M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 14, No. 4, Fall 2012, pp. 670–684
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs®
http://dx.doi.org/10.1287/msom.1120.0394
© 2012 INFORMS

Appointment Scheduling Under Patient No-Shows and Service Interruptions

Jianzhe Luo, Vidyadhar G. Kulkarni, Serhan Ziya
Department of Statistics and Operations Research, University of North Carolina at Chapel Hill,
Chapel Hill, North Carolina 27599 {jzluo@email.unc.edu, vkulkarn@email.unc.edu, ziya@email.unc.edu}

We consider an appointment-based service system (e.g., an outpatient clinic) for which appointments need to be scheduled before the service session starts. Patients with scheduled appointments may or may not show up for their appointments. The service of scheduled patients can be interrupted by emergency requests that have a higher priority. We develop a framework that can be utilized in determining the optimal appointment policies under different assumptions regarding rewards, costs, and decision variables. We propose two methods to evaluate the objective function for a given appointment schedule. We specifically consider two different formulations, both of which aim to balance the trade-off between the patient waiting times and server utilization and carry out a numerical study to provide insights into optimal policies. We find that policies that ignore interruptions perform quite badly, especially when the number of appointments to be scheduled is also a decision variable. We also find that policies that require equally spaced appointments perform reasonably well when the interruption rate is constant. However, their performance worsens significantly when the interruption rate is time dependent.

Key words: healthcare operations management; service operations; stochastic methods
History: Received: March 7, 2011; accepted: February 16, 2012. Published online in *Articles in Advance* July 13, 2012.

1. Introduction

In healthcare, appointment systems mainly work to regulate the patient demand for various services. They help reduce the variability in the patients' arrival process so that patients wait less and the system is kept highly utilized. Clearly, however, it is not possible to eliminate the variability completely. Patients may arrive earlier or later than their scheduled appointment times, or they may simply not show up at all. It may take longer than expected to serve a particular patient, or the service can be interrupted for various reasons, including arrivals of emergency patients who need to be attended to right away. Some of these factors have been considered within the large and growing body of work on appointment scheduling, but to the best of our knowledge, little attention has been paid to how to schedule appointments when the scheduled service can be interrupted. The objective of this paper is to fill this gap in the literature by proposing methods for determining appointment times in the presence of service interruptions, evaluating the importance of incorporating service interruptions in the decision models, and identifying the structural properties of the optimal policies.

Service interruptions are prevalent in many service systems, and the formulation we consider in this paper does not have any features that make it exclusively appropriate for a specific setting. Therefore, the results in this paper are relevant to a wide class of appointment-based service systems. However, we are primarily motivated by applications from healthcare, where service interruptions are mostly caused by emergency patients who need immediate attention. For example, physicians and dentists can be called to attend to or to be consulted for emergency patients (see, e.g., Kenny and Barrett 2005, Alderman 2011). Many healthcare clinics of various specialties and dental offices warn their patients in advance that in case of emergencies, they may experience longer waiting times (see, e.g., http://www.nwh.org/clinical-centers/spine-center/patient-faq/, http://www.northfultonpediatrics.com/policies, http://princetonpediatricdentist.com/faqs). In fact, interruptions to scheduled appointments are not necessarily caused by patients in need of immediate attention. According to Klassen and Yoogalingam (2008), such interruptions may include calls from other doctors or pharmacists and problems that require dealing with the administrative staff. The interrupted process can also be the service provided by a diagnostic machine. Typically, patients make appointments for access to computerized tomography (CT) scans or magnetic resonance imaging (MRI) machines at hospitals. However, frequently, emergency physicians find it necessary to use these machines for emergency patients who cannot "afford" to wait. Such patients, when sent for a diagnostic scan, get higher priority than and cause additional delays for the regular patients who have scheduled appointments (see, e.g., Green et al. 2006). Recent studies suggest that the rate of such emergency use of these machines is quite high and has been increasing significantly over the last years. For example, Korley et al. (2010) conducted a national survey of patient visits to emergency departments within the United States and found that the percentage of emergency department visits that required the use of CT scan or MRI increased from 6% in 1998 to 15% in 2007. Broder and Warshauer (2006) analyzed adult patient data from the emergency department of a university hospital. They found that from 2000 to 2005, the adult emergency department volume increased by 13%; head CT scans increased by 51%, cervical spine CTs by 463%, chest CTs by 226%, abdominal CTs by 72%, and miscellaneous CTs by 132%. These numbers clearly show the significant increase in the use of the CT scan at this hospital over a five-year period. The authors have also observed that except for the abdominal CT, which seemed to level off over the last year of the five-year period, the numbers of all types of CT scans have consistently increased at a rate higher than that for the adult patient volume. Another study carried out at the emergency department of the HealthAlliance Hospital in Leominster, Massachusetts, found that roughly half of the patients who go through the radiology department originated from the emergency department and that these patients were given top priority for access to radiology services with no delay (Anderson et al. 2010). The disruption of regularly scheduled appointments by emergencies could be prevented if emergency departments had dedicated diagnostic machines. However, because it is prohibitively costly to have a separate diagnostic machine for the exclusive use of emergency patients, this solution is not feasible for most hospitals. Therefore, these machines are typically shared by regular patients, who schedule appointments in advance, and emergency patients, who come without appointments and get higher priority, and there is usually a strong incentive to keep them as highly utilized as possible (see Green et al. 2006).

Earlier work on appointment scheduling has provided many useful insights on how appointments should be scheduled over a given period of time when there are no interruptions to the service process. However, it is not known whether or how explicit consideration of interruptions (e.g., emergency cases) would change these earlier insights. One of the objectives of this paper is to investigate this question. For example, we know that when the service times are independent and identically distributed, and there are no interruptions to the service process, the optimal scheduling policy has a "dome" shape, meaning that the times between consecutive appointments that are scheduled early or late in the day are small, whereas the times between those scheduled midday are larger (see Hassin and Mendel 2008). We also know that requiring the time between any two consecutive appointments to be the same does not degrade the policy performance significantly (see Stein and Côté 1994). The question is whether these observations are valid in the presence of interruptions as well. When there are interruptions, does the optimal policy still have a dome shape, and does the optimal policy under the restriction that appointments are scheduled at equally spaced intervals perform sufficiently well? Perhaps there is no good reason to suspect that the answers to these questions would be any different when emergency cases arrive with some fixed rate, but what if the rate changes depending on the time of the day? For example, the rate of arrivals to emergency departments is known to be time dependent, which means that the arrival rate of emergencies to the diagnostic machines is also very likely to be time dependent. In such cases, one can see that insisting that the appointments be equally spaced could be more "costly," as it might make more sense to schedule fewer appointments around the times when the arrival rates of emergency cases are higher.

In this paper, we first develop an appointment-scheduling model that differs from prior models mainly in that the service of regularly scheduled patients can be temporarily suspended because of interruptions. Our model can be seen as a generalization of the model of Hassin and Mendel (2008), who implicitly assumed that there are no interruptions. We assume that interruptions occur according to a Poisson process, but we allow the interruption rate to change with time. This is one of the important features of our formulation, as it fits nicely with our motivating applications. The complexity of our formulation makes it very difficult if not impossible to characterize the optimal policy analytically. This is not surprising, because even for the simpler case, where there are no interruptions, Hassin and Mendel (2008) resort to numerical analysis to generate insights on the problem. In fact, even a simple computation of the objective function for a given appointment schedule is a significant challenge in our optimization problem; therefore, the core of our analysis is devoted to the question of how this computation can be done. In particular, the computation requires the solution of a system of differential equations, which is not readily available. However, we provide two different methods, either of which can be used to find a solution and thus compute the objective function. After developing these solution methods, we use them for a numerical study to quantify the potential benefits of incorporating the interruption process into the formulation, we then investigate how explicit consideration of interruptions influences the key insights on optimal appointment scheduling. We find that ignoring interruptions when they are in fact prevalent can result in appointment schedules that demonstrate significantly worse performances.

The remainder of this paper is organized as follows: Section 2 gives a review of the related literature. In §§3 and 4, we introduce the formulation. In §5, we develop the two methods that can be used to compute the objective function. Section 6 demonstrates how the method for computing the objective function can also be used in computing the expected patient waiting time and server overtime. In §7, we show how our formulation can be generalized to allow the interruption time to have a phase-type distribution. Section 8 provides our numerical results. Finally, we conclude with §9.

2. Literature Review

The operations management literature on appointment scheduling is vast and rapidly expanding. For an extensive review of this literature, as well as discussions on directions for future research, we refer the reader to Cayirli and Veral (2003) and Gupta and Denton (2008). Here, we only mention those papers that are either very closely related to this paper or very recent.

Gupta and Denton (2008) propose a useful classification scheme for appointment-scheduling models, depending on the type of waiting that is formulated. They define a patient's *direct waiting time* as the time the patient spends in the clinic on the day of his appointment and *indirect waiting time* as the time between the patient's call for an appointment and the scheduled appointment time. There is some relatively recent work on indirect waiting (see Gupta and Wang 2008, Green and Savin 2008, Liu et al. 2010), but the vast majority of the papers focus on performance measures related to direct waiting. This paper also contributes to this literature.

When determining appointment times on a given day, there are a number of objectives, including keeping the server (e.g., physician) busy, keeping waiting times short, and avoiding or minimizing overtime. Papers that deal with direct waiting time typically consider one or more of these objectives, in many cases by minimizing an objective function that is a weighted sum of a subset of these various performance measures (weighted by their relative "costs") and/or adding them as constraints into the formulation. For some examples, see Bailey (1952), Fries and Marathe (1981), Wang (1997, 1999), Denton and Gupta (2003), Kaandorp and Koole (2007), Robinson and Chen (2003), Muthuraman and Lawley (2008), Chakraborty et al. (2010), and Jouini and Benjaafar (2012).

The three papers that appear to be the closest to our work are Pegden and Rosenshine (1990), Stein and Côté (1994), and Hassin and Mendel (2008). These three papers consider models that are special cases of our model. Pegden and Rosenshine (1990) obtain a closed-form solution for the optimal schedule for the case where there are only two appointments; they develop a method to compute the optimal schedule for the general case with more than two appointments. What is mainly different in the model of Pegden and Rosenshine (1990) (with respect to our formulation) is that all customers show up for their appointments and the service process never gets interrupted. Stein and Côté (1994) mainly build on Pegden and Rosenshine (1990) and study the effect of requiring equally spaced appointment times. On the other hand, Hassin and Mendel (2008) generalize the model of Pegden and Rosenshine (1990) by allowing no-shows. They carry out a numerical study and generate insights on the structure of the optimal appointment policy, the importance of modeling no-shows, the effects of no-shows on the optimal policy and its performance, and the "cost" of forcing equidistant appointments. We generalize the model of Hassin and Mendel (2008) by allowing the service of scheduled patients to possibly be interrupted. In our analysis, we mainly investigate the importance of modeling interruptions and how their existence changes the main insights obtained earlier in the literature, mostly in these three papers.

Although, in general, limited work on interruptions has appeared within the context of appointment scheduling, we are aware of three other papers that share our primary motivation, as they also deal with service interruptions at outpatient clinics and diagnostic machines. However, these papers use completely different analytical techniques and/or structurally different formulations. In particular, Klassen and Yoogalingam (2008) study nonemergency physician interruptions in an outpatient clinic using simulation optimization. Fiems et al. (2012) develop a queueing model and carry out steady-state analysis to investigate the impact of emergency requests on the waiting time of regularly scheduled patients in the radiology department. On the other hand, Vasanawala and Desser (2005) develop a simple mathematical model to obtain the number of schedule slots to leave open for emergency CT scan or ultrasonography requests.

There are also papers (many from the traditional job-scheduling and queueing literature) that analyze models in which the service might get interrupted because of a server failure or vacation. However, with one exception, which we discuss below, these papers make assumptions that do not fit well with the appointment-scheduling problem. For example, Federgruen and Green (1986), Takine and Sengupta (1997), and Gray et al. (2000) all consider queueing models in which the server can go on and off, but they assume that customers arrive according to some stationary process (e.g., Poisson) and carry out steady-state analysis. Glazebrook (1984), Adiri et al. (1989), and Birge et al. (1990), on the other hand, assume that all jobs are available to be processed at the beginning of the service session and the decision to be made is the order in which these jobs will be processed. One exception from the job-scheduling literature is Wang (1994), who develops an algorithm that determines the optimal release times of a finite number of jobs to an unreliable machine. His model is in fact almost the same as ours, with one difference being our consideration of the possibility of no-shows. However, there is one important error in the analysis of Wang (1994) that affects the resulting expressions and methods significantly. The error is related to the author's implicit independence assumption in the derivation of an equation when, in fact, there is dependence. We provide details on this in Appendix B of the online supplement (available at http://dx.doi.org/10.1287/msom.1120.0394).

3. Model Description

The methodology we use in this paper can be used in a variety of formulations that consider the scheduling of a finite number of appointments over a finite or infinite horizon. We consider two such formulations, one of which has received significant attention in the literature, but with the restriction that there are no service interruptions. To keep the presentation simple, we introduce the models assuming that interruptions occur according to a homogeneous Poisson process. In §8, we explain how our analysis can easily be extended to the case where interruptions occur according to a nonhomogeneous Poisson process, and we provide numerical results under that generalization.

3.1. Model I: Restricted Scheduling Horizon

Suppose that there is a predetermined scheduling horizon $[0, T]$ where $T < \infty$. At time zero, we need to decide N, the number of appointments to be scheduled over this time interval, as well as the times for these N appointments. We define d_k as the appointment time scheduled for the kth patient, $k = 1, 2, \ldots, N$. The vector $\mathbf{d} = (d_1, d_2, \ldots, d_N)$, where $0 \le d_1 \le d_2 \le \cdots \le d_N \le T$, is called a schedule for these N patients. Scheduled patients either show up punctually at their appointment times with probability p or become no-shows in an independent manner. Patients who show up are served on a first-come first-served basis. The service times for patients are assumed to be independent and identically distributed according to an exponential distribution with mean $1/\mu$. However, services can be interrupted by certain events, which we assume to occur according to a Poisson process with rate η. Once the server is interrupted, it stays in that stage for an amount of time that is exponentially distributed with rate θ, and during that time any new interruptions are assumed to have no effect. In §7, we show how the exponential distribution assumption on the interruption times can be relaxed by allowing them to have a phase-type distribution, which also makes it possible to model more explicit connections between interruption events and the interruption durations (e.g., explicit modeling of emergency patients who queue up). The service for scheduled patients is preemptive resume; that is, the service for a scheduled patient is suspended immediately in the presence of interruptions and resumes with no loss of work when the server becomes available again.

It might be helpful to think of the whole service horizon as a sequence of "on" and "off" periods. During the "on" periods, the server is available to work on regularly scheduled patients. During the "off" periods, the server is not available and is engaged in other activities, such as attending to emergency patients. At time zero, the server is available for serving scheduled patients; that is, the service session starts with an "on" period. An interruption ends this "on" period and starts an "off" period, during which no scheduled patients can be served. Once this period is over, the server becomes available for scheduled patients again and another "on" period starts. The server status alternates between these "on" and "off" states until the services of all the scheduled patients who show up are completed. Even though all appointments need to be scheduled some time between 0 and T, it is possible that some of the scheduled patients will be served after time T. Note that, even after time T, services of the regular patients can still be interrupted. However, if all the scheduled patients who show up are served by T, the server is turned off and no more interruptions occur.

The system incurs the waiting cost from scheduled patients (the waiting time of a scheduled patient is the total time the patient spends in the system minus the time in service) and the server overtime cost if the service completion time of all the patients who show up is later than T. We use c_w to denote the patient waiting cost per unit of time and c_l to denote the server overtime cost per unit of time beyond T. In addition, the system earns a reward r from each scheduled patient who receives service. The objective is to find the optimal policy $(N^*, \mathbf{d}^*)$ to maximize $\Pi(N, \mathbf{d})$, the total expected net profit, which is the reward from serving scheduled patients minus the patient waiting and the server overtime cost.

3.2. Model II: Unrestricted Scheduling Horizon

Model II makes the same assumptions as Model I regarding patients' service times, no-show behavior, and service interruption process, but differs from Model I in a few important aspects. Most importantly, N, the total number of appointments to be scheduled, is not a decision variable and there is no restriction on when these N appointments can be scheduled (i.e., $T=\infty$). In other words, the number of appointments to be scheduled is given and the decision to be made at time zero is at what time to schedule these appointments. The system keeps operating until the appointment time assigned to the Nth patient or the service completion time of the last patient who shows up, whichever is later. We consider two types of cost: c_w, as defined in Model I, and c_s, the cost of operating the system per unit of time (service availability cost). The objective is to find the optimal schedule $\mathbf{d}^*$ for these N patients to minimize the total expected cost. Note that this model reduces to the model of Hassin and Mendel (2008) if we assume that there are no service interruptions and the server is available at all times; it reduces to the model of Pegden and Rosenshine (1990) if we further assume that (in addition to the no interruption assumption) all patients show up for their appointments.

4. Complete Description of the Optimization Problem

In this section, we provide a more complete description of the optimization problem for Model I. It is important to note that the treatment of the optimization problem in Model II is similar, with some minor differences; therefore, we skip it for brevity. The proofs of all the analytical results are given in Appendix A of the online supplement.

4.1. Formal Statement of the Optimization Problem

Our optimization problem can briefly be stated as follows:

$$\max_{N,\,\mathbf{d}} \quad \Pi(N,\mathbf{d})$$
$$\text{s.t.} \quad 0\le d_1\le d_2\le\cdots\le d_N\le T,$$

where $\Pi(N,\mathbf{d})$ is the total expected net profit.

An analytical characterization of the optimal policy does not appear to be possible because of the complexity of the problem. Therefore, a more realistic goal, which we pursue in this paper, is to develop a numerical solution method. In fact, even the computation of the objective function $\Pi(N,\mathbf{d})$ is a significant challenge because it does not have a closed-form expression. We can, however, obtain $\Pi(N,\mathbf{d})$ by solving a system of differential equations, as we demonstrate in the following.

4.2. Effective Service Time

Even though the time it takes to serve a scheduled patient has an exponential distribution, the time between the start of a given patient's service and its end, called the *effective service time*, is not exponentially distributed because of the possibility of interruptions. Let X be the effective service time of a scheduled patient who shows up, and let $G(t)=P\{X\le t\}$. Recall that η is the Poisson arrival rate of interruptions, $1/\mu$ is the mean service time, and $1/\theta$ is the mean interruption time. One can show that (see the proof of Proposition 1 in Appendix A of the online supplement)

$$G(t)=(1-\beta)(1-e^{-at})+\beta(1-e^{-bt}), \tag{1}$$

where $\beta=(\mu-a)/(b-a)$, and

$$a=\tfrac{1}{2}\big[\eta+\mu+\theta+\sqrt{(\eta+\mu+\theta)^2-4\theta\mu}\big]>0, \tag{2}$$

$$b=\tfrac{1}{2}\big[\eta+\mu+\theta-\sqrt{(\eta+\mu+\theta)^2-4\theta\mu}\big]>0. \tag{3}$$

Hence, X is a mixture of two exponential distributions and its mean is given by

$$E(X)=\int_0^\infty(1-G(t))\,\mathrm{d}t=\frac{\eta+\theta}{\theta\mu}. \tag{4}$$

4.3. Recursive Expression for the Objective Function

In this section, we derive the system of differential equations, which needs to be solved to evaluate the objective function $\Pi(N,\mathbf{d})$. To that end, first denote the server state by 0 if it is available for scheduled patients and by 1 if not. Let $d_0=0$ and $d_{N+1}=T$. Also define the net profit function associated with each appointment interval $[d_k,d_{k+1})$, $k=0,1,\ldots,N$ as follows: $0<t\le d_{k+1}-d_k$, $R^k_{n,i}(t)$ is the total expected net profit earned by operating the system over $[d_{k+1}-t,\infty)$ if at time $d_{k+1}-t$ there are n scheduled patients in the system and the server is in state i, where $n=0,1,\ldots,k$, and $i=0,1$. We assume that the server is available for scheduled patients at time zero, and thus $R^0_{0,0}(d_1)$ is the net profit the system earns over $[0,\infty)$. Consequently, we have $\Pi(N,\mathbf{d})=R^0_{0,0}(d_1)$.

To obtain $R^0_{0,0}(d_1)$ (or $\Pi(N,\mathbf{d})$), we first need to characterize the expected net profit function $R^k_{n,i}(t)$ for each $k=0,1,\ldots,N$ and for $t\in(0,d_{k+1}-d_k]$, that is, between any two consecutive appointment times, in the interior of the appointment interval. In addition, we need to establish how $R^k_{n,i}(t)$ for different values of k, n, and i are related. To do this, for each $k=0,1,\ldots,N$, $n=0,1,\ldots,k$, and $t\in(0,d_{k+1}-d_k]$, denote

$$R^k_n(t)=\begin{bmatrix}R^k_{n,0}(t)\\ R^k_{n,1}(t)\end{bmatrix}\quad\text{and}\quad\frac{\mathrm{d}R^k_n(t)}{\mathrm{d}t}=\begin{bmatrix}\dfrac{\mathrm{d}R^k_{n,0}(t)}{\mathrm{d}t}\\[2ex] \dfrac{\mathrm{d}R^k_{n,1}(t)}{\mathrm{d}t}\end{bmatrix}.$$

Also let

$$A=\begin{bmatrix}\mu & 0\\ 0 & 0\end{bmatrix},\quad B=\begin{bmatrix}-(\eta+\mu) & \eta\\ \theta & -\theta\end{bmatrix},$$

$$E=\begin{bmatrix}-\eta & \eta\\ \theta & -\theta\end{bmatrix},\quad \text{and}\quad C_w=\begin{bmatrix}c_w\\ c_w\end{bmatrix}.$$

We can prove the following theorem.

Theorem 1. *For each $k=0,1,\ldots,N$, the vector of the net profit functions $R_n^k(t)$, $0<t\le d_{k+1}-d_k$, satisfies the following differential equations*:

$$\frac{\mathrm{d}R_0^k(t)}{\mathrm{d}t}=ER_0^k(t), \tag{5}$$

$$\frac{\mathrm{d}R_n^k(t)}{\mathrm{d}t}=-\begin{bmatrix}n-1 & 0\\ 0 & n\end{bmatrix}C_w+AR_{n-1}^k(t)+BR_n^k(t),\qquad n=1,\ldots,k, \tag{6}$$

with boundary conditions

$$R_{0,0}^N(0^+)=R_{0,1}^N(0^+)=0, \tag{7}$$

$$R_{n,0}^N(0^+)=-c_l nE(X)-c_w\left[\frac{n(n+1)}{2}E(X)-\frac{n}{\mu}\right],\qquad n=1,\ldots,N, \tag{8}$$

$$R_{n,1}^N(0^+)=-c_l\left(\frac{1}{\theta}+nE(X)\right)-c_w\left[\frac{n}{\theta}+\frac{n(n+1)}{2}E(X)-\frac{n}{\mu}\right],\qquad n=1,\ldots,N, \tag{9}$$

$$R_{n,i}^k(0^+)=p\big(r+R_{n+1,i}^{k+1}(d_{k+2}-d_{k+1})\big)+(1-p)R_{n,i}^{k+1}(d_{k+2}-d_{k+1}),$$
$$k=0,1,\ldots,N-1,\ n=0,\ldots,k,\ i=0,1. \tag{10}$$

Theorem 1 states the differential equations that the functions $R_n^k(\cdot)$ need to satisfy, but the solution to these equations is not directly available. In §5, we describe how to solve them.

5. Two Methods for Computing the Objective Function

In this section, we propose two methods, the method of Laplace transform (LT) and the method of integrating factor, both of which can be used to evaluate the objective function, $\Pi(N,\mathbf{d})$, for given N and $\mathbf{d}$.

5.1. Method I: Using Laplace Transforms

For each $k=0,1,\ldots,N$, $R_n^k(t)$ is defined on $t\in(0,d_{k+1}-d_k]$. To apply the method of LT, the domain of $R_n^k(t)$ is extended to be $t\in(0,\infty)$. After $R_n^k(t)$ is obtained, we only need its values on $t\in(0,d_{k+1}-d_k]$. Let $\tilde{R}_n^k(s)$ denote the LT of $R_n^k(\cdot)$ for $k=0,1,\ldots,N$ and $n=0,1,\ldots,k$. Then we can show that $\tilde{R}_n^k(s)$ can be obtained recursively as stated in the following theorem.

Theorem 2. *For each $k=0,1,\ldots,N$, we have*

$$\tilde{R}_0^k(s)=(sI-E)^{-1}R_0^k(0^+), \tag{11}$$

$$\tilde{R}_n^k(s)=\big[(sI-B)^{-1}A\big]^n(sI-E)^{-1}R_0^k(0^+)+\sum_{j=0}^{n-1}\left\{\big[(sI-B)^{-1}A\big]^j(sI-B)^{-1}\cdot\left[-\begin{bmatrix}n-1-j & 0\\ 0 & n-j\end{bmatrix}\frac{C_w}{s}+R_{n-j}^k(0^+)\right]\right\},$$
$$n=1,\ldots,k, \tag{12}$$

where $R_0^k(0^+)$ and $R_{n-j}^k(0^+)$, $j=0,\ldots,n-1$, can be obtained using the boundary conditions (7)–(10).

For a given schedule $\mathbf{d}=(d_1,d_2,\ldots,d_N)$, Theorem 2 suggests a recursive procedure that can be used to obtain the LT $\tilde{R}_n^k(s)$ for each $k=0,1,\ldots,N$ and $n=0,1,\ldots,k$, which can then be inverted to obtain $R_n^k(t)$. In particular, $R_{0,0}^0(t)$ is equal to $\Pi(N,\mathbf{d})$ for $t=d_1$. The following algorithm is a detailed description of this recursive procedure.

Algorithm 1

Step 1. Initialize: Set $k=N$. Compute $E(X)$, the expected length of the effective service time, and use (8) and (9) to evaluate $R_n^N(0^+)$ for $n=0,1,\ldots,N$.

Step 2. Apply (11), (12), and the boundary constraint (10) to compute $\tilde{R}_n^k(s)$, the LT of $R_n^k(t)$, for $n=0,1,\ldots,k$.

Step 3. For each $n=0,1,\ldots,k$, invert $\tilde{R}_n^k(s)$ to obtain $R_n^k(t)$ and evaluate its value at $t=d_{k+1}-d_k$, which will be used in Step 2 of the next iteration.

Step 4. If $k>0$, set $k=k-1$ and go to Step 2. Otherwise, stop. The objective function $\Pi(N,\mathbf{d})=R_{0,0}^0(d_1)$ has been obtained in the last iteration of the algorithm.

In Step 3 of the algorithm, $\tilde{R}_n^k(s)$ needs to be inverted to obtain $R_n^k(t)$. It can be shown that all the terms appearing in (11) and (12) are rational functions of s. Hence, one can use the method of partial fraction decomposition (see Horowitz 1971) to invert $\tilde{R}_n^k(s)$ to obtain $R_n^k(t)$ for $k=0,1,\ldots,N$ and $n=0,\ldots,k$. The details are omitted for brevity.

5.2. Method II: Using an Integrating Factor

An alternative and more direct way of determining the solution to the system of differential equations given in Theorem 1 is to use the method of integrating factor. According to this method, we multiply both sides of (6) by e^{-Bt}, the "integrating factor," and solve the differential equations. The use of this solution method for solving differential equations leads to the following theorem, which suggests a recursive procedure that can be used to determine $R_n^k(t)$, $k = 0, 1, \dots, N$ and $n = 0, \dots, k$.

Theorem 3. *Let*

$$H = \begin{bmatrix} \dfrac{-a+\theta}{\theta} & \dfrac{\eta}{\theta} \\ 1 & \dfrac{b-\theta}{\theta} \end{bmatrix},$$

$$J = \begin{bmatrix} \dfrac{b-\theta}{\theta} & \dfrac{-\eta}{\theta} \\ -1 & \dfrac{-a+\theta}{\theta} \end{bmatrix}, \quad \text{and} \quad L = \begin{bmatrix} \theta & \eta \\ \theta & \eta \end{bmatrix},$$

where a and b are given by (2) *and* (3), *respectively. For each $k = 0, 1, \dots, N$, and $\eta > 0$,*

$$R_n^k(t) = D_n^k(t) + z_n^k, \quad n = 0, 1, \dots, k,$$

where z_n^k and $D_n^k(t)$ are given as follows:

$$z_0^k = [0, 0]',$$

$$z_n^k = B^{-1}\left(\begin{bmatrix} n-1 & 0 \\ 0 & n \end{bmatrix} C_w - A z_{n-1}^k\right), \quad n = 1, 2, \dots, k,$$

$$D_0^k(t) = u_0^{0,k} + v_0^{0,k} e^{-(\eta+\theta)t} + m_{0,0}^{0,k} e^{-at} + q_{0,0}^{0,k} e^{-bt},$$

$$D_n^k(t) = \sum_{j=-n}^{n} u_j^{n,k} e^{j(a-b)t} + \sum_{j=-n}^{n} v_j^{n,k} e^{[-(\eta+\theta)+j(a-b)]t}$$
$$+ \sum_{j=0}^{n-1} \sum_{i=0}^{n-1-j} m_{i,j}^{n,k} t^i e^{-[a+j(a-b)]t}$$
$$+ \sum_{j=0}^{n-1} \sum_{i=0}^{n-1-j} q_{i,j}^{n,k} t^i e^{-[b+j(b-a)]t}, \quad n = 1, 2, \dots, k.$$

In the above equations, $m_{0,0}^{0,k} = q_{0,0}^{0,k} = [0, 0]'$,

$$u_0^{0,k} = \frac{LR_{0,k}(0^+)}{\eta+\theta}, \quad v_0^{0,k} = \frac{-ER_{0,k}(0^+)}{\eta+\theta},$$

and $u_j^{n,k}$, $v_j^{n,k}$, $m_{i,j}^{n,k}$, $q_{i,j}^{n,k}$, $n = 1, \dots, k$, can be obtained recursively, as described in Appendix A of the online supplement.

The statement of Theorem 3 is not complete because the recursive expressions for $u_j^{n,k}$, $v_j^{n,k}$, $m_{i,j}^{n,k}$, and $q_{i,j}^{n,k}$ are not provided. We provide these long expressions in Appendix A of the online supplement as part of the complete statement of this theorem along with its proof.

6. Computing the Expected Patient Waiting Time and Server Overtime

The expected waiting time of each patient with a scheduled appointment and the expected server overtime are not obtained explicitly when the optimization problems for Models I and II are solved. However, one could easily come up with alternative formulations in which one may want to put constraints such as keeping the maximum expected waiting time or the server overtime below a certain level while maximizing or minimizing a particular objective. Here we show that our methodology can be used to compute such performance measures as well, because our reward function reduces to the patient waiting time or the server overtime when model parameters are set appropriately.

Given a schedule $\mathbf{d} = (d_1, d_2, \dots, d_N)$, suppose we want to compute the expected waiting time of the kth scheduled patient if he shows up, $1 \le k \le N$. Note that the waiting time of the kth patient depends only on the schedule of the first $k-1$ patients. Hence, the problem of finding the mean waiting time of the kth patient (assuming he shows up) can be formulated as a modified version of the original problem. Specifically, consider the first k patients, the schedule of whom is a subvector of $\mathbf{d}$—that is, $(d_1, \dots, d_{k-1}, d_k)$, set $r = c_w = c_l = 0$—and change boundary constraints (8) and (9) to $R_{n,0}^{k-1}(0^+) = (n+1)E(X) - 1/\mu$, $n = 0, 1, \dots, k-1$, and $R_{n,1}^{k-1}(0^+) = 1/\theta + (n+1)E(X) - 1/\mu$, $n = 0, 1, \dots, k-1$, respectively.

By making the above changes and keeping everything else in Model I unchanged, the system incurs no cost or reward from the first $k-1$ patients, but only from the waiting time of the kth patient at rate 1. Thus, in this case, $R_{0,0}^0(d_1)$ is the expected waiting time of the kth patient if he shows up.

To compute the expected server overtime, we need to set $r = c_w = 0$ and $c_l = -1$ in the original model. Then $R_{0,0}^0(d_1)$ is equal to the expected server overtime.

7. An Extension on the Interruption Time Distribution

Models I and II both assume that once the server is interrupted, it stays "off" for an exponentially distributed amount of time. In this section, we show how we can generalize our formulation so that the length of each "off" period has a phase-type distribution (see Fackrell 2009). To keep the presentation simpler and highlight one way of using this generalization, we focus on a specific phase-type distribution. However, generalization to any phase-type distribution would be similar.

Specifically, each "off" period is modeled as a continuous-time Markov chain with the state space

$\{0, 1, 2, \ldots, m\}$, where state 0 represents the absorbing state that indicates the end of an "off" period. The "off" period starts at state 1 and has the following rate matrix:

$$Q = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ \theta & -(\theta+\eta) & \eta & 0 & 0 & \cdots & 0 & 0 \\ 0 & \theta & -(\theta+\eta) & \eta & 0 & \cdots & 0 & 0 \\ 0 & 0 & \theta & -(\theta+\eta) & \eta & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & \theta & -\theta \end{bmatrix}.$$

The reason for choosing this particular matrix is that this transition rate matrix naturally arises if we assume that the service time of an individual emergency patient has an exponential distribution and emergency patients who find the server busy with another emergency patient join the emergency queue, which has some finite capacity of m. We can, in fact, choose a more general form for this matrix and thus allow the interruption time to have any phase-type distribution. In particular, we can easily generalize our analysis to the case where the rate θ becomes phase dependent, which would allow us to capture possible changes in service speed depending on the number of emergency patients waiting.

Then the length of the service "off" period has a phase-type distribution denoted by (α, M), where $\alpha = [1, 0, \ldots, 0]$, being an m-dimensional vector, and M is the submatrix of Q, corresponding to the states in $\{1, 2, \ldots, m\}$ (see Neuts 1981 for more on phase-type distributions). Define $\hat{X}$ as the effective service time. Its mean is given by the following proposition:

Proposition 1. *We have*

$$E(\hat{X}) = \frac{\theta(1-(\eta/\theta)^{m+1})}{\mu(\theta-\eta)},$$

where $\hat{X}$ denotes the effective service time for a random patient with a scheduled appointment.

When $m = 0$, which corresponds to the case where there are no interruptions, the expected effective service time simplifies to $E(\hat{X}) = 1/\mu$, the mean service time for a scheduled patient. On the other hand, when $m = 1$, the expression simplifies to (4), the expected effective service time when the interruption takes an exponentially distributed amount of time.

For each $k=0, 1, \ldots, N$, define $\hat{R}^k_{n,i}(t), 0<t \le d_{k+1}-d_k$, as the total expected net profit over $[d_{k+1}-t, \infty)$ if at time $d_{k+1}-t$ there are n scheduled patients in the system, and the server is in state i, where $n=0, 1, \ldots, k$, $i=0, 1, \ldots, m$. Also define

$$\hat{R}^k_n(t) = \begin{bmatrix} \hat{R}^k_{n,0}(t) \\ \hat{R}^k_{n,1}(t) \\ \vdots \\ \hat{R}^k_{n,m}(t) \end{bmatrix} \quad \text{and} \quad \frac{\mathrm{d}\hat{R}^k_n(t)}{\mathrm{d}t} = \begin{bmatrix} \dfrac{\mathrm{d}\hat{R}^k_{n,0}(t)}{\mathrm{d}t} \\ \dfrac{\mathrm{d}\hat{R}^k_{n,1}(t)}{\mathrm{d}t} \\ \vdots \\ \dfrac{\mathrm{d}\hat{R}^k_{n,m}(t)}{\mathrm{d}t} \end{bmatrix}.$$

Then we can state the generalized version of Theorem 1 as follows:

Theorem 4. *For each $k = 0, 1, \ldots, N$, the vector of the net profit functions $\hat{R}^k_n(t)$, $0 < t \le d_{k+1} - d_k$, satisfies the following differential equations*:

$$\frac{\mathrm{d}\hat{R}^k_0(t)}{\mathrm{d}t} = \begin{bmatrix} -\eta & \eta & 0 & 0 & \cdots & 0 & 0 \\ \theta & -(\eta+\theta) & \eta & 0 & \cdots & 0 & 0 \\ 0 & \theta & -(\eta+\theta) & \eta & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \theta & -\theta \end{bmatrix} \hat{R}^k_0(t),$$

and for $n = 1, \ldots, k$,

$$\frac{\mathrm{d}\hat{R}^k_n(t)}{\mathrm{d}t} = \begin{bmatrix} -(n-1)c_w \\ -nc_w \\ \vdots \\ -nc_w \end{bmatrix} + \begin{bmatrix} \mu & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & 0 \end{bmatrix} \hat{R}^k_{n-1}(t)$$

$$+ \begin{bmatrix} -(\eta+\mu) & \eta & 0 & 0 & \cdots & 0 & 0 \\ \theta & -(\eta+\theta) & \eta & 0 & \cdots & 0 & 0 \\ 0 & \theta & -(\eta+\theta) & \eta & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \theta & -\theta \end{bmatrix} \hat{R}^k_n(t).$$

The generalized boundary conditions are

$$\hat{R}^N_{0,i}(0^+) = 0, \quad i = 0, 1, \ldots, m,$$

$$\hat{R}^N_{n,0}(0^+) = -c_l n E(\hat{X}) - c_w\left[\frac{n(n+1)}{2}E(\hat{X}) - \frac{n}{\mu}\right], \quad n = 1, \ldots, N,$$

$$\hat{R}^N_{n,i}(0^+) = -c_l[-e_i M^{-1} e + nE(\hat{X})] - c_w\left[-ne_i M^{-1} e + \frac{n(n+1)}{2}E(\hat{X}) - \frac{n}{\mu}\right], \quad n = 1, \ldots, N,\ i = 1, \ldots, m,$$

$$\hat{R}^k_{n,i}(0^+) = p[r + \hat{R}^{k+1}_{n+1,i}(d_{k+2} - d_{k+1})] + (1-p)\hat{R}^{k+1}_{n,i}(d_{k+2} - d_{k+1}),$$
$$k = 0, 1, \ldots, N-1,\ n = 0, \ldots, k,\ i = 0, 1, \ldots, m,$$

where $e_i = [0, \ldots, 1, \ldots, 0]$ *with* 1 *being the ith element,* $e = [1, 1, \ldots, 1]^T$, *and* $-e_i M^{-1} e$ *being the mean length of the remaining "off" period when the service interruption process is in phase* i, $i = 1, 2, \ldots, m$.

The proof of Theorem 4 is very similar to that of Theorem 1 and hence is omitted from this paper for brevity. The system of differential equations stated in Theorem 4 can be solved using the methods introduced in §5. The algorithm to be used is very similar to that for the original formulation; therefore, we omit the details for brevity.

Note that if we use this generalization to formulate the queue of emergency patients as described, the queue capacity m should be finite but can be arbitrarily large. However, it is important to note that although the mathematical analysis does not change with m, the methods we propose become increasingly computationally expensive with larger m. In particular, the method of LT and integrating factor are both $O(m^3)$.

8. Numerical Results

In this section, we report the results of our numerical study. This study has two main objectives: first, to investigate the potential benefits of incorporating service interruptions when determining optimal appointment times, and second, to study how the main insights on optimal appointment-scheduling policies that have been reported in the literature change when service interruptions are considered. In our numerical study, when solving the optimization problems, we used a built-in function *fmincon* in Matlab with the interior-point algorithm option. It is important to note that we have identified instances of Model I for which the objective function has multiple local maxima and instances of Model II for which the objective function has multiple local minima. Therefore, there is no guarantee that the solutions that the *fmincon* function found are in fact globally optimal. To at least partially overcome this issue, for each instance of the problem, we used the *fmincon* function starting with various initial points. Each initial point was obtained by first randomly generating a vector of size N (the number of appointments to be scheduled) whose components take values between 0 and 1, and then multiplying this random vector by a scalar K, which is set to 0, 3, 6, 9, and 12 in turn. We then identified the locally optimal solution corresponding to each initial point and compared the objective function values at these local optima to determine the "best" solution, which we believe is very likely to be the global optimum. It is also important to note that this uncertainty on the global optimality of the solutions we obtained does not prevent us from generating insights regarding the importance of taking into account service interruptions, because the improvements we obtained are already significant, as we report in the following. Under the globally optimal solution, which is possibly different from what we obtained, the improvements can only be greater.

As we stated in §3, one of the desirable features of our formulation and the solution methods is that the interruption rate can be allowed to be time-dependent. More precisely, the arrival rate of interruptions can be a stepwise constant function. The way that this generalization is handled in our solution methods is somewhat tedious but straightforward. Specifically, we use the following procedure: For a given stepwise-constant interruption rate function, the problem horizon consists of a sequence of time intervals in which the interruption rate is constant. Because of this constant interruption rate, within each interval one can use the methods we developed in §5 with no changes. In a sense, the appointment scheduling over each interval can be seen as a separate problem in which the interruption rate is a constant. Clearly, however, the separated problems over these intervals are not independent of each other, but that can be taken care of by adding boundary conditions—which "transfer" the "accumulated" reward (cost) from one interval to the next—into the system of differential equations that describe the evolution of the expected net profit (cost) function.

Having the capability of handling time-dependent interruption rates is crucial because of its practical relevance. As we stated in §1, we are primarily motivated by interruptions caused by emergency patients, and empirical studies have consistently found that the arrival rates of emergency patients depend highly on the time of day. In our numerical study, although we considered constant interruption rates for Model I, we considered time-dependent rates for Model II. Note that one could easily carry out a time-dependent study for Model I as well.

8.1. Numerical Results for Model I

First, recall that T is the length of the service session during which all appointments should be scheduled, $1/\mu$ is the mean service time, $1/\theta$ is the mean duration for an interruption, p is the show-up probability, and c_w and c_l are the patient waiting costs per unit of time and the server overtime cost per unit of time beyond T, respectively. In our numerical study for Model I, we considered three different scenarios. For Scenario 1, we set $T = 8$, $\mu = 1$, $\theta = 0.5$, $p = 0.75$, $c_w = 1$, $c_l = 1$, and $r = 2$. For Scenario 2, we simply increased the overtime unit cost to $c_l = 2$, and for Scenario 3, we kept $c_l = 1$ but decreased the no-show probability to 0. Note that μ and θ are fixed in all three scenarios. However, although we do not report any details here, in our numerical study, we observed that the way the system costs change with μ and θ is as expected.

Table 1 **Numerical Results for Scenario 1**

η	R^*	R_{nointer}	R_{approx}	R_{eq}
0.00	7.9422 (8)	7.9422 (8)	7.9422 (8)	7.8584 (8)
0.10	3.8440 (5)	2.2876 (8)	3.6408 (6)	3.7363 (5)
0.15	2.7369 (4)	−0.4731 (8)	2.4316 (5)	2.6328 (4)
0.20	1.9999 (3)	−3.2036 (8)	1.6683 (4)	1.9328 (3)
0.25	1.4750 (2)	−5.9116 (8)	0.7449 (4)	1.4750 (2)
0.30	1.1951 (2)	−8.6030 (8)	0.6449 (3)	1.1951 (2)

Note. Numbers in parentheses indicate N^*, the optimal number of appointments to be scheduled in each setting.

Because the increase in either essentially makes the server faster, the optimal costs decrease if either of these two parameters increases.

For each scenario, we considered six different values for η, the arrival rate of interruptions: 0, 0.1, 0.15, 0.2, 0.25, and 0.3. For each instance of the problem, we first determined the optimal policy and the objective function value under the optimal policy, which we denote by R^*. In addition, we also determined the performance of the following policies: the policy that ignores interruptions; the policy that considers interruptions approximately by assuming that service times are exponentially distributed with mean adjusted to be equal to the mean effective service time given in (4); and the policy that considers the interruptions but has the restriction that the times between all consecutive appointments are the same (equally spaced appointments). In the following, we use R_{nointer}, R_{approx}, and R_{eq} to denote the value of the objective function under these three policies, respectively.

The results are given in Tables 1–3 for Scenarios 1–3, respectively. We can observe immediately from the three tables that completely ignoring interruptions can be quite costly, particularly when the interruption rate is high. It is important to note that in Model I, in addition to the appointment schedule, we also determine the total number of appointments to be scheduled. Ignoring interruptions clearly overestimates the number of appointments the system can reasonably handle and results in negative values for the objective function. (In the tables, the numbers in parentheses are the optimal number of appointments to be scheduled associated with each policy under each case.) Capturing the interruptions approximately by extending the mean service time appropriately seems to work reasonably well when the interruption rate is small, but for high interruption rates, the difference between the performance of the optimal policy R^* and the performance of the approximation R_{approx} is significantly greater. For brevity, we do not report the optimal schedule **d** here, but we observe that when the interruption rate is high the optimal policy does not have a dome shape. It has a monotone structure; more specifically, the time between two consecutive appointments is larger for appointments scheduled later in the day.

Table 2 **Numerical Results for Scenario 2**

η	R^*	R_{nointer}	R_{approx}	R_{eq}
0.00	7.0223 (7)	7.0223 (7)	7.0223 (7)	6.9340 (7)
0.10	3.2216 (4)	1.5562 (7)	3.0319 (5)	3.1439 (4)
0.15	2.2711 (3)	−1.1313 (7)	1.9768 (4)	2.2224 (3)
0.20	1.6406 (3)	−3.7985 (7)	0.8578 (4)	1.6205 (2)
0.25	1.3024 (2)	−6.4509 (7)	0.6503 (3)	1.3024 (2)
0.30	0.9838 (2)	−9.0928 (7)	0.0561 (3)	0.9838 (2)

Note. Numbers in parentheses indicate N^*, the optimal number of appointments to be scheduled in each setting.

Table 3 **Numerical Results for Scenario 3**

η	R^*	R_{nointer}	R_{approx}	R_{eq}
0.00	9.0101 (7)	9.0101 (7)	9.0101 (7)	8.9160 (7)
0.10	4.4608 (4)	2.1134 (7)	4.2718 (5)	4.3855 (4)
0.15	3.1885 (3)	−1.2710 (7)	3.0102 (4)	3.1344 (3)
0.20	2.3879 (3)	−4.6284 (7)	1.6520 (4)	2.3055 (3)
0.25	1.8552 (2)	−7.9668 (7)	1.3644 (3)	1.8552 (2)
0.30	1.4535 (2)	−11.2923 (7)	0.5314 (3)	1.4535 (2)

Note. Numbers in parentheses indicate N^*, the optimal number of appointments to be scheduled in each setting.

Finally, we observe that requiring the times between two consecutive appointments to be the same throughout the day does not degrade the performance significantly. Interestingly, the performance gap is smaller when the interruption rate is higher. This might be because regardless of whether or not one has the restriction, when there are frequent interruptions, the system will incur significant overtime costs, and thus the difference between any two policies will be small, as long as they both take interruptions into account and thus choose N the total number of appointments to be scheduled reasonably. However, it is also important to note that this relatively small difference between the two policies is likely to be caused partially by the fact that the interruptions occur at a constant rate throughout the day. The difference would likely be more significant when the interruption rate is time dependent, which we demonstrate for Model II in the next section.

8.2. Numerical Results for Model II

Studies on the arrivals of patients to emergency departments have found that the arrival rate function is typically such that the rate makes a single peak in the late morning or early afternoon (see Duguay and Chetouane 2007, McCarthy et al. 2008, Pitts et al. 2008) or makes two peaks, one during late morning hours or early afternoon and the other during late afternoon or early evening (Draeger 1992, Rossetti et al. 1999, Channouf et al. 2007). All studies find that the rate typically increases rapidly during the early morning hours and decreases rapidly starting with late evening.

Based on these findings, we considered two different emergency arrival rate (interruption rate) functions for our numerical study. In Scenario 1, the arrival rate function for emergency patients is given by $\eta(t) = 0.3$ for $t \in [0, 3)$, 0.5 for $t \in [3, 5)$, 0.4 for $t \in [5, 11)$, 0.2 for $t \in [11, 17)$, and 0.1 for $t \in [17, \infty)$. Thus, in Scenario 1, the interruption rate has a single peak. In Scenario 2, the interruption rate has two peaks. More specifically, $\eta(t) = 0.2$ for $t \in [0, 4)$, 0.5 for $t \in [4, 6)$, 0.3 for $t \in [6, 10)$, 0.4 for $t \in [10, 12)$, and 0.1 for $t \in [12, \infty)$. For both scenarios, we assumed that there were seven appointments to be scheduled and we chose $\mu = 1$, $\theta = 0.5$, and $p = 0.75$. We let $\gamma = c_s/(c_s + c_w)$, and for each scenario we varied it from 0.1 to 0.9.

Under each scenario, for each fixed value of γ, we determined the optimal schedule, the optimal schedule when interruptions are ignored, and the optimal schedule under the restriction that appointments are equally spaced. Figure 1 provides a visual description of the optimal appointment schedule for Scenario 1 with $\gamma = 0.5$. It also shows the optimal equally spaced schedule and the optimal schedule under the assumption that interruptions are ignored. Each curve was obtained by connecting the corresponding points (i, x_i), $i = 2, \ldots, 7$, where i is the appointment number, with appointment 1 being the first appointment of the day, and x_i is the time between the ith and the $(i-1)$th appointments. Note that the plots start with $i = 2$ because the first appointment is always scheduled for $t = 0$ in all cases.

The optimal schedule obtained by ignoring interruptions underestimates the load on the system. As a result, under this policy, appointments are scheduled close to each other. When interruptions are considered, appointments are scheduled more sparsely. On the other hand, the equally spaced schedule captures the interruption effect to a certain extent but does not respond to changes in interruption rate throughout the day. Because of that, the middle portion of the curve for the equally spaced schedule stays below the optimal schedule curve during the period when interruptions are most likely to happen. In the optimal schedule, appointments are more frequent early and late in the day and less frequent in the middle of the day. This is expected because in Scenario 1, the interruption rate is higher in the middle of the day. One would expect that when the interruption rate function has a different shape, the optimal policy would have a different structure as well. That is indeed the case. For example, one can easily find examples in which interruption rate functions are monotone in time of the day, and optimal appointment policies are also monotone (times between consecutive appointments increase or decrease throughout the day). Such interruption rate functions can be seen in hospitals on days that have special events such as football games, which are known to increase the demand for emergency response services. Thus, whether one would observe a dome-shaped structure for the optimal appointment schedule depends significantly on the shape of the interruption rate function.

We also evaluated the performance of each policy using our model where interruptions are present and computed the percentage improvement one gets by explicitly considering interruptions when scheduling appointments and the percentage improvement that one gets by not requiring equally spaced appointment times. (Note that, as in §8.1, when we find the optimal policy under the restriction that appointments are equally spaced, we do consider the interruption process so that the observed performance improvement results only by not requiring the times between the appointments to be the same.)

Figure 1 **Optimal Schedules Under Different Policies**

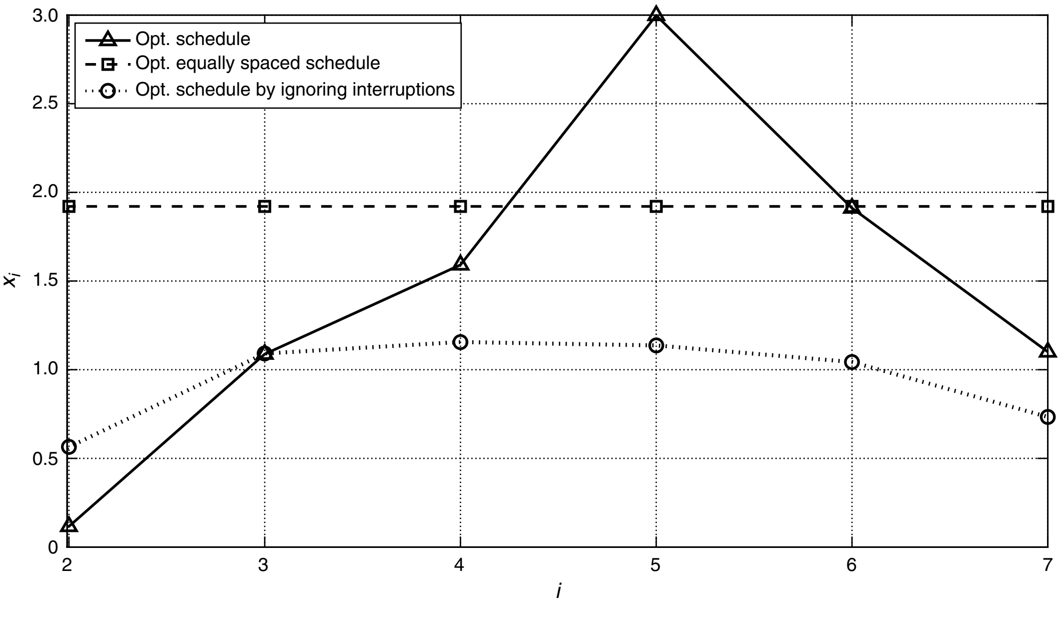

Table 4 Benefits of Considering Interruptions and Allowing Flexible Appointment Times Under Scenario 1

γ	Considering vs. ignoring interruptions (%)	Unconstrained vs. equally spaced appointments (%)
0.1	64.26	15.34
0.3	20.78	5.77
0.5	3.76	3.98
0.7	0.73	2.43
0.9	0.17	1.19

Tables 4 and 5 present the results for Scenarios 1 and 2, respectively. We can immediately observe from the first column in both tables that the "cost" of ignoring interruptions could be significant, particularly when patient waiting cost is high. When $\gamma = 0.1$, the benefit from modeling the interruptions can be more than 64%. When the waiting cost is small, the interruptions are much less of a concern. Even if patients end up waiting for a long time because of interruptions, that does not affect the objective function in a significant way. As a result, explicit consideration of the interruption process does not gain us much. Note that the improvements are not as dramatic as those in Model I, because N is not a decision variable in Model II.

Looking at the second column in both tables, we observe the percentage improvement one would get by allowing appointments to be scheduled at any time, as opposed to requiring them to be scheduled at equally spaced time points. We observe that there are modest improvements in all cases (approximately 15% when waiting cost is high). It is difficult to make a strong statement as to whether these improvements are large enough to not recommend equally spaced appointments, because such simple appointment-scheduling policies might have some additional advantages—such as ease of implementation—that are not captured in our formulation. Nevertheless, the improvements are clearly larger than when there are no interruptions (see Hassin and Mendel 2008). This is particularly the case when the interruption rate changes with time as in our numerical study. (Although we do not report it here, we found that when the interruption rate is constant, the improvements are less significant.) This is not surprising, because in that case, one can see that there could potentially be more benefits in asking more patients to come when the interruption probability is low, resulting in more frequent appointments during certain times of the day. The benefits would potentially be higher when the variations in the interruption rate throughout the day are more significant.

Table 5 Benefits of Considering Interruptions and Allowing Flexible Appointment Times Under Scenario 2

γ	Considering vs. ignoring interruptions (%)	Unconstrained vs. equally spaced appointments (%)
0.1	50.17	13.9
0.3	16.12	7.21
0.5	2.37	3.98
0.7	0.88	2.25
0.9	0.14	1.26

8.3. Simulation Study: Systems with Nonexponential Service and Interruption Times

So far, we assumed that the service times and the interruption times are exponentially distributed. Although, as we discussed in §7, one can use phase-type distributions for the interruption times and a similar generalization can be made for service times as well, assuming exponential distribution for service and interruption times significantly simplifies computational requirements. However, empirical studies have mostly found that service times typically follow a lognormal—not exponential—distribution (see, e.g., Cayirli and Veral 2003, Klassen and Yoogalingam 2008). It is thus important to investigate how the appointment-scheduling policies that are obtained through our mathematical models would perform in settings where the exponential distribution assumption does not hold.

Specifically, we consider Model II and assume that there are seven appointments to be scheduled under the restriction that the times between the appointments are the same. The only reason we concentrate on the problem with equally spaced appointment intervals is that with this restriction there is only one decision variable, the time between two consecutive appointments; this makes simulation optimization relatively a more viable option. Without the restriction, there would be six decision variables, and finding the optimal values for these variables using simulation would be computationally expensive. In the simulation model, we assumed that both the service times and interruption times have lognormal distribution with mean 2. We considered four different values for the coefficient of variation (CV) (0.6, 0.8, 1.0, and 1.2). We also considered three different values for the arrival rate of interruptions (0.1, 0.15, and 0.2) and three different values for the cost ratio $\gamma = c_s/(c_s + c_w)$ (0.1, 0.5, and 0.9). All these different choices for the three parameters resulted in 36 different scenarios. In all the scenarios, we assumed that the no-show probability is zero.

For each scenario, we used a relatively primitive simulation optimization method that uses line search over a discrete set of values ($\{0.05, 0.10, \ldots, 11.00\}$) for d, the time between two consecutive appointments. With each d, we ran 100,000 independent replications to obtain the mean cost and its associated 95% confidence interval. We then identified the "optimal" policy (d) as the one under which the mean cost is the smallest. In the following, we use $C_{\log}$ to denote the mean cost under this policy. It is important to note that the actual "optimal" policy is possibly different from the one we obtained here, because our study has not established optimality at some statistically significant level, even among the set of discrete choices. It appears that significantly more replications are needed to conclude any policy as being "optimal" at some statistically significant level. However, it is clear that the performance of the policy we obtained would be very close to that of the actual optimal policy and thus would not change the conclusions we reach.

After determining the "optimal" policy and the mean cost under this policy, we then identified the policy that would be optimal if the service times and interruption times were exponentially distributed with the same mean as in the simulation model. We then used this policy in the simulation model to determine how that policy would perform in the lognormal setting. In the following, we use $C_{\exp}$ to denote the mean cost (obtained via simulation) under the policy that is optimal for the system with exponentially distributed service and interruption times.

The results are summarized in Table 6. We can observe that the policies obtained by assuming exponential distributions perform quite well. In many cases, the performance difference is less than 1%, although it is closer to 5% when the coefficient of variation is 0.6. The performance difference is smallest when the coefficient of variation is high (1 and 1.2). This may not be surprising, given that the exponential distribution has a coefficient of variation of 1. Thus, our findings suggest that the exact shape of the distribution may not be very important in predicting how well the policies obtained using our formulation would perform, but their performance is likely to be better when the coefficient of variation is high.

Table 6 **Simulation Results**

CV	γ	η	$C_{\log}$ and 95% confidence interval	$C_{\exp}$ and 95% confidence interval	Percentage difference in the mean
0.6	0.1	0.10	8.0962 [8.0637, 8.1288]	8.4957 [8.4711, 8.5203]	4.93
		0.15	10.3303 [10.2934, 10.3672]	10.7668 [10.7376, 10.7960]	4.23
		0.20	12.4425 [12.4033, 12.4817]	12.9300 [12.8971, 12.9629]	3.92
	0.5	0.10	15.7692 [15.7274, 15.8110]	15.8562 [15.8178, 15.8946]	0.55
		0.15	18.2025 [18.1560, 18.2491]	18.2740 [18.2293, 18.3188]	0.39
		0.20	20.5166 [20.4617, 20.5714]	20.5306 [20.4810, 20.5801]	0.07
	0.9	0.10	17.7846 [17.7511, 17.8180]	18.0671 [18.0311, 18.1032]	1.59
		0.15	19.5262 [19.4873, 19.5651]	19.9070 [19.8656, 19.9484]	1.95
		0.20	21.2808 [21.2370, 21.3247]	21.6873 [21.6412, 21.7334]	1.91
0.8	0.1	0.10	9.1742 [9.1292, 9.2191]	9.3187 [9.2802, 9.3573]	1.58
		0.15	11.5924 [11.5440, 11.6409]	11.7027 [11.6587, 11.7468]	0.95
		0.20	13.8340 [13.7785, 13.8895]	14.0024 [13.9524, 14.0524]	1.22
	0.5	0.10	17.1306 [17.0735, 17.1878]	17.1850 [17.1313, 17.2388]	0.32
		0.15	19.6467 [19.5807, 19.7128]	19.7675 [19.7051, 19.8298]	0.61
		0.20	22.1568 [22.0856, 22.2280]	22.1973 [22.1293, 22.2652]	0.18
	0.9	0.10	18.2144 [18.1714, 18.2573]	18.3201 [18.2753, 18.3648]	0.58
		0.15	19.9871 [19.9385, 20.0357]	20.1175 [20.0668, 20.1683]	0.65
		0.20	21.7668 [21.7124, 21.8213]	21.8974 [21.8411, 21.9536]	0.60
1.0	0.1	0.10	10.3574 [10.2963, 10.4184]	10.3638 [10.3056, 10.4220]	0.06
		0.15	12.9443 [12.8761, 13.0126]	13.0213 [12.9532, 13.0894]	0.59
		0.20	15.3906 [15.3208, 15.4604]	15.5055 [15.4290, 15.5819]	0.75
	0.5	0.10	18.4510 [18.3753, 18.5268]	18.4869 [18.4163, 18.5574]	0.19
		0.15	21.1498 [21.0658, 21.2338]	21.1985 [21.1175, 21.2795]	0.23
		0.20	23.7796 [23.6906, 23.8687]	23.8111 [23.7213, 23.9009]	0.13
	0.9	0.10	18.5569 [18.5030, 18.6107]	18.5782 [18.5243, 18.6321]	0.11
		0.15	20.3532 [20.2937, 20.4128]	20.3653 [20.3050, 20.4257]	0.06
		0.20	22.1373 [22.0712, 22.2033]	22.1529 [22.0860, 22.2199]	0.07
1.2	0.1	0.10	11.5939 [11.5125, 11.6753]	11.6645 [11.5789, 11.7502]	0.61
		0.15	14.3977 [14.3076, 14.4877]	14.5031 [14.4068, 14.5994]	0.73
		0.20	17.0908 [16.9917, 17.1900]	17.1819 [17.0753, 17.2884]	0.53
	0.5	0.10	19.6490 [19.5529, 19.7451]	19.7608 [19.6704, 19.8513]	0.57
		0.15	22.5548 [22.4485, 22.6611]	22.5992 [22.4971, 22.7012]	0.20
		0.20	25.2419 [25.1279, 25.3558]	25.4963 [25.3793, 25.6134]	1.01
	0.9	0.10	18.8049 [18.7421, 18.8677]	18.8189 [18.7555, 18.8823]	0.07
		0.15	20.5560 [20.4852, 20.6268]	20.6123 [20.5416, 20.6830]	0.27
		0.20	22.4270 [22.3498, 22.5042]	22.4606 [22.3819, 22.5392]	0.15

When service and interruption times are not exponentially distributed, for a given appointment schedule, simulation of an appointment system would most likely give a more reliable estimate on the mean performance as opposed to using our numerical methods that assume exponential distributions. However, simulation is very inefficient when it comes to identifying the "optimal" policy. When there is one single decision variable, as in the simulation study we conducted in this section, simulation could be a reasonable choice. However, when the times between appointments are not restricted to be the same and/or when the number of appointments to be scheduled is also a decision variable, there are so many different policies to compare that finding the "optimal" policy by simulation is impractical. Even if one is interested in using simulation optimization, our numerical methods provide a fast way of obtaining a good policy that can be served as a good starting point, and they help in making the process much more efficient. Furthermore, our methods would help in carrying out a quick what-if analysis and providing insights into various research questions of interest (as we did in §§8.1 and 8.2) in an efficient manner. Finally, incorporation of various constraints into the optimization problem, such as putting a bound on the expected patient waiting time, is much more straightforward and efficient when using our numerical methods than when using simulation optimization.

9. Concluding Remarks

The vast literature on appointment scheduling has provided both various methods to schedule appointments and valuable insights into the type of policies that should be expected to perform well. However, the possibility that the service of scheduled patients can be interrupted has largely been ignored. In some settings, such as outpatient clinics that are located outside of the hospitals, it might be reasonable to ignore interruptions if they tend to be short in duration or are so rare that it is difficult to predict when and why they would occur. However, there are many appointment-based services within hospitals (such as electronic imaging machines) for which interruptions are not only common but can also be predicted to a certain extent based on past data. For example, the rate of patient traffic to MRI or CT scan machines from the emergency departments, which interrupts scheduled services, can easily be determined as a function of the time of day. This paper focuses on appointment scheduling for systems for which service interruption is a regular phenomenon.

One of the major contributions of this paper is that it develops a general framework that can be used in the analysis of various appointment-scheduling models that make different assumptions regarding performance measures and decision variables. We have concentrated on only two formulations, but one can easily come up with alternative models that can be analyzed in almost the same way. Introducing the possibility that services can be interrupted brings significant difficulty into mathematical modeling and analysis. One of the main challenges is to come up with an expression for the objective function. We overcome this difficulty by first writing the total accumulated net profit after T in Model I, or the total accumulated cost after the appointment time of the last patient in Model II, and going backward in time, appointment by appointment, until time zero, the beginning of the day. This approach gives us a system of differential equations whose solution provides the objective function value for a given appointment schedule. The solution to the differential equations is not readily available, but we propose two different solution methods, either of which can be used to obtain the solution and determine the objective function value.

Having a formulation and a method for determining the optimal policy is important for two main reasons. First, they can be used in practice if model assumptions are believed to fit reasonably well with the practical setting considered. Second, they can be used to solve various problem instances in order to obtain some general insights into scheduling policies that perform well when interruptions are present. Part of this paper is devoted to this second potential use of our framework. As a result of this analysis, we made a number of observations: We found that ignoring interruptions can lead to policies that perform very poorly, especially when the number of appointments to be scheduled on a given day is also a decision variable. One way of considering interruptions approximately could be by adjusting the mean service time appropriately in the model that ignores interruptions. However, that approach appears to work well only when the interruption rate is small. These two observations point to the importance of explicit formulation of the interruption process. We also observed that policies that require the time between consecutive appointments to be the same have a decent performance when the interruption rate is constant, but their performances worsen when the interruption rate is time dependent. This suggests, for example, that when scheduling appointments for electronic imaging machines that are shared by emergency patients, it might be worthwhile to drop the convenience of having equally spaced intervals and distribute appointments over time so that there are fewer scheduled appointments around times when the arrival rate of emergency patients typically peaks.

An important assumption made in our models is that interruptions are preemptive. This would be a reasonable assumption in cases where interrupting a regular service is practically possible (for example, MRI machines), but there are also settings in which preemption may not be an option (e.g., surgeries). Thus, a potentially useful direction for future research is the analysis of appointment systems with nonpreemptive interruptions.

Electronic Companion

An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/msom.1120.0394.

Acknowledgments

The work of the third author was supported by the National Science Foundation [Grant CMMI-0927607].

References

Adiri I, Bruno J, Frostig E, Rinnooy Kan AHG (1989) Single machine flow-time scheduling with a single breakdown. *Acta Informatica* 26(7):679–696.

Alderman L (2011) The doctor will see you . . . eventually. *New York Times* (August 1), http://www.nytimes.com/2011/08/02/health/policy/02consumer.html.

Anderson C, Butcher C, Moreno A (2010) Emergency department patient flow simulation at HealthAlliance. Project proposal, Worcester Polytechnic Institute, Worcester, MA.

Bailey NTJ (1952) A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *J. Royal Statist. Soc. Series B* (*Methodological*) 185–199.

Birge J, Frenk JBG, Mittenthal J, Rinnooy Kan AHG (1990) Single-machine scheduling subject to stochastic breakdowns. *Naval Res. Logist.* 37(5):661–677.

Broder J, Warshauer DM (2006) Increasing utilization of computed tomography in the adult emergency department, 2000–2005. *Emergency Radiology* 13(1):25–30.

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

Chakraborty S, Muthuraman K, Lawley M (2010) Sequential clinical scheduling with patient no-shows and general service time distributions. *IIE Trans.* 42(5):354–366.

Channouf N, L'Ecuyer P, Ingolfsson A, Avramidis AN (2007) The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Management Sci.* 10(1):25–45.

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

Draeger MA (1992) An emergency department simulation model used to evaluate alternative nurse staffing and patient population scenarios. *Proc. 24th Conf. Winter Simulation* (ACM, New York), 1057–1064.

Duguay C, Chetouane F (2007) Modeling and improving emergency department systems using discrete event simulation. *Simulation* 83(4):311–320.

Fackrell M (2009) Modelling healthcare systems with phase-type distributions. *Health Care Management Sci.* 12(1):11–26.

Federgruen A, Green L (1986) Queueing systems with service interruptions. *Oper. Res.* 34(5):752–768.

Fiems D, Koole G, Nain P (2012) Waiting times of scheduled patients in the presence of emergency requests. Working paper, VU University Amsterdam, Amsterdam. Accessed June 20, 2012, http://www.math.vu.nl/∼koole/articles/2005report1/art.pdf.

Fries BE, Marathe VP (1981) Determination of optimal variable-sized multiple-block appointment systems. *Oper. Res.* 29(2): 324–345.

Glazebrook KD (1984) Scheduling stochastic jobs on a single machine subject to breakdowns. *Naval Res. Logist. Quart.* 31(2): 251–264.

Gray WJ, Wang PP, Scott MK (2000) A vacation queueing model with service breakdowns. *Appl. Math. Model.* 24(5–6):391–400.

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

Green LV, Savin S, Wang B (2006) Managing patient service in a diagnostic medical facility. *Oper. Res.* 54(1):11–25.

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

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

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

Horowitz E (1971) Algorithms for partial fraction decomposition and rational function integration. *Proc. Second ACM Sympos. Symbolic and Algebraic Manipulation* (ACM, New York), 441–457.

Jouini O, Benjaafar S (2012) Queueing systems with appointment-driven arrivals, non-punctual customers, and no-shows. Working paper, École Centrale Paris, Châtenay-Malabry, France. Accessed June 20, 2012, http://www.isye.umn.edu/faculty/pdf/jobe-5-17-10.pdf.

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

Kenny DJ, Barrett EJ (2005) Emergency trauma: Treating the unexpected. *J. Calif. Dental Assoc.* 33(5):383–386.

Klassen KJ, Yoogalingam R (2008) An assessment of the interruption level of doctors in outpatient appointment scheduling. *Oper. Management Res.* 1(2):95–102.

Korley FK, Pham JC, Kirsch TD (2010) Use of advanced radiology during visits to US emergency departments for injury-related conditions, 1998–2007. *J. Amer. Medical Assoc.* 304(13): 1465–1471.

Liu N, Ziya S, Kulkarni VG (2010) Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing Service Oper. Management* 12(2):347–364.

McCarthy ML, Zeger SL, Ding R, Aronsky D, Hoot NR, Kelen GD (2008) The challenge of predicting demand for emergency department services. *Academic Emergency Medicine* 15(4): 337–346.

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

Neuts MF (1981) *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach* (Johns Hopkins University Press, Baltimore).

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

Pitts SR, Niska RW, Xu J, Burt CW (2008) National hospital ambulatory medical care survey: 2006 emergency department summary. *National Health Statist. Rep.* 7(7):1–38.

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

Rossetti MD, Trzcinski GF, Syverud SA (1999) Emergency department simulation and determination of optimal attending physician staffing schedules. *Proc. 31st Conf. Winter Simulation* (ACM, New York), 1532–1540.

Stein WE, Côté MJ (1994) Scheduling arrivals to a queue. *Comput. Oper. Res.* 21(6):607–614.

Takine T, Sengupta B (1997) A single server queue with service interruptions. *Queueing Systems* 26(3):285–300.

Vasanawala SS, Desser TS (2005) Accommodation of requests for emergency US and CT: Applications of queueing theory to scheduling of urgent studies. *Radiology* 235(1):244–249.

Wang PP (1994) Releasing N jobs to an unreliable machine. *Comput. Indust. Engrg.* 26(4):661–671.

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

Wang PP (1999) Sequencing and scheduling N customers for a stochastic server. *Eur. J. Oper. Res.* 119(3):729–738.