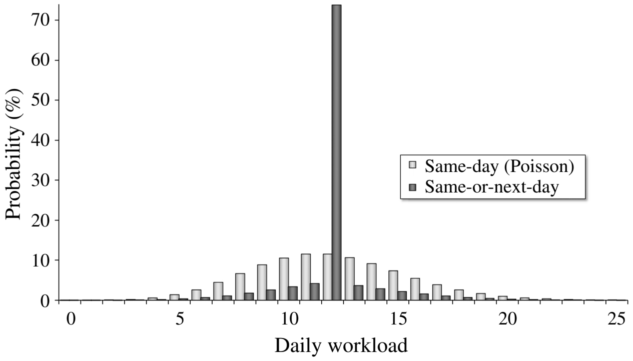
$$\begin{cases} \displaystyle\sum_{j=0}^{\min\{d,k\}} q_j \cdot p(k-j \mid \bar{n}) & \text{for } k \le T-1, \\ \displaystyle\sum_{j=0}^{d} q_j \cdot [P(d+T-j \mid \bar{n}) - P(T-j-1 \mid \bar{n})] & \\ & \text{for } k = T, \\ \displaystyle\sum_{j=0}^{d} q_j \cdot p(d+k-j \mid \bar{n}) & \text{for } k \ge T+1. \end{cases} \quad (17)$$

The ability to defer some patients to the following day significantly reduces the variability of the daily workload. For example, if $\bar{n} = T = 12$, the distribution of the number of patients seen per day is given in Figure 2 both for $d = 12$ (maximum deferrals) and $d = 0$ (no deferrals—the original Poisson distribution). The standard deviation of the number of patients seen per day drops by half, from $\sqrt{12} = 3.464$ to 1.746, while the probability that the number of patients seen is precisely equal to the daily capacity T jumps dramatically, from 11.44% to 73.75%.

The expected overtime for this same-or-next-day policy is

$$\bar{O}^{\text{SND}}(d) = \sum_{k=T+1}^{\infty} (k-T) \cdot \left(\sum_{j=0}^{d} q_j \cdot p(d+k-j \mid \bar{n}) \right); \quad (18)$$

Figure 2 **Distribution of the Number of Patients Seen in a Day** ($d = \{0, 12\}$; $\bar{n} = T = 12$)



so that

$$C^{\text{SND}}(d) = \beta \bar{O}^{\text{SND}}(d). \quad (19)$$

The advantage of the same-or-next-day policy over the same-day policy is seen to be largest when $\bar{n} \approx T$, as will be shown in §4.3.3.

4. Numerical Analysis

Rather than attempt to exhaustively analyze the universe of parameter combinations, in this section we will instead use marginal analysis to explore the effects of parameter changes within a specific, moderate-sized base problem. In particular, our base problem sets $T = \bar{n} = 12$, which could correspond to filling (in expectation) 12 40-minute time slots over an 8-hour working day. We assume that the no-show probability $p = 25\%$, so that the number of patients scheduled per day is $N = \bar{n}/(1-p) = 16$. We examine 101 values of α ranging between 0.01 and 1, with $\alpha_k \doteq 10^{-k/50}$ for $k = 0, \ldots, 100$. Except in §4.3.4, we assume that the overtime surcharge β will equal its default value of 50%.

4.1. Optimal Traditional Policy

Because of the difficulty of optimizing a nonlinear integer program, we instead used enumeration to find the optimal policy, using Proposition 1 to reduce the solution space to $2^{N-1} = 32{,}768$ different schedules. The optimal policies for 101 different values of α are shown in Figure 3, in which the shading of each time slot corresponds to the number of patients scheduled then. We see that the overbooked time slots are distributed evenly throughout the day, to balance patient waiting with the doctor's idle time. Triple and

Figure 3 **Optimal Assignment of Patients to Time Slots** ($\bar{n} = 12$, $T = 12$, $p = 25\%$, $N = 16$, $\beta = 0.5$)

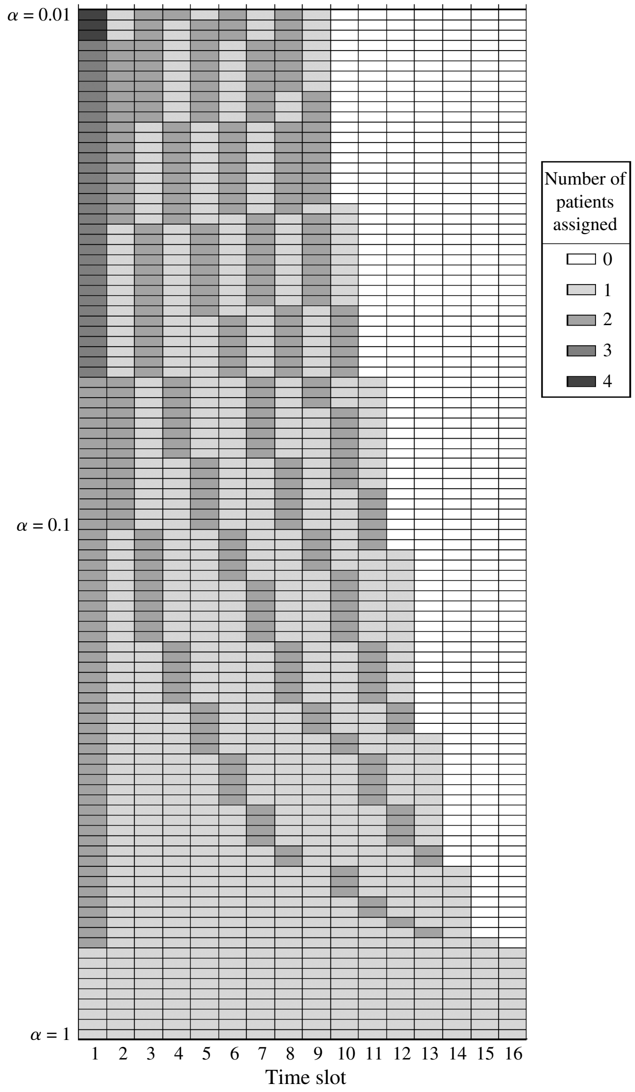
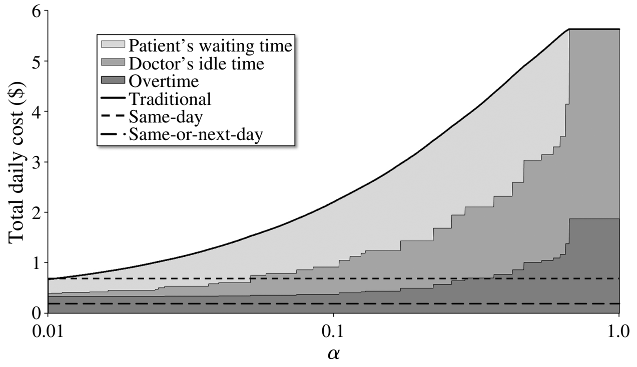


Figure 4 **Cost of the Optimal Policies** ($\bar{n} = 12$, $T = 12$, $p = 25\%$, $N = 16$, $\beta = 0.5$)



quadruple bookings are optimal only for the first time slot of the day, and then only for smaller values of α ($\alpha < 0.05$). Any optimal schedule with overbooking will overbook the first time slot. The number of time slots scheduled, $t_{\max}$, is seen to be increasing in the value of the patient's time α, as the amount of overbooking decreases. In particular, note that for $\alpha < 0.11$ it will be optimal to overbook some time slots while leaving other time slots unscheduled at the end of the day (i.e., $t_{\max} < T$). Although this overbooking will increase patient waiting time, it will also reduce the doctor's idle time (and the expected length of the day). Finally, note from Proposition 2 that the optimal policy will be to single-book all N patients for $\alpha \geq (1+\beta)p/(1-p)^2 = 2/3$, when $p = 25\%$ and $\beta = 0.5$.

The cost of the optimal policy is shown in Figure 4, and is a piecewise-linear concave increasing function of α. (It is not concave in $\ln(\alpha)$, which is how Figure 4 is constructed.) The kink at $\alpha = 2/3$, where the optimal policy becomes single booking ($t_{\max} = N$) and patient waiting (and therefore the effect of α) is eliminated, is clearly observable.

4.2. Optimal Open-Access Policies

Because of our assumption that every patient will always show up on the day when he or she schedules an appointment, we can simply schedule them sequentially as they call in, avoiding both patient waiting time (so that α is irrelevant) and the doctor's idle time; the only cost will be the expected overtime surcharge. For our base problem, the cost of the same-day policy from (11) and (12) becomes $C^{\text{SD}} = 0.6862$, which is also shown in Figure 4. It is clear that the same-day scheduling policy outperforms—substantially for larger values of α—the traditional scheduling policy in almost all cases. Only for the smallest values of α ($\alpha < 0.0105$), when patient waiting time is unimportant, will the traditional scheduling policy be very slightly preferable.

Furthermore, in situations where a patient is amenable to sometimes postponing his or her appointment to the following day, costs can be further reduced. For our base problem, we can use (18) and (19) to calculate $C^{\text{SND}}(12) = 0.1865$, which is also graphed in Figure 4. This policy is clearly dominant and reduces the cost of the same-day policy by 72.8%.

Even though the standard deviation of the daily workload is twice as high ($\sqrt{\bar{n}} = 3.464$ versus $\sqrt{p\bar{n}} = 1.732$), the same-day scheduling policy performs consistently better than the traditional policy—dramatically so when patient waiting is an important component of system performance. Furthermore, if patients can be deferred to the following day, much of the remaining costs are also eliminated. This somewhat counterintuitive result stems from the different effects that variability has in the two policies. Under the traditional scheduling policy, variation caused by no-shows affects performance throughout the day, causing idle time for the doctor and/or waiting time for the patients. By contrast, under the open-access scheduling policy, variation due to patient arrivals affects only the length of the day and the doctor's overtime; there is neither idle time for the doctor nor waiting time for the patients.

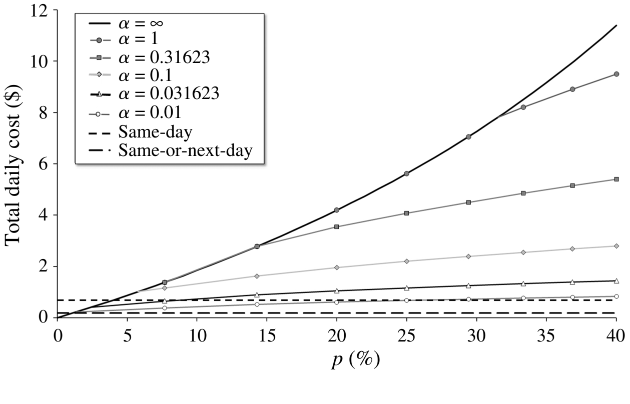
Until this point, we have compared the two policies when the expected workload $\bar{n}$ had been set exogenously in advance. That is, for a given $\bar{n}$ (and so given expected revenue), we have seen that the open-access scheduling policy is generally less costly than the traditional scheduling policy. We now consider whether this result will hold as well under the objective of profit maximization, where the expected workload $\bar{n}$ is endogenous. Suppose that the observed superiority of the open-access scheduling policy does hold for the particular value of $\bar{n}$ that maximizes the profits of the traditional scheduling policy. With revenues the same for both policies, the lower cost of the open-access policy will of course result in higher profits. By allowing the open-access policy to shift to the value of $\bar{n}$ that maximizes its profits, the open-access policy's superiority in profits over the traditional policy can only increase. Thus our analysis provides a conservative estimate of the advantage of the open-access policy under profit maximization.

4.3. Marginal Analysis

This section analyzes the effects of the various parameters on the performance of the two types of scheduling policies. Although our analysis examined all 101 different values of α, for graphical clarity we only present the results for five: $\alpha = 10^{-2}$, $10^{-1.5}$, 10^{-1}, $10^{-0.5}$, and 10^{0}. We will examine in turn the effect of the no-show probability p, the expected workload $\bar{n}$, the length of the day T, and the overtime surcharge β. We explore how much the panel size can be increased by moving to an open-access policy that has the same cost as the current traditional scheduling policy. Finally, we examine the effect of the alternative assumption that the doctor must remain until the end of the day, even when all patients have been seen.

4.3.1. Effects of the No-Show Probability p. Here, we let N, the number of patients scheduled per day, vary between 12 and 20 while keeping the expected workload $\bar{n}$ at 12, which generated values of the no-show probability $p = 1 - \bar{n}/N$ between 0% and 40%. We maintain the length of the day T at its base value of 12 and the overtime surcharge β at its default value of 0.5. The results displayed in Figure 5 show (unsurprisingly) that the cost of the traditional policy is increasing in p for all α. When $p = 0\%$, there will be no uncertainty—and no cost—for the traditional scheduling policy. The single-booking policy that eliminates patient waiting provides an upper bound and is optimal as $\alpha \to \infty$; its cost is convex in the no-show probability p. This single-booking policy will be optimal for small values of p and large values of α, as given in Proposition 2. When the optimal traditional scheduling policy employs some overbooking (for larger p and smaller α), costs will instead be concave in p. Note that the costs of the optimum same-day and same-or-next-day policies depend only on $\bar{n}$ (held constant at 12 here) and so are unaffected by p. The cost of the same-day scheduling policy remains lower than the cost of the traditional policy for all but the smallest values of α ($\alpha < 0.03$) as long as there is

Figure 5 **Effect of the No-Show Probability p on Schedule Cost** ($\bar{n} = 12$, $T = 12$, $N = \bar{n}/(1-p)$, $\beta = 0.5$)



at least a small chance of no-shows. As in our base model, when it is possible to postpone some open-access patients until the following day, the same-or-next-day policy will be dominant.

4.3.2. Effects of the Expected Workload $\bar{n}$. We let the expected workload $\bar{n}$ vary between 3 and 24, keeping the day length $T = \bar{n}$, the no-show probability $p = 25\%$, and the overtime surcharge $\beta = 0.5$, with $N = \bar{n}/(1-p)$. To facilitate meaningful comparisons, Figure 6 presents the cost per unit of expected workload: $C/\bar{n}$. Under the traditional scheduling policy, this ratio is increasing with $\bar{n}$ for the higher values of α, showing that total costs are increasing more than linearly with $\bar{n}$. However, for $\alpha \leq 0.018$ this ratio is actually decreasing, albeit very slightly, with $\bar{n}$. Essentially, the negative consequences of no-shows generally increase with the workload, as the day grows longer and the number of patients affected increases.

The total cost per unit of expected workload for the same-day and same-or-next-day policies is seen to decrease with $\bar{n}$. In fact, when $\bar{n} = T$ (as here), we can use Stirling's formula,

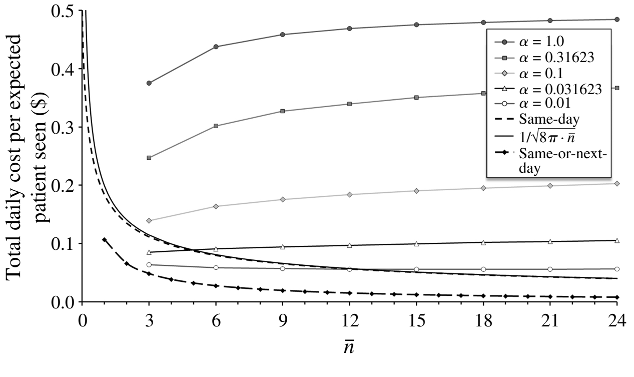
$$\lim_{\bar{n}\to\infty}\left(\frac{e^{\bar{n}}\bar{n}!}{\bar{n}^{\bar{n}}\sqrt{\bar{n}}}\right) = \sqrt{2\pi},$$

to show that

$$\lim_{\bar{n}\to\infty}\left(\frac{C^{\mathrm{SD}}}{\bar{n}}\right) = \frac{1}{\sqrt{8\pi\bar{n}}},$$

which will asymptotically approach zero. This limit is also graphed in Figure 6 and is quite close to $C^{\mathrm{SD}}/\bar{n}$ for $\bar{n} \geq 3$. Compared to the traditional scheduling policy, the same-day policy is preferable for medium and large workloads and number of patients (e.g., $\bar{n} \geq 9$, with $N \geq 12$) and all but the smallest weights α on the patient's waiting time. The same-or-next-day policy is, yet again, universally dominant.

Figure 6 **Effect of the Expected Workload $\bar{n}$ on Schedule Performance: Cost per Expected Patient Seen** ($T = \bar{n}$, $p = 25\%$, $N = \bar{n}/(1-p)$, $\beta = 0.5$)



4.3.3. Effects of the Length of the Day T. Here, we keep the expected workload $\bar{n} = 12$, the no-show probability $p = 25\%$, and the overtime surcharge $\beta = 0.5$, with $N = 16$ patients. We allow the length of the day T to vary between 7 and 17 time slots, to analyze its effect on policy performance. The costs are shown in Figure 7 and are decreasing in T as overtime is reduced. This decrease is initially linear as unavoidable overtime is eliminated, before leveling off at $T = N$, at which point overtime is never incurred. The theoretical lower bound of $\beta(\bar{n} - T)^+$ will hold in the absence of variability. Note that as long as the cost of patient waiting is not too high, the optimal policy will still employ overbooking to reduce the doctor's idle time by allowing him or her to leave early on occasion, even when $T \geq N$. However, if the doctor operates under the no-golf policy that we consider in §4.3.6, then he or she will need to stay until the end of the day T. Thus for $T \geq N$ the optimal policy will be to single-book all patients (incurring neither patient waiting nor overtime), for a daily cost of $T - \bar{n}$, regardless of α.

Both the same-day and same-or-next-day policies dominate the traditional policy for all α considered, for all T besides 12 or 13. In particular, the same-or-next-day policy is distinct from the theoretical lower bound only when T is close to $\bar{n}$ (12). The ability to

Figure 7 **Effect of the Workday Length T on Schedule Cost** ($\bar{n} = 12$, $p = 25\%$, $N = 16$, $\beta = 0.5$)

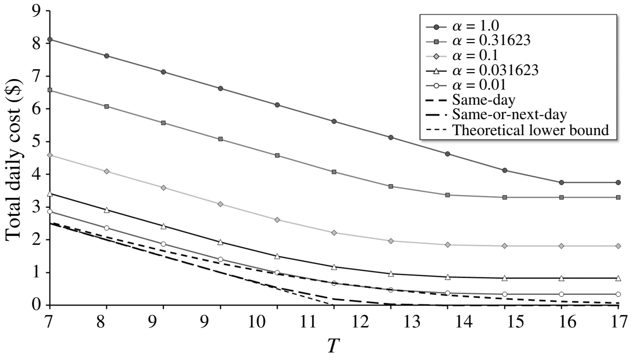


Figure 8 **Effect of the Overtime Surcharge β on Schedule Cost** ($\bar{n} = 12$, $T = 12$, $p = 25\%$, $N = 16$)

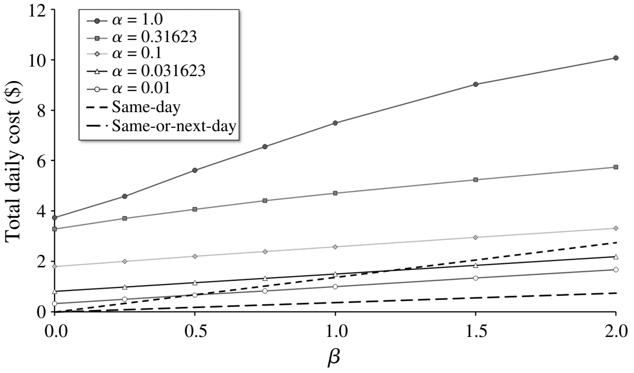
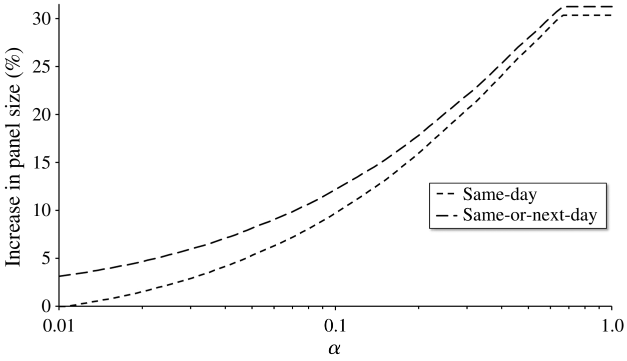


Figure 9 **Increase in Panel Size by Moving to an Open-Access Policy at the Same Cost** ($\bar{n} = 12$, $T = 12$, $p = 25\%$, $N = 16$, $\beta = 0.5$)



postpone patients to the following day is most advantageous near this point as well: when $T \gg \bar{n}$, the necessity of postponement is small, whereas when $T \ll \bar{n}$, it will be unusual for any day to be less than fully booked, with or without postponement.

4.3.4. Effects of the Overtime Surcharge β. In this section we keep the expected workload $\bar{n} = 12$, the standard day length $T = 12$, and the no-show probability $p = 25\%$, with $N = 16$ patients. We allow the overtime surcharge β to range between 0% and 200%; the resulting costs are graphed in Figure 8. Note that because the same and same-or-next-day policies incur only overtime costs, their costs are linear in β. In contrast, although the cost of the traditional scheduling policy is of course also increasing with the surcharge, it is concave in β since for higher values it can compensate by scheduling more patients at the beginning of the day, trading off increased patient waiting times against the increasingly expensive overtime surcharge. However, the same-day policy does perform well: even when β is four times the typical amount (2.0 versus 0.5), the same-day policy will still be preferable for all $\alpha > 0.06$. Yet again, the same-or-next-day policy dominates all other types of schedules.

4.3.5. Increasing the Panel Size ($\propto \bar{n}$)**.** Instead of calculating the reduction in cost that a physician could achieve by moving from a traditional to an open-access schedule, we could instead calculate the new value of the expected daily workload $\bar{n}$ for which each open-access policy will incur the same expected cost as the traditional policy. In other words, we can calculate the panel size (the number of patients covered by the physician, which is proportional to $\bar{n}$) that can be achieved by moving to an open-access policy. The results shown in Figure 9 indicate that for the base problem (day length $T = 12$, no-show probability $p = 25\%$, overtime surcharge $\beta = 0.5$, and $N = 16$ patients), the change in the panel size is almost always ($\alpha > 0.0105$) positive and is increasing with α, because the patients' waiting times will be eliminated under a same-day scheduling policy. For example, for $\alpha \geq 0.1032$ (0.0713), the panel size can be increased by at least 10% by moving from a traditional to a same-day (same-or-next-day) scheduling policy. These results are consistent with the 10% increase in patient count and 20% increase in revenue at Wisconsin's Dean Health System, as reported by Grandinetti (2000). Finally, when patient waiting is very important ($\alpha \geq 2/3$), moving from a traditional to an open-access scheduling policy could accommodate an increase of more than 30% in the doctor's panel size.

Intuitively, when α is large, single-booking will be optimal and overtime will be unavoidable; the cost of the traditional policy will be $(N - p) + \beta(N - p - T)$. (Because $\bar{n}$ is being changed, we must look at the length of the day rather than the idle time.) When overtime is necessary for an open-access schedule with an expected workload of $\bar{n} + \Delta\bar{n}$, costs (again looking at the length of the day instead of idle time) will be $(\bar{n} + \Delta\bar{n}) + \beta(\bar{n} + \Delta\bar{n} - T)$; equating these two cost functions yields a relative increase of

$$\frac{\Delta\bar{n}}{\bar{n}} = \left(\frac{p}{1-p}\right)\left(1 - \frac{1}{N}\right),$$

which is equal to 31.25% for our numerical example. (Recall that $\bar{n} = (1-p)N$.) Because overtime is not always necessary in same-day scheduling, this result holds only as an approximation but is quite close to the actual relative increase of 31.2% for the same-or-next-day policy and 30.3% for the same-day policy. Although the same-or-next-day policy is only slightly better than the same-day policy for large α, it is substantially better for smaller values of α, for which the allowable increase in panel size is much smaller.

4.3.6. The "No-Golf" Scheduling Policy. Until this point we have followed the standard literature in assuming the doctor's day concludes with the examination of the final patient, after which he or she is free to leave. However, there are realistic situations that would require the doctor to remain throughout the session, e.g., when patients can randomly "walk in" throughout the day. In this section, we examine an alternative assumption, where the doctor may not leave early, even after all the patients have been seen. In this case, the definitions of the expected patient waiting time $\overline{W}$ and the expected overtime $\bar{O}$ will not change. The doctor can leave at the latter of the actual day length D and the scheduled day length T, so that his or her expected idle time $\bar{I}^{\mathrm{NG}}$ can be written as

$$\bar{I}^{\mathrm{NG}} = E_D \max\{D, T\} - \bar{n} = T + \bar{O} - \bar{n}.$$

In other words, the doctor's day length is equal to the scheduled day length T plus the expected overtime $\bar{O}$.

Under this no-golf scheduling policy there will be no benefit to finishing early, so it is optimal to single-book patients whenever $N \le T$. As can be seen from Figure 10, the cost increase from imposing this restriction is much larger for the same-day policy than for either the traditional or the same-or-next-day policy. The same-day policy outperforms the traditional policy for $\alpha \ge 0.0758$, which is considerably more restrictive than the cutoff point of 0.0105 when the doctor leaves after seeing all the patients.

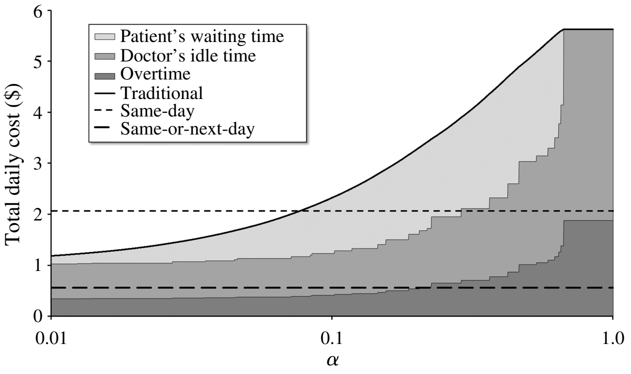
Proposition 4. *Consider the optimal scheduling policy* $\{x_t^*\}$ *for the alternate assumption on the doctor's length of day, where he or she must always stay through period* T. *Then*

(a) *Proposition* 1 *holds for this assumption as well.*

(b) *If* $N \le T$, *then* $x_t^* = 1$ *for* $t = 1, \ldots, N$ *and* 0 *otherwise.*

(c) *If* $N \ge T$, *then* $t_{\max}^* \ge T$.

Figure 10 **Cost of the Optimal Policies Under the No-Golf Model** ($\bar{n} = 12$, $T = 12$, $p = 25\%$, $N = 16$, $\beta = 0.5$)



4.3.7. Summary of Marginal Analysis. Two major results consistently hold over the various marginal analyses performed. First, the same-day policy outperforms the traditional policy except when α is very small ($\alpha < 0.03$) and so will be preferred whenever patient waiting is a moderate or significant component of the overall system performance. The same-day policy performs better for larger no-show probabilities p, for larger workloads $\bar{n}$, and for smaller overtime surcharges. Although the costs under the same-day and same-or-next-day policies are linearly increasing with the overtime surcharge, they remain well below the concave cost of the traditional scheduling policy when α is not too small ($\alpha > 0.06$). Additionally, moving from a traditional schedule to an open-access schedule will allow the physician to increase his or her panel size by up to 30%.

Second, the same-or-next-day scheduling policy is significantly less costly than the same-day policy when $T \approx \bar{n}$; otherwise (when overtime is either unusual or unavoidable), the option of postponing service makes little difference.

5. Extension to a Hybrid System

The traditional and open-access scheduling policies represent two extreme policies, corresponding to the cases when either all or none of the patients are previously scheduled to arrive each day. More realistic policies will of course fall somewhere in between, where most of the patient appointments are made ahead of time, but some random number of patients will also need to be seen that day. Following previous studies (e.g., Dobson et al. 2006), we assume that each of these unscheduled patients will call before the start of the day to make an appointment for that day. This assumption is reasonable for general practitioners and other professionals serving customers who must be seen the same day but who will rarely if ever require preemptive emergency care. We assume that patients are assigned to time slots as they call in, without the scheduler knowing the total number of unscheduled patients that will ultimately need to be seen.

This problem can be formulated as a nonlinear integer program. Essentially, we would introduce additional binary variables y_t^s that are equal to 1 if time slot t is assigned to one of the first s patients to call in and are equal to 0 otherwise, append the additional constraints $y_t^s \geq y_t^{s-1}$, and then calculate expected costs conditional on the number s of patients who call in that morning. However, the extreme intractability of this problem nullifies whatever small contribution its precise formulation might make. For example, Proposition 1 will no longer hold, as it may well become optimal to leave the first few time slots for same-day patients who will almost surely arrive. Although the formulation is intractable, we can use a lower bound to establish its inferiority to a same-day schedule.

Proposition 5. *If the same-day scheduling policy is preferable to the traditional scheduling policy, then it will also be preferable to any hybrid system.*

In other words, the more realistic hybrid policies that can accommodate a combination of same-day and previously scheduled patients will also be generally dominated by open-access policies.

6. Conclusions and Directions for Future Research

Under a traditional scheduling policy, a patient makes an appointment well ahead of time. This allows the variability in the demand arrival process to be smoothed out, at the expense of a long lead time until the patient's appointment. This in turn often results in patient no-shows, increasing the variability in the number of patients seen per day. Under an open-access policy, a patient calls in to make an appointment each morning and is seen by the doctor later that day (or possibly the next). This policy virtually eliminates the doctor's idle time due to no-shows, as well as the patients' waiting time caused by overbooking policies, but it is subject instead to the variability in demand arrivals.

This paper is the first to compare the performances of traditional and open-access scheduling policies under their respective sources of variabilities. It is also the first to find the optimal traditional scheduling policy with no-shows and deterministic service times. Our numerical results show that open-access scheduling is, in the wide majority of cases, clearly preferable to traditional appointment scheduling. In particular, open-access scheduling will perform better when patients' waiting times matter at least slightly, or when the no-show probability exceeds 5%. The same-day policy performs better for larger no-show probabilities, for larger workloads, and for most reasonable overtime surcharges. Comparing the same-day policy and same-or-next-day policy, we find that the ability to defer patients to the following day can help substantially, especially when the length of the workday is close to the expected workload. Finally, we find that the panel size (and revenues) can be increased by as much as 30% (when α is large) by moving to an open-access schedule, without exceeding the cost of the current traditional appointment policy.

Although our results are based on the assumptions of deterministic service times and punctual patient arrivals, the strong dominance of the open-access policy over the traditional scheduling policy, which was found in our numerical tests, suggests that it may well remain dominant in more realistic, if less tractable, settings. In particular, we show that whenever the traditional scheduling policy is dominated by the open-access policy, then the more realistic hybrid scheduling policy will be as well. The dominance of the same-day policy comes despite the higher variance of its daily workload, because the effects of this variance are limited to the end of the schedule. We recognize that these strong results are somewhat at variance with the dominance of traditional scheduling policies that we observe in practice. This might be explainable by the effect of other (unstudied) sources of variability, a truly negligible weight (i.e., $\alpha \ll 0.01$) placed on the value of patients' time, or perhaps a nonlinear cost function for the doctor (and staff) overtime costs.

One immediate extension that we plan to pursue is the inclusion of other types of variability (e.g., random service time, patient lateness, doctor interruptions) in the appointment system. Such an extension will, obviously and realistically, degrade the performance of both the traditional and the same-day scheduling policies. However, it is unclear whether the extent of such degradation will differ so much between the two policies that it will reverse our results here. Another extension would be to allow for multiple doctors; we expect such systems to perform better because of risk pooling among the doctors, but the relative improvement between traditional and open-access scheduling systems is not clear.

Electronic Companion

An electronic companion to this paper is available on the *Manufacturing & Service Operations Management* website (http://msom.pubs.informs.org/ecompanion.html).

References

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

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

Chen, R. 2007. Appointment scheduling under unpunctual arrivals. Working paper, Graduate School of Management, University of California at Davis, Davis.

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

Dobson, G., S. Hasija, E. J. Pinker. 2006. Reserving capacity for urgent cases in primary care. Working paper, Simon School of Business, University of Rochester, Rochester, NY.

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

Galucci, G., W. Swartz, F. Hackerman. 2005. Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatric Services* **56**(3) 344–346.

Grandinetti, D. A. 2000. You mean I can see the doctor today? *Medical Econom.* **77**(6) 102–114.

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

Green, L. V., S. Savin, M. Murray. 2007. Providing timely access to care: What is the right patient panel size? *Joint Commission J. Quality Patient Safety* **33**(4) 211–218.

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

Hixon, A. L., R. W. Chapman, J. Nuovo. 1999. Failure to keep clinic appointments: Implications for residency education and productivity. *Family Medicine* **31**(9) 627–630.

Ho, C.-J., H.-S. Lau. 1992. Minimizing total cost in scheduling outpatient appointments. *Management Sci.* **38**(12) 1750–1764.

Ho, C.-J., H.-S. Lau. 1997. Evaluating the impact of operating conditions on the performance of appointment scheduling rules in service systems. *Eur. J. Oper. Res.* **112**(3) 542–553.

Jansson, B. 1966. Choosing a good appointment system—A study of queues of the type $(D, M, 1)$. *Oper. Res.* **14**(2) 292–312.

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

LaGanga, L. R., S. R. Lawrence. 2007a. An appointment overbooking model to improve client access and provider productivity. *Proc. New Challenges Service Oper., POMS College Service Oper., EurOMA, London*.

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

Mercer, A. 1960. A queuing problem in which the arrival times of the customers are scheduled. *J. Roy. Statist. Soc. Ser. B* **22**(1) 108–113.

Mercer, A. 1973. Queues with scheduled arrivals: A correction, simplification and extension. *J. Roy. Statist. Soc. Ser. B* **35**(1) 104–116.

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

Mondschein, S. V., G. Y. Weintraub. 2003. Appointment policies in service operations: A critical analysis of the economic framework. *Production Oper. Management* **12**(2) 266–286.

Murray, M., D. M. Berwick. 2003. Advance access: Reducing waiting and delays in primary care. *J. Amer. Medical Assoc.* **289**(8) 1035–1039.

Murray, M., C. Tantau. 2000. Same-day appointments: Exploding the access paradigm. *Family Practice Management* **7**(September) 45–50.

Muthuraman, K., M. Lawley. 2008. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Trans.* **40**(9) 820–837.

O'Brien, G., R. Lazebnik. 1998. Telephone call reminders and attendance in an adolescent clinic. *Pediatrics* **101**(6) e6.

Pegden, C. D., M. Rosenshine. 1990. Scheduling arrivals to queues. *Comput. Oper. Res.* **17**(4) 343–348.

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

Qu, X., R. L. Rardin, J. A. A. Williams, D. R. Willis. 2007. Matching daily healthcare provider capacity to demand in advanced access scheduling systems. *Eur. J. Oper. Res.* **183**(2) 812–826.

Robinson, L. W., R. Chen. 2003. Scheduling doctors' appointments: Optimal and empirically-based heuristic policies. *IIE Trans.* **35**(3) 295–307.

Rust, C. T., N. H. Gallups, W. S. Clark, D. S. Jones, W. D. Miller. 1995. Patient appointment failures in pediatric resident continuity clinics. *Arch. Pediatrics Adolescent Medicine* **149**(6) 693–695.

Shuster, M. 2003. Advanced-access scheduling in primary care. *J. Amer. Medical Assoc.* **290**(3) 332–333.

Soriano, A. 1966. Comparison of two scheduling systems. *Oper. Res.* **14**(3) 388–397.

Steinbauer, J. R., K. Korell, J. Erdin, S. J. Spann. 2006. Implementing open-access scheduling in an academic practice. *Family Practice Management* **13**(3) 59–64.

Ulmer, T., C. Troxler. 2006. The economic cost of missed appointments and the open access system. Working paper, University of Florida Area Health Education Centers Program, Gainesville.

Vissers, J. 1979. Selecting a suitable appointment system in an outpatient setting. *Medical Care* **17**(12) 1207–1220.

Vissers, J., J. Wijngaard. 1979. The outpatient appointment system: Design of a simulation study. *Eur. J. Oper. Res.* **3**(6) 459–463.

Wang, P. P. 1993. Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Res. Logist.* **40**(3) 345–360.

Wang, P. P. 1997. Optimally scheduling *N* customer arrival times for a single-server system. *Comput. Oper. Res.* **24**(8) 703–716.

Weiss, E. N. 1990. Models for determining estimated start times and case orderings in hospital operating rooms. *IIE Trans.* **22**(2) 143–150.

Xakellis, G. C., A. Bennett. 2001. Improving clinic efficiency of a family medicine teaching clinic. *Family Medicine* **33**(7) 533–538.

Zeng, B., A. Turkcan, J. Lin, M. Lawley. 2009. Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. *Ann. Oper. Res.*, ePub ahead of print June 12.