Comparison of Traditional and Open-Access Appointment Scheduling for Exponentially Distributed Service Time

Chongjun Yan[1], PhD; Jiafu Tang[1*], PhD; Bowen Jiang[2], PhD and Richard Y.K. Fung[3], PhD

[1]*College of Management Science & Engineering, Dongbei University of Finance and Economic, Dalian, China.*

[2]*Department of Systems Engineering, Northeastern University, Shenyang, China.*

[c]*Department of System Engineering & Engineering Management, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China.*

Submitted June 2014. Accepted for publication April 2015.

ABSTRACT

This paper compares the performance measures of traditional appointment scheduling (AS) with those of an open-access appointment scheduling (OA-AS) system with exponentially distributed service time. A queueing model is formulated for the traditional AS system with no-show probability. The OA-AS models assume that all patients who call before the session begins will show up for the appointment on time. Two types of OA-AS systems are considered: with a same-session policy and with a same-or-next-session policy. Numerical results indicate that the superiority of OA-AS systems is not as obvious as those under deterministic scenarios. The same-session system has a threshold of relative waiting cost, after which the traditional system always has higher total costs, and the same-or-next-session system is always preferable, except when the no-show probability or the weight of patients' waiting is low. It is concluded that open-access policies can be viewed as alternative approaches to mitigate the negative effects of no-show patients.

Keywords: healthcare management, appointment scheduling, no-show

1. INTRODUCTION

1.1. Background

With the rapid development of the healthcare industry, how to reduce operational costs and improve service quality is becoming an essential task for service providers in outpatient services [1]. It is common for patients to book an appointment in advance. It has made the appointment scheduling (AS) system an active research issue over the past six decades (Cayirli and Veral [2], Gupta and Denton [3], LaGanga [4],

*Corresponding author: Prof. Jiafu Tang, College of Management Science & Engineering, Dongbei University of Finance and Economic, Shahekou, Dalian, 116026, China. Phone: (86) 0411-84713593. Fax: (86) 0411-84710475. Email: jftang@mail.neu.edu.cn. Other authors: yanchongjun0607@163.com, ddxx100@sina.cn, meykfung@cityu.edu.hk.

Rinder et al. [5]). A well-designed AS system can achieve a balance between the efficiency of service providers and the satisfaction of patients. The objective of AS is to find an effective policy to optimize its performance measures, e.g., the waiting time, idle time and overtime. Two widely used AS systems are the traditional AS system and the open-access appointment scheduling (OA-AS) system [6]. In the traditional AS system, a patient who books an appointment in advance may not show up. On the other hand, a random number of patients call to make appointments at the beginning of the session, and all patients are served during that session or the next session in the OA-AS system.

In the traditional AS system, a considerable number of patients who have made an appointment do not show up. This phenomenon cannot be ignored because it results in idle time of physicians and other expensive physical assets and equipment in some intervals. Clinics usually overbook additional patients in expectation of patients' no-shows, and it may lead to undue congestion and dissatisfaction of patients. This situation indicates that no-shows waste substantial operational costs and resources in outpatient AS systems [7]. Although Huang and Zuniga [8] showed that the policies allowing cancellation up to several hours in advance could reduce the negative impact of no-shows, these approaches could not entirely eliminate no-shows for clinics. Most practitioners resort to overbooking or compressing the appointment intervals to offset no-shows. However, it leads to longer waiting times for patients. Hence, a reasonable tradeoff decision is needed to balance the interests of patients and those of physicians.

The so-called OA-AS system was introduced by Murray and Tantau [9] as an alternative approach to reduce no-show probabilities and improve patients' satisfaction. The target of the OA-AS system is to accept appointments on the calling session or the session after rather than keeping the sick waiting for several days before the appointment time. The OA-AS systems have received more attention from both practitioners and academics. Two widely used OA-AS policies are investigated, i.e., same-session policy and same-or-next-session policy, depending on whether the patients can be postponed to the next session. Under the same-session policy, all patients who call before the session begins will be scheduled to the same session. On the contrary, some of the patients have to wait until the next session in order to smooth the arrival rate fluctuation under the same-or-next-session policy. There are both successful and failure stories reported about the OA-AS systems. Therefore, quantitative comparison is needed to help clinic administrators answer some essential questions, such as how effective the OA-AS systems are compared to the traditional AS system, what the most influential factor is on the performance measures of OA-AS systems, and how these policies should be better implemented. For this purpose, Robinson and Chen [10] compare the traditional AS and the OA-AS systems under deterministic service time. Their performance measures include the waiting cost, the idle time cost during working hours, and the overtime cost to serve the remaining patients beyond the normal working hours. The results show that the open access policies outperform the traditional one substantially, but they also conclude that the differences may be less significant if other types of variability are exploited.

1.2. Objectives and Significance of the Study

By extending the work of Robinson and Chen [10], the goal of this paper is to make a comparison of the performance measures of AS and OA-AS under exponentially distributed service time. Both systems aim to minimize the total weighted costs of waiting time of patients, idle time and overtime of the healthcare systems. To this end, a queueing model is first formulated in this paper for the traditional AS system with a no-show probability, and the stochastic programming model under a no-golf policy is also presented. By no-golf policy, it means the physician is not allowed to go off duty even if he has no patients waiting. This research extends the work by Hassin and Mendel [11] to consider overtime cost [12]. Thereafter, the mathematical models for the OA-AS systems under the same-session policy and same-or-next-session policy are developed. To facilitate the comparison, it is assumed that all patients who call before the session begins will show up for the appointment on time in the OA-AS queueing models, while the expected workload remains the same as the traditional AS model. The comparison of system performance between the traditional AS and the OA-AS is made numerically. The effects of factors including session-length, expected work-load, ratios of overtime cost and no-show probability on system performance measures are examined.

This is the first paper to compare different AS systems under exponentially distributed service time. The results indicate that the superiority of open-access policies is not as explicit as those under deterministic assumptions. The same-session AS system outperforms the traditional AS system in terms of the total costs beyond a threshold of relative waiting cost in most of the cases considered. The same-or-next-session system is always preferable, except when the no-show probability or weight of the patients waiting is low.

1.3. Literature Review on Appointment Scheduling

The components involved in the AS queueing systems are the arrival process (distribution, punctual arrival and no-show), the service process (number of servers, service rules and service time), patients' and providers' preferences, incentives and performance measures criteria (Gupta and Denton [3]). The basic factors considered in optimizing AS systems are the number of servers, deterministic or stochastic service time and the no-show probability. Cayirli and Veral [2] performed a comprehensive review of environmental factors and performance measures that have been investigated in previous studies and categorized solution methodologies and decision variables in the design of an appointment system. The most relevant studies including the modeling approaches of different appointment systems and methods to mitigate the impact of no-shows are summarized as follows.

From the perspective of the modeling approach, analytical studies based on queuing models (Creemers and Lambrecht [13, 14]) are used to obtain the steady-state distribution of performance. Researchers also pay attention to provide a better representative of outpatient queue in a fixed length of a day (Klassen and Yoogalingam [15], Kong et al. [16], De Vuyst et al. [17], Mak et al. [18]). Meanwhile, some researchers developed quantitative models for OA-AS systems (Qu and Shi [19], Wang and Gupta [20], Huang et al. [21]).

There is a significant body of evidence that illustrates how no-show becomes a crucial factor decreasing the performance of healthcare services all over the world. Cayirli et al. [22] report that the average no-show probability amounts to 38% and varies from 0% for colonoscopies to 67% for pediatric neurology. The first to address the appointment system with no-show is Mercer [23], who provides an analytical approach based on queueing theory. Previous studies on appointment systems with no-shows fall into three categories: 1) estimating the no-show probability and the social characteristics of such patients based on empirical data, e.g., Cayirli et al. [22], Kim and Giachetti [24]; 2) analyzing the impact of no-show patients on the performance measures of outpatient clinics, e.g., Green and Savin [25], Cayirli et al. [26]; and 3) proposing a series of recommendations to mitigate the negative impacts of no-show, e.g., Erdogan and Denton [27], Chakraborty et al. [28, 29], Berg et al. [30]. Hassin and Mendel [11] found the optimal schedule in terms of a weighted sum of patients' waiting time and doctors' working hours for exponentially distributed service times without consideration of overtime; the impact of no-show on the system performance was also investigated. Finally, attention is moved to the situation of equal-spaced interval lengths. Tang et al. [12] extended a previous work by integrating overtime cost into the numerical optimal solution. Zeng et al. [31] studied the clinic overbooking problem for patients with heterogeneous no-show probabilities, identified the properties of optimal schedule and designed a heuristic search algorithm to yield a local optimal solution. Begen and Queyranne [32] determined an optimal appointment schedule for a given sequence of patients with the objective of minimizing the expected total operational costs for discrete distribution service time. Their model can be easily adjusted to handle overtime, no-show and emergencies as well. Dome pattern is introduced and demonstrated as an effective rule in many situations [33, 34]. Klassen and Yoogalingam [15, 35, 36] showed that the dome and plateau dome pattern perform quite well even if both service interruption and doctor lateness are considered at the same time. Cayirli et al. [37] formulated a dome appointment rule as a function of the environmental factors and used simulation and nonlinear regression to derive the planning constant by parameterization. The subsequent procedure explicitly minimizes the negative impacts of no-shows and walk-ins. All these papers propose approaches to mitigate no-show effects under certain conditions by adjusting the interval length allocated to the patients or overbooking more than one patient in the same interval.

There are examples of successful application of open-access policy in the literature using short lead time to avoid the wasted capacity caused by no-shows. Kopach et al. [38] apply discrete simulation to investigate the impacts of four different parameters, including the lowest percentage of open appointments, the time horizon for fixed appointments, the provider care group and the overbooking horizon, on the performance of an open-access AS system in terms of continuity of care provided to patients under various settings. Green and Savin [25] proposed a single-server queueing system to analyze the AS system with state-dependent no-show probability, identifying the patient panel size for the clinic implementing the open-access policy. Liu et al. [39] proposed a heuristic procedure that outperforms all other benchmark

policies with no-shows and late cancellations when the workload is high. Their simulation results indicate that the open-access policy is appropriate when the patient load is relatively low. Dobson et al. [40] reserved time intervals for urgent patients in order to maximize the revenue of a physician with two service quality measures, i.e., the average number of urgent patients who are not handled during normal hours and the queue length of routine patients. Wang and Gupta [20] designed a new outpatient AS system with patient choice, matching the random arrivals of appointment requests with doctors' capacity reservation in a way that maximizes the clinics' revenue and patients' satisfaction simultaneously. LaGanga and Lawrence [7] proposed a gradient search method to find quality solutions for joint capacity and scheduling problems considering no-show patients based on the submodularity of the objective function. Qu et al. [41] develop a mean-variance model and an efficient solution approach to determine the Pareto-optimal open appointment percentages, increasing the average number of patients seen in a session while reducing the uncertainty caused by no-shows. Over a wide range of scenarios and clinics, Patrick [42] demonstrated that the MDP methods perform as well as or better than the OA-AS policy in terms of the clinic revenue. Lee et al. [43] compared the open access policy and the overbooking policy under some commonly used rules by simulation, but optimization is not applied in the comparison. The current literature survey shows that there is a need to compare these traditional methods with the open-access policy in terms of commonly used performance measures in order to determine the condition under which open-access performs more effectively than the dome pattern.

2. METHODS

2.1. Assumption and Notations for Traditional AS System

The prime objective of most of the traditional AS model is to search for an optimal schedule to minimize the weighted sum of patients' waiting cost and doctors' idle cost and overtime cost. Common medical practice allows doctors to have their own waiting lists, as it can provide a one-on-one doctor-patient experience. Therefore, the appointment system is formulated as a single-server queueing system with scheduled arrivals. The number of patients to be scheduled in a working session is determined in advance by the statistics of historical operation data, and unscheduled arrivals are not taken into consideration [10]. It is assumed that each patient in the queue shows up punctually with a probability. The no-show probability is the same for all patients and is obtained by statistics, and the arrival processes of individual patients are independent of one another. The service time of patients follows an independent and identical exponential distribution (Kopach et al. [38] justify this assumption by simulation). Patients are served in the order of their scheduled appointments. If some patients are scheduled to arrive simultaneously, the one with a lower scheduled position in line is served first if he/she arrives. Although the working hours of a doctor are planned in advance by the clinic, he/she can leave the office only after seeing all patients that show up. Hence, the doctor has to be available until the last scheduled appointment. The notations for the traditional AS system are the following:

c_i: Doctor's idle cost per unit time (dollar/min).
c_o: Doctor's overtime cost per unit time (dollar/min).
c_w: Patients' waiting cost per unit time (dollar/min).
I: Expected idle time (min) that a doctor wastes during a standard working session.
N: Number of patients to be scheduled in a standard working session.
N_i: Number of patients in the queue just before the arrival time of the ith scheduled patient.
O: Expected overtime (min) of the doctor in a standard working session.
p: No-show probability of each patient.
T: The predetermined length (min) of a standard working session.
t_i: Time (min) of the ith scheduled arrival, $t_i = \sum_{j=1}^{i-1} x_j$.
w_i: Expected waiting time (min) of the ith patient if he/she shows up, $w_1 = 0$.
W: Expected waiting time (min) of all patients scheduled in a session if a patient shows up, $W = \sum_{i=1}^{N} w_i$.
x: A vector of scheduled intervals, $x = (x_1, x_2, \ldots, x_{N-1})$.
x_i: Time interval (min) between the ith and $(i + 1)$th scheduled patient.
$\alpha = c_w/c_i$: The relative cost of patients' waiting, as a fraction of the idle time cost per unit time (min).
$\beta = c_o/c_i$: The relative cost of doctor's overtime, as a fraction of the idle time cost per unit time (min).
$1/\mu$: Mean of exponentially distributed service time.

By definition, t_i, time of the ith scheduled arrival, is the function of schedule x, vector of scheduled intervals; the following subsection shows that variables w_i (expected waiting time of the ith patient if he shows up), N_i (number of patients in the queue just before the arrival time of the ith scheduled patient), W (expected waiting time of all patients scheduled in a session if a patient shows up), O (expected overtime of the doctor in a standard working session), I (expected idle time that a doctor wastes during a standard working session) are all functions of schedule x.

2.2. The General Model for Traditional AS System

The objective is to find an optimal schedule x that can minimize the weighted sum of expected total patients' waiting time valued at α (relative cost of patients' waiting), the doctor's idle time standardized at 1, and the doctor's overtime valued at β (relative cost of doctor's overtime):

$$C(x) = c_w(1-p)W + c_i I + c_o O = c_i\left(\alpha(1-p)W + I + \beta O\right) \tag{1}$$

The expected waiting time of a show-up patient in the queue is determined by the number of patients upon his/her arrival [11]:

$$w_i = \sum_{j=1}^{i-1}\left(\frac{j}{\mu}\cdot \Pr\{N_i = j\}\right), i \geq 2 \tag{2}$$

The probability that j patients remain in the system just before t_i, denoted by $\Pr\{N_i = j\}$, is derived recursively from system state just before $t_{i-1}(i \geq 2)$:

For $1 \leq j < i$,

$$\begin{aligned}\Pr\{N_i = j\} = (1-p)\sum_{k=0}^{i-j-1}\Pr\{N_{i-1} = j+k-1\}\frac{(\mu x_{i-1})^k}{k!}e^{-\mu x_{i-1}} \\ + p\sum_{k=0}^{i-j-2}\Pr\{N_{i-1} = j+k\}\frac{(\mu x_{i-1})^k}{k!}e^{-\mu x_{i-1}}\end{aligned} \tag{3}$$

$$\begin{aligned}\Pr\{N_i = 0\} = (1-p)\sum_{k=1}^{i-1}\Pr\{N_{i-1} = k-1\}\sum_{l=k}^{\infty}\frac{(\mu x_{i-1})^l}{l!}e^{-\mu x_{i-1}} \\ + p\sum_{k=0}^{i-2}\Pr\{N_{i-1} = k\}\sum_{l=k}^{\infty}\frac{(\mu x_{i-1})^l}{l!}e^{-\mu x_{i-1}}\end{aligned} \tag{4}$$

Idle time is defined as the difference between the expected working time and the actual closing time of a session. Given that the former is a constant $(1\text{-}p)N/\mu$, the expected idle time is derived by subtracting it from the expected off duty time:

$$\begin{aligned}I &= \left(\sum_{i=1}^{N-1}x_i\right) + E\left[\textit{service time after } t_N\right] - \frac{(1-p)N}{\mu} \\ &= \left(\sum_{i=1}^{N-1}x_i\right) + w_N + \frac{1-p}{\mu} - \frac{(1-p)N}{\mu}\end{aligned} \tag{5}$$

Overtime is defined as a positive deviation between the working hours and the actual closing time. The expected overtime is categorized into two cases based on the appointment time of the last patient and is denoted by O_1 and O_2, respectively:

1. If all patients are scheduled before the end of session, i.e., $\sum_{i=1}^{N-1}x_i \leq T$, the expected overtime is given as:

$$O_1 = \sum_{k=0}^{N-1}\Pr\{N_N = k\}\left[(1-p)o_{1k} + po_{2k}\right] \tag{6}$$

In this case, the doctor has to remain available until t_N to see whether the last patient shows up for the appointment. Using the memoryless property of exponential distribution, the expected overtime when the last patient shows up, represented by o_{1k}, is the probability of l departures before the closing time multiplied by $(k-l+1)$s

expectations of service time. The expected overtime when the last patient does not show up, o_{2k}, is computed in a similar way, except that there are only $(k-l)$s expectations of service time. Thus, O_1 is derived by:

$$\begin{aligned} O_1 &= \sum_{k=0}^{N-1} \Pr\{N_N = k\} \left((1-p) \sum_{l=0}^{k} \frac{\mu^l (T-t_N)^l}{l!} e^{-\mu(T-t_N)} \right. \\ &\qquad \left. \frac{k-l+1}{\mu} + p \sum_{l=0}^{k-1} \frac{\mu^l (T-t_N)^l}{l!} e^{-\mu(T-t_N)} \frac{k-l}{\mu} \right) \\ &= \sum_{k=0}^{N-1} \Pr\{N_N = k\} \left(\sum_{l=0}^{k-1} \frac{\mu^l (T-t_N)^l}{l!} e^{-\mu(T-t_N)} \frac{k-l}{\mu} \right. \\ &\qquad \left. + (1-p) \sum_{l=0}^{k} \frac{\mu^l (T-t_N)^l}{l!} e^{-\mu(T-t_N)} \frac{1}{\mu} \right) \end{aligned} \tag{7}$$

2. If not all patients are scheduled before the end of session, i.e., $\sum_{i=1}^{N-1} x_i > T$ the expected overtime is given as:

$$O_2 = \left(\sum_{i=1}^{n-1} x_i \right) + E\left[service\ time\ after\ t_n \right] - T = \left(\sum_{i=1}^{N-1} x_i \right) + w_N + \frac{1-p}{\mu} - T \tag{8}$$

Thus, the traditional AS model for an arbitrary schedule is formulated as follows:

$$\begin{aligned} \Phi_1(x) &= c_i \left(\alpha(1-p)W + I + \beta O_1 \right) \\ &= c_i \left(\alpha(1-p)W + \left(\sum_{i=1}^{N-1} x_i \right) + w_N + \frac{1-p}{\mu} - \frac{(1-p)N}{\mu} + \beta O_1 \right) \\ &\quad s.t. \sum_{i=1}^{N-1} x_i \le T \end{aligned} \tag{9}$$

$$\begin{aligned} \Phi_2(x) &= c_i \left(\alpha(1-p)W + I + \beta O_2 \right) = c_i \left(\alpha(1-p)W + \left(\sum_{i=1}^{N-1} x_i \right) + w_N \right. \\ &\quad \left. + \frac{1-p}{\mu} - \frac{(1-p)N}{\mu} + \beta \left(\left(\sum_{i=1}^{N-1} x_i \right) + w_N + \frac{1-p}{\mu} - T \right) \right) \\ &\quad s.t. \sum_{i=1}^{N-1} x_i > T \end{aligned} \tag{10}$$

The objective functions are simplified by neglecting the constants $(1-p)(1-N)/\mu$, and c_i:

$$\begin{cases} \Phi_1(x) = \alpha(1-p)W + \left(\sum_{i=1}^{N-1} x_i\right) + w_N + \beta O_1 \\ s.t. \sum_{i=1}^{N-1} x_i \le T \end{cases} \tag{11}$$

$$\begin{cases} \Phi_2(x) = \alpha(1-p)W + \left(\sum_{i=1}^{N-1} x_i\right) + w_N + \beta\left(\left(\sum_{i=1}^{N-1} x_i\right) + w_N + \frac{1-p}{\mu} - T\right) \\ s.t. \sum_{i=1}^{N-1} x_i > T \end{cases} \tag{12}$$

2.3. Traditional AS Model under the No-Golf Policy

This section considers a more realistic scenario, the so-called no-golf policy, in which the doctor is not free to leave before the closing time of a session even if he finishes seeing all patients appointed. This assumption is more reasonable when auxiliary jobs are taken into consideration, such as going through records of the patients of the next session.

Neither the expected waiting time of each show-up patient nor the doctor's overtime is influenced by this assumption. However, the expected idle time is determined by the closing time of the doctor, which is the larger of the actual off duty time and the predetermined end of the session.

If not all patients are scheduled before the end of session, the doctor certainly has to work overtime to see all the patients. In this case, the expected idle time is the same as eqns. 5.

If all patients are served before the end of session, the idle time is

$$I = \sum_{k=0}^{N-1} \Pr\{N_N = k\}\left[(1-p)I_{1k} + pI_{2k}\right] \tag{13}$$

where I_{1k}, and I_{2k} represent the cases when the last patient shows up and doesn't show up, respectively:

$$\begin{aligned} I_{1k} = \Bigg(& \sum_{l=0}^{k} \frac{\mu^l (T-t_n)^l}{l!} e^{-\mu(T-t_n)} \left(T + \frac{k-l+1}{\mu}\right) \\ & + \sum_{l=k+1}^{\infty} \frac{\mu^l (T-t_n)^l}{l!} e^{-\mu(T-t_n)} T - \frac{N(1-p)}{\mu} \Bigg) \end{aligned} \tag{14}$$

$$I_{2k} = \left(\sum_{l=0}^{k-1} \frac{\mu^l (T-t_n)^l}{l!} e^{-\mu(T-t_n)} \left(T + \frac{k-l}{\mu} \right) + \sum_{l=k}^{\infty} \frac{\mu^l (T-t_n)^l}{l!} e^{-\mu(T-t_n)} T - \frac{N(1-p)}{\mu} \right) \tag{15}$$

Substituting eqns. 14 and 15 into eqns. 3 leads to the expected idle time:

$$I = T - \frac{N(1-p)}{\mu} + O_1 \tag{16}$$

The expected idle time is equal to subtracting the expected working hours from the actual time off duty, which is consistent with the intuition. The objective function under no-golf policy is rewritten as follows after omitting the negative constant $(1-p)N/\mu$:

$$\begin{cases} \Phi_1(x) = \alpha(1-p)W + T + O_1(1+\beta) \\ s.t. \sum_{i=1}^{N-1} x_i \le T \end{cases} \tag{17}$$

$$\begin{cases} \Phi_2(x) = \alpha(1-p)W + \left(\sum_{i=1}^{N-1} x_i \right) + w_N + \frac{1-p}{\mu} \\ \quad + \beta \left(\left(\sum_{i=1}^{N-1} x_i \right) + w_N + \frac{1-p}{\mu} - T \right) \quad s.t. \sum_{i=1}^{N-1} x_i > T \end{cases} \tag{18}$$

2.4. Notations and Assumptions for OA-AS System

To facilitate the comparison, it is assumed that all patients who call before the beginning of a session will show up for the appointment on time under open access policy, and the expected workload is the same as that of the traditional AS system. As Robison and Chen [10] noted, it is theoretically sensible to use a Poisson distribution to model the number of booking requests in the session. From a realistic point of view, an upper bound is set for the maximum number of patients scheduled in a session. As for the same-or–next-session policy, the number of patients who can be delayed to the next session is no more than the expected workload because deferring too many patients provides no benefit in smoothing the demand fluctuation between different sessions. The formulation is given only for the case in which the doctor can only leave the office after seeing all of the patients on his/her schedule. The notations for the OA-AS system are the following:

d: The maximum number of patients that can be postponed to the next session.
$I(m)$: The idle time (min) of the doctor if m patients are scheduled in a session.
M: The maximum number of patients to be scheduled in a session.

$\bar{n}$: The expected workload in a session, $\bar{n} = N(1-p)$.
$O(m)$: Expected overtime (min) of the doctor if m patients are scheduled in a session.
q_i: The probability of starting a session with i patients delayed from the previous session.
$w_i(m)$: Expected waiting time (min) of the ith patient if he/she shows up on the condition that m patients are seen in a session.
$w(m)$: Expected waiting time (min) if m patients are scheduled in a session.
$\varphi(m)$: The probability mass function that m patients call for appointments before the session begins.
$\Phi(m)$: The cumulative distribution function of $f(m)$.
$\phi(m)$: The probability distribution that m patients are seen in a session.

Other notations and assumptions are the same as the traditional model.

2.5. Model for Same-Session Policy

Under the same-session policy, the number of patients scheduled in a session follows a Poisson distribution, and any requests exceeding the upper bound of the capacity will be ignored; all appointments are scheduled to the session the patients request. When M is large enough, the abandonment has little impact on the expectation of the distribution.

$$\varphi(m) = \frac{(\bar{n})^m}{m!} e^{-\bar{n}}, \forall m, \quad \phi(m) = \begin{cases} \dfrac{(\bar{n})^m}{m!} e^{-\bar{n}}, & m = 0, \ldots, M-1 \\ 1 - \displaystyle\sum_{k=0}^{M-1} \dfrac{(\bar{n})^k}{k!} e^{-\bar{n}}, & m = M \end{cases} \tag{19}$$

The performance when m patients call before the beginning of the session is evaluated in the same way as the traditional AS model, except $N = m$, $p = 0$.

The probability distribution of system state at t_i can be derived from eqns. 3 and 4:

For $$1 \le j < i, \quad \Pr\{N_i = j\} = \sum_{k=0}^{i-j-1} \Pr\{N_{i-1} = j+k-1\} \frac{(\mu x_{i-1})^k}{k!} e^{-\mu x_{i-1}} \tag{20}$$

$$\Pr\{N_i = 0\} = \sum_{k=1}^{i-1} \Pr\{N_{i-1} = k-1\} \sum_{l=k}^{\infty} \frac{(\mu x_{i-1})^l}{l!} e^{-\mu x_{i-1}} \tag{21}$$

The overall performance measures W, I, O are the weighted sums of $w(m)$, $I(m)$, $O(m)$, where

$$\begin{gathered} w_i(m) = \sum_{j=1}^{i-1} \left(\frac{j}{\mu} \cdot \Pr\{N_i = j\} \right), W(m) = \sum_{i=1}^{m} w_i(m), \\ W = \sum_{m=1}^{M} W(m)\phi(m) \end{gathered} \tag{22}$$

$$I(m)=\sum_{i=1}^{m-1}x_i+w_m+\frac{1-m}{\mu},m\geq 1,\ I=\sum_{m=1}^{M}I(m)\phi(m) \tag{23}$$

$$O(m)=\begin{cases}\sum_{i=1}^{m-1}x_i+w_m+\frac{1}{\mu}-T, & if\ \sum_{i=1}^{m-1}x_i>T\\ \sum_{k=0}^{m-1}\Pr\{N_m=k\}\left(\sum_{l=0}^{k}\frac{\mu^l(T-t_m)^l}{l!}e^{-\mu(T-t_m)}\frac{k-l+1}{\mu}\right), & if\ \sum_{i=1}^{m-1}x_i\leq T\end{cases}$$
$$O=\sum_{m=1}^{M}O(m)\phi(m) \tag{24}$$

The objective of the same-session policy is to find an optimal schedule to minimize the total costs:

$$C(x)=c_wW+c_iI+c_oO=c_i(\alpha W+I+\beta O) \tag{25}$$

2.6. Model for Same-Or-Next-Session Policy

As for the same-or-next-session policy, the performance measures are calculated by the same approach described in eqns. 19-24, but the number of patients seen in a session need to be re-evaluated. As demonstrated by Robinson and Chen [10], the probability mass function of the number of patients delayed to the next session can be obtained by solving the transition equation of the Markov chain.

No deferred patients means that the workload of the session before is no more than $\lfloor T\times\mu\rfloor$:

$$q_0=\sum_{j=0}^{d}q_j\cdot\Phi(\lfloor T\times\mu\rfloor-j)$$

i deferred patients means $\lfloor T\times\mu\rfloor+i$ patients arrive in the last session:

$$q_i=\sum_{j=0}^{d}q_j\cdot\varphi(\lfloor T\times\mu\rfloor+i-j),i\leq d-1$$

The sum of all probability is equal to 1: $\sum_{j=0}^{d}q_j=1$

With the solution to the above-mentioned equations, the number of patients seen in a session is obtained as follows:

$$\phi(m)=\begin{cases}\sum_{j=0}^{\min\{d,m\}} q_j\cdot\varphi(m-j), & m\le\lfloor T\times\mu\rfloor-1\\ \sum_{j=0}^{d} q_j\cdot\sum_{k=0}^{d}\varphi(\lfloor T\times\mu\rfloor-j+k), & m=\lfloor T\times\mu\rfloor\\ \sum_{j=0}^{d} q_j\cdot\varphi(d+m-j), & \lfloor T\times\mu\rfloor+1\le m<M\\ \sum_{j=0}^{d} q_j\cdot\sum_{M}^{\infty}\varphi(d+m-j), & m=M\end{cases} \tag{26}$$

3. RESULTS

3.1. Solution for the General Traditional AS Model

The traditional appointment scheduling problem is a convex minimization problem [44] of which the optimal numerical solution can be found by the '*fmincon*' function in the commercial software Matlab implementing sequential quadratic programming algorithm [11, 12].

The effects of the parameters are analyzed based on a medium-sized problem. In particular, the expected service time is normalized at 1, and the basic problem assumes $N = 16$, $T = 12$. That is to say, the day length is 12 times as long as the expected service time. In practice, this case corresponds to the situation where 16 patients are scheduled to a three-hour working session with an average service time of fifteen minutes. All performance measures in the following experiment are presented by means of this normalized dimensionless time for simplicity. Other input parameters are given as follows: the no-show probability p is 0.25, relative cost of overtime β is 2, and the relative cost of waiting α ranges from 0 to 1. Because of the scarcity of the healthcare resources, the situation that waiting cost is more important is not considered.

Figure 1 shows the commonly dome-pattern interval lengths under different values of α, with the interval numbers labeled on the horizontal axis [36]. It is found that the interval length increases in the earlier part of session, and then remains relatively steady afterwards. Finally, it decreases in the latter part of session under each α considered except 0 because the doctor compresses the first intervals in order to reduce the possibility of idle time at the beginning of the session. The last few patients (e.g., 3) are scheduled closer in order to mitigate idle time and overtime incurred by no-show patients at the end of session. Appointment intervals are kept almost the same in most of the intervals in order to maintain steady workload. It is also noted that the more patients' waiting are weighted, the larger the interval lengths are allocated. Intuitively speaking, this phenomenon occurs in that the objective is to strike a balance between the interests of patients and the doctor.

Figure 2 presents the expected waiting time of each patient showing up, the expected idle time and overtime of the doctor. Figure 2 displays that the impacts of no-shows

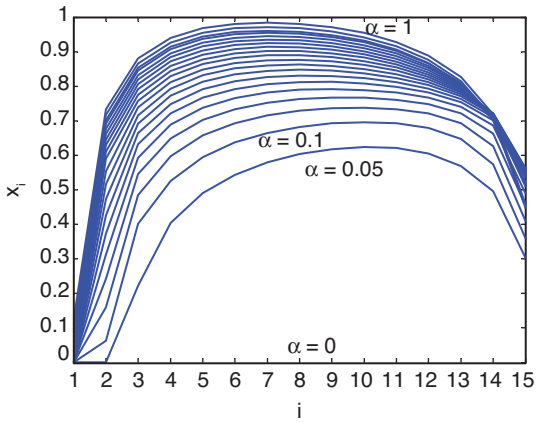

Figure 1. Scheduled intervals of traditional policy under different α values.

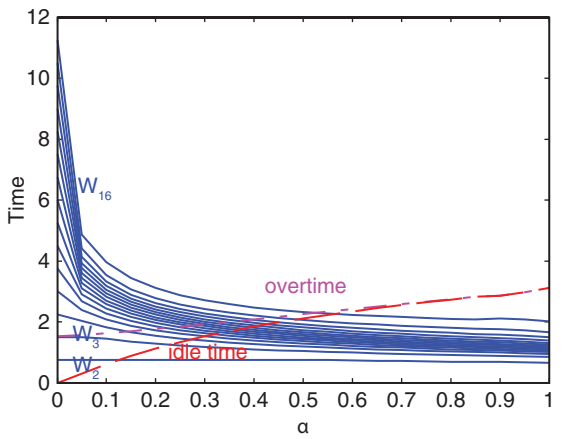

Figure 2. System performance of traditional policy under different α values.

accumulate in the queue, and the delay of each show-up patient increases with the position in the waiting list. Another intuitive result is the monotonically non-increasing curve of waiting for each patient in line, reflecting that the more weight the clinic gives to waiting time, the less time the patients have to spend. On the contrary, the idle time and overtime curves are monotonically increasing with the weight of waiting, reflecting that the objective function is aimed at balancing the efficiency of doctors and the satisfaction of patients. More attention paid to the waiting time of the patients means that the relative weight of idle time and overtime in the objective function will decline. From the standpoint of the appointment system, the solid line in Figure 3 depicts the trend of the objective function under various values of α in traditional AS system. The fact that total costs monotonically increases with the weight given to waiting time is not

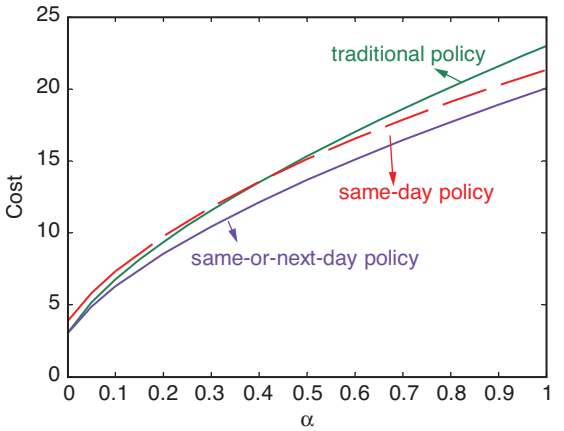

Figure 3. Total operational costs of AS systems under different α values.

only because of the higher coefficient of the patients, but also due to the decrease of doctor's efficiency caused by the attempts to reduce patients' waiting time. The superiority of open-access policies is not as explicit as the results under the deterministic service time. The same-session system has a threshold of α in terms of total operational costs and the same-or-next-session system is always preferable.

3.2. Numerical Solution for the Open-Access AS Model

The parameters under open-access policies are consistent with those under the traditional policy for a reasonable comparison. As for the same-or-next-session policy, the value of d is set equal to T in all of the subsequent experiments. The key point for finding the optimal interval length is to determine how many appointments can be scheduled within the working hours. Therefore, the original problem is decomposed into $M+1$ sub-problems as follows:

1. all of the patients can be scheduled before the end of session, i.e., $\sum_{i=1}^{M-1} x_i \le T$;
2. only M-1 patients can be scheduled before the end of session, i.e., $\sum_{i=1}^{M-2} x_i \le T$, $\sum_{i=1}^{M-1} x_i > T$;

…

$M+1$) all of the patients are scheduled in the overtime, i.e., $x_i > T$.

In each sub-problem, only one overtime expression of eqn. 22 holds for any given m. The minimum of all local optimal solutions derived by the same approach described in section 3.1 is the global numerical optimal solution for an open-access policy.

Figures 4 and 5 display the OA-AS system performance measures under the same-session and same-or-next-session policies, respectively. The curves in red represent the

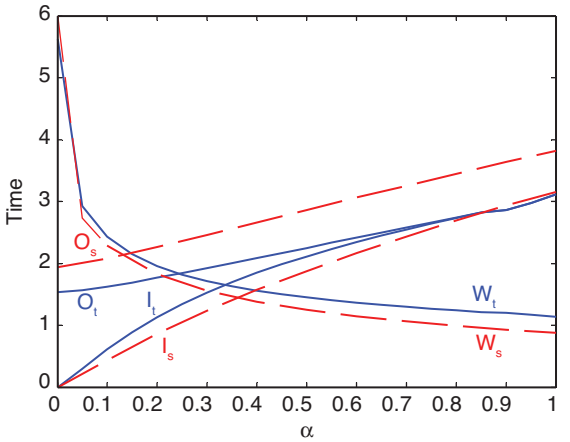

Figure 4. System performance of same-session policy under different α values.

performance of the same-session and same-or-next-session systems with the same expected workload. The phenomenon that the average expected waiting time of all patients is monotonically non-increasing coincides with the traditional results. Higher waiting parameters decrease the delays in the queue and increase both the idle time and overtime because the objective is to balance the waiting time of patients and the working time of doctors.

Another observation is that the OA-AS policy is superior to the traditional policy in the basic problem in terms of average waiting time when α is larger than 0.05. The same-session policy gives rise to higher overtime for all values of α and higher idle time when $\alpha > 0.85$, whereas the same-or-next-session policy brings about less idle time all the time. When $\alpha < 0.2$, the overtime is almost the same for the both traditional and same-or-next-session policies; otherwise, the same-or-next-session policy requires slightly less overtime when α grows larger. From the standpoint of a clinic, the same-session policy lowers the total operational cost when $\alpha \geq 0.4$, while the same-or-next-session policy maintains lower cost for all values of α considered as shown in Figure 3. The source of uncertainty for OA-AS policies is the random number of patients. The above observations imply that the negative effect of random patients is less than that of no-show patients for the open-access policy when greater importance is attached to the waiting time of patients.

Contrasting Figure 4 and Figure 5, it is noted that the same-or-next-session policy obviously outperforms the same-session policy in terms of overtime and idle time. As noted by Robinson and Chen [9], the fluctuation of patients seen in each session under the same-or-next-session policy is much smaller than that of same-session policy.

3.3. Numerical Solution under No-Golf Policy

This section examines the situation when the doctor cannot leave the office before the predetermined end of the session. Compared with Figure 1, Figure 6 shows that the no-golf policy only affects the last few intervals in the traditional AS system. It is observed

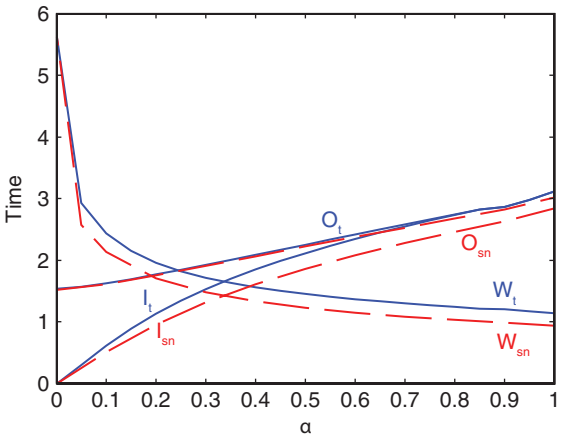

Figure 5. System performance of same-or-next-session policy under different α values.

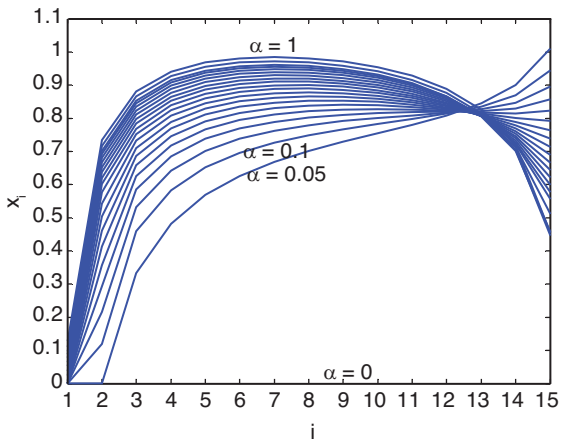

Figure 6. Scheduled intervals of traditional no-golf policy under different α values.

that the last few patients are scheduled scarcely because there is no benefit to finish consultation ahead of time. However, increasing the interval length cuts down the patients' waiting time significantly without influencing the idle time of doctors. When the waiting time is given more weight, the scheduled intervals are in the same pattern as Figure 1 in order to mitigate the patients' waiting by longer idle time and overtime.

Figure 7 shows that the waiting time of each patient increases with his/her position in the queue, and the expected idle time is equal to the expected overtime for each value of α because of the parameters satisfying the equation $T = N(1-p)$. Otherwise, the curve of overtime and idle time will be parallel. Since the session length is just enough for the

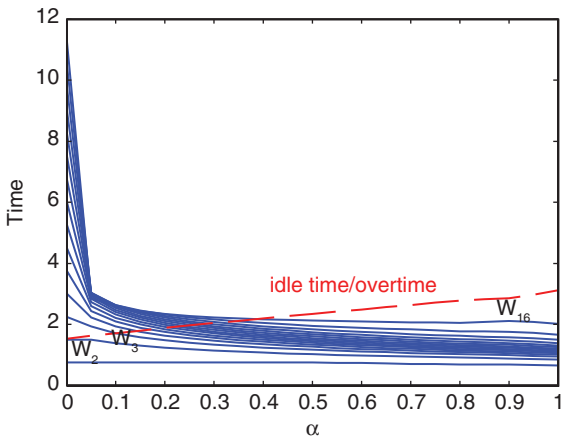

Figure 7. System performance of traditional no-golf policy under different α values.

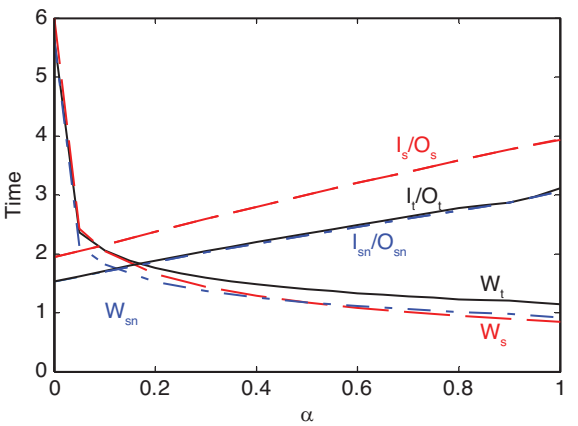

Figure 8. System performance of open-access no-golf policy under different α values.

expected workload, the above equality reveals that idle time often causes overtime. This phenomenon is also viewed as an expense of the clinic to trade off the uncertainty of service time. Compared with Figure 2, it is found that longer working hours reduce delays in the waiting list. In addition, considering the possibility that the doctor may finish the examination of the last patient in advance, idle time is inevitable even if patients' waiting is negligible.

Figures 8 and 9 display the results of the same-session no-golf policy and the same-or-next-session no-golf policy, respectively. The trends of the curves are the same as

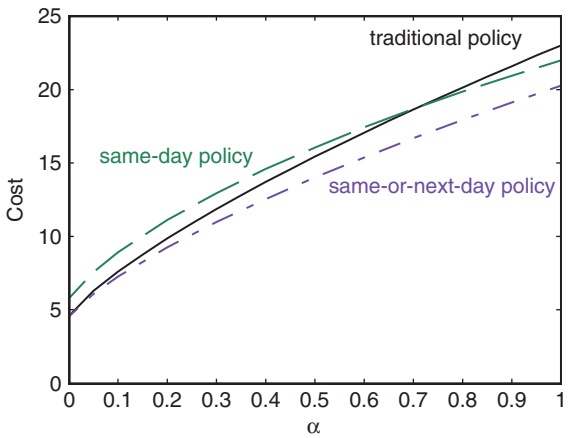

Figure 9. Total operational costs of no-golf AS systems under different α values.

those in Figures 6 and 7 except when the idle time is longer than usual. Idle time is nearly the same as overtime for any given α because the physician is not allowed to go off duty in advance, and this will cause irreducible idle time in a particular session. However, the reason for the equality is quite different from that depicted in Figure 7. In fact, the overtime and idle time is not exactly the same if the precision of the numerical computing is improved. For a given number of arrivals, they are not necessarily the same; therefore, the analogy of the ultimate results is viewed as a coincidence. Figure 9 shows a similar conclusion as Figure 4; i.e. the open-access policy lowers the total operational costs when α is larger than a certain threshold, e.g., 0.7, for the same-session policy; the same-or-next-session policy is always preferable.

3.4. Effects of the Session Length

The session length is varied from 12 to 20 in order to analyze its effects on performance measures, while the expected workload, no-show probability and relative overtime cost are kept the same as those in the basic problem. Figures 10 to 12 display the average expected waiting time with a black solid line, expected overtime with a blue dotted line and expected idle time with a red dashed line as functions of T for the traditional policy, same-session policy and same-or-next-session policy, respectively. To clarify the variation trends of the performance, the curves are plotted only with α equal to 0, 0.2, 0.4, 0.6, 0.8, 1. It is noted that the average waiting time and expected overtime decrease, the expected idle time increases with the session length for any specific α under all policies. Overtime cannot be eliminated thoroughly due to the stochastic characteristics of service time, even if the session length is adequate for all patients to show up. The average waiting time is almost the same for different T values when α is zero because increasing the session length reduces the overtime instead of delays in the queue when the waiting cost of patients is negligible. Similar to the basic case, it is observed that the average waiting time and idle time are reduced by OA-AS policies when α is large, and

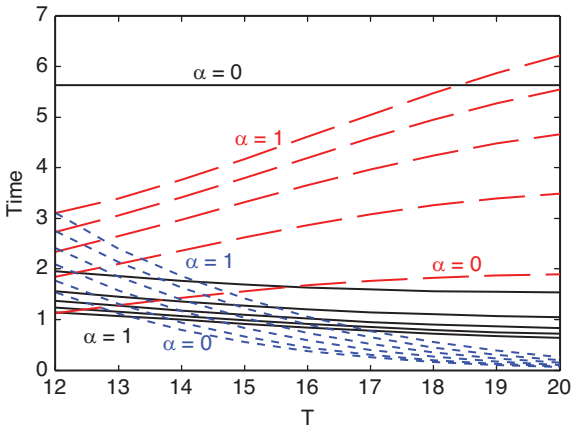

Figure 10. System performance of traditional policy under different T values.

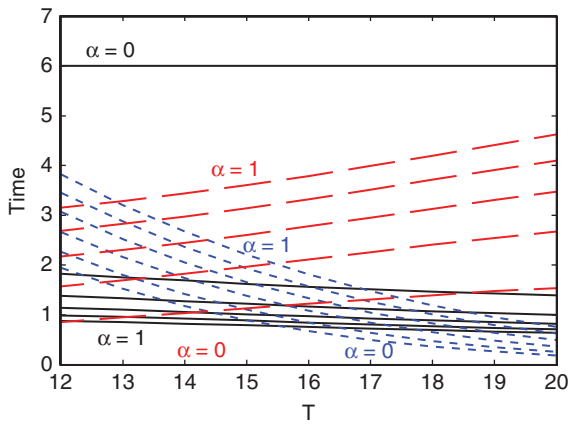

Figure 11. System performance of same-session policy under different T values.

more overtime are incurred by the same-session policy for any given T. Regarding the total operational costs shown in Figure 13, the same-or-next-session policy (denoted by the blue dashed line) has lower costs than those of the traditional policy (denoted by the black solid line) except when α is small. For all of the T values considered, the same-session policy (denoted by the red dotted line) reduces the total costs when α is larger than a certain threshold.

3.5. Effects of the No-Show Probability

To analyze the impacts of no-show probability, the expected workload $\overline{n}$ is kept constant at 12, while the number of patients scheduled per session varies from 12 to 24. The no-show probability can be calculated as $p = 1 - \overline{n} / N$ between 0% and 50%; the session length and relative cost of overtime are the same as those in the basic problem.

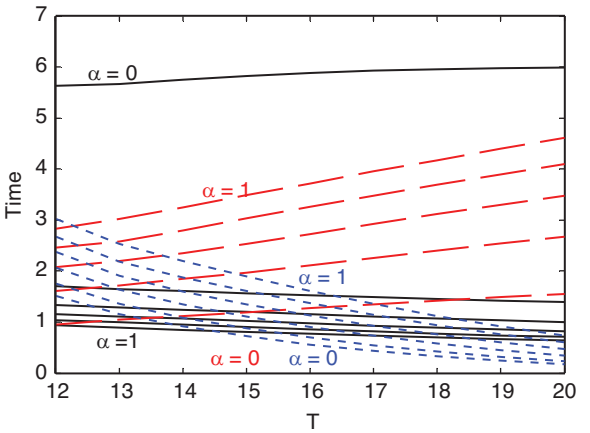

Figure 12. System performance of same-or-next-session policy under different *T* values.

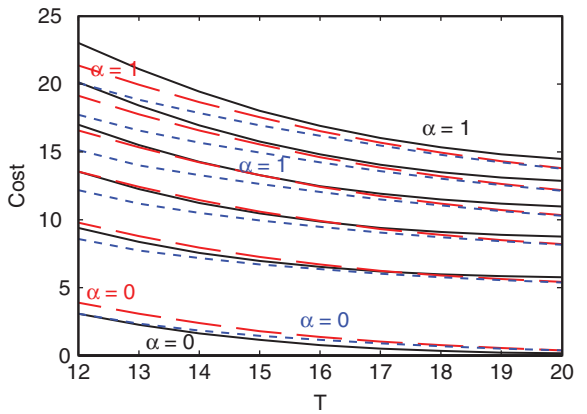

Figure 13. Total operational costs of AS systems under different *T* values.

Figures 14 to 17 show the average expected waiting time, expected idle time, expected overtime and total operational cost as a function of α under different no-show probabilities, respectively. The performance measures under the traditional policy are plotted with black solid lines, while those under the same-session and same-or-next-session policies are represented by the red dashed lines and purple dotted lines, respectively. If all no-show patients notify the clinic of their cancellation of appointment in advance, the AS system with the same expected workload (e.g., $p = 0$) can reduce the patients' waiting time as well as the doctor's idle time. It is also noted that all three operational costs increase monotonously with no-show probability. The larger the no-show probability, the worse the performance measure. As intended by the administrator of the clinics, the OA-AS policy can be viewed as an alternative way to

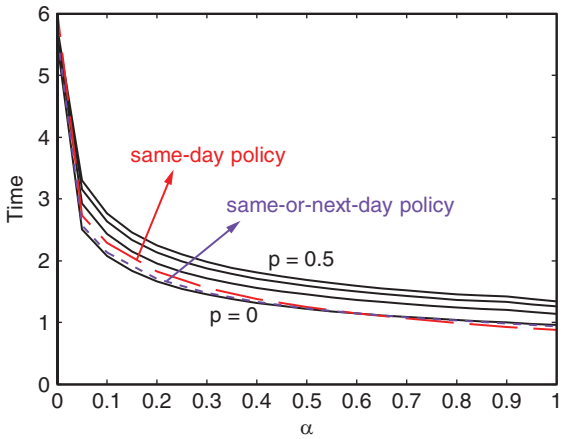

Figure 14. Average waiting time of AS systems under different p values.

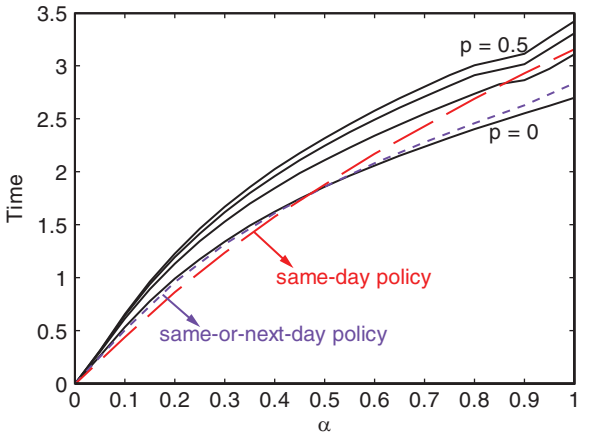

Figure 15. Idle time of AS systems under different p values.

lower operational costs when the no-show probability is high. Similar to the basic cases, the same-or-next-session policy is preferable to the same-session policy in terms of total operational costs. Once again, Figure 17 indicates that the same-session system outperforms the traditional system beyond a certain threshold of α, and the same-or-next-session system is always preferable except when the no-show probability or the weight of the patients is low.

3.6. Effects of the Expected Workload

To investigate the impacts of the expected workload, the number of patients scheduled in a session is varied from 4 to 28, while the expected workload is still calculated by

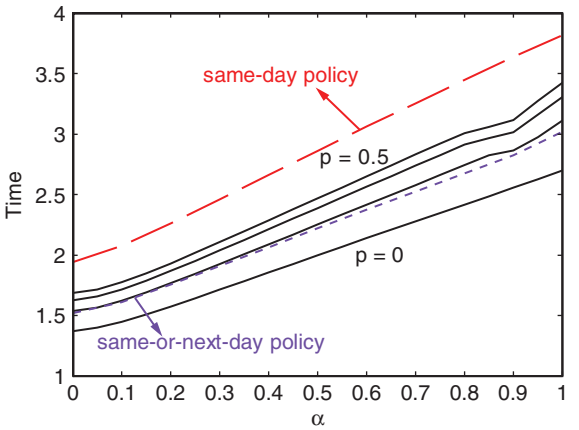

Figure 16. Overtime of AS systems under different p values.

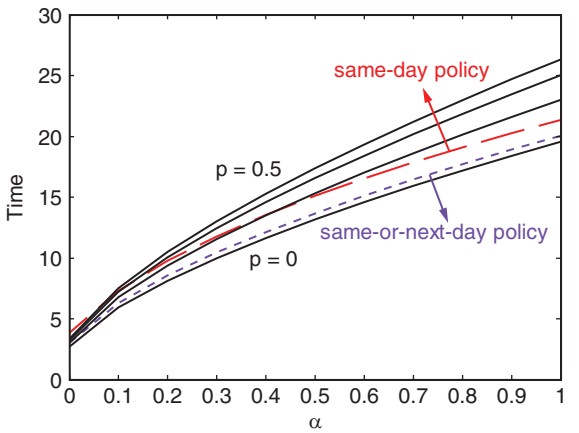

Figure 17. Total operational costs of AS systems under different p values.

$\overline{n} = N(1-p)$. To facilitate comparison, the session length remains the same as the expected workload to eliminate the influences of other factors, e.g., $T = \overline{n}$; other parameters are kept the same as those in the basic problem.

Figures 18 to 20 present the system performance measures as a function of N under different values of α for traditional, same-session and same-or-next-session policies, respectively. Although the value of α is tested from 0 to 1 with an increment of 0.05, the curves are depicted only with α values of 0, 0.2, 0.4, 0.6, 0.8, 1 for illustration. As shown in the figures, the trends of average waiting time denoted by the black solid lines, idle time by the red dashed lines and overtime by the blue dotted lines are the same as those in the basic problem for any given N. For any given α, the idle time and overtime increase with the number of patients scheduled per session because the negative

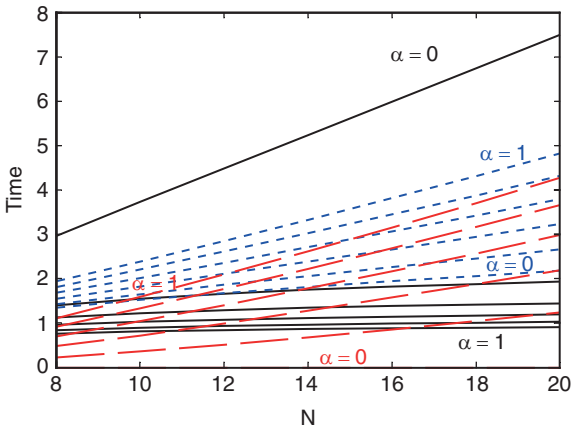

Figure 18. System performance of traditional policy under different N values.

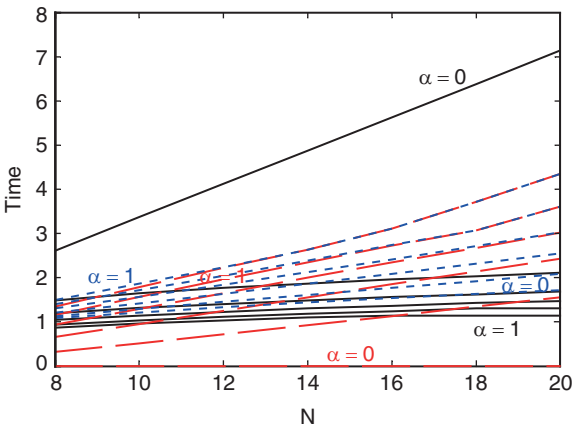

Figure 19. System performance of same-session policy under different N values.

influence of no-show is accumulated through the waiting list, just like the bullwhip effect in supply chain. However, when $\alpha \geq 0.2$, the average waiting time does not change significantly when N is large enough; this can be viewed as the preliminary period before the steady state. This phenomenon also reveals that the clinic has to make additional efforts in maintaining the same service level as the expected workload increases. Figure 21 presents the average operational cost per patient for traditional, same-session and same-or-next-session policies with the black solid, red dashed and blue dotted line, respectively. It is noted that the average operational costs increase with the expected workload except when $\alpha = 0$, given that only overtime is incurred in this situation with lower increment than that of the expected workload.

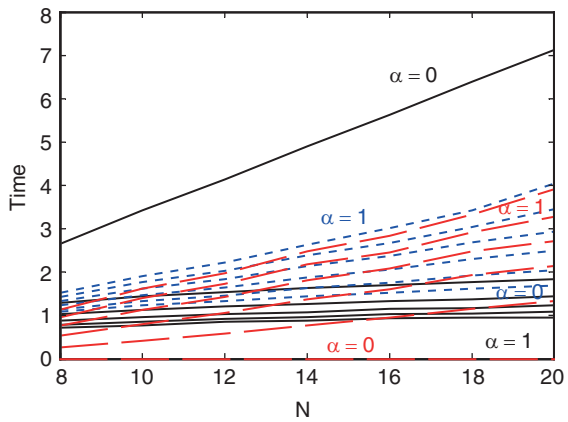

Figure 20. System performance of same-or-next-session policy under different N values.

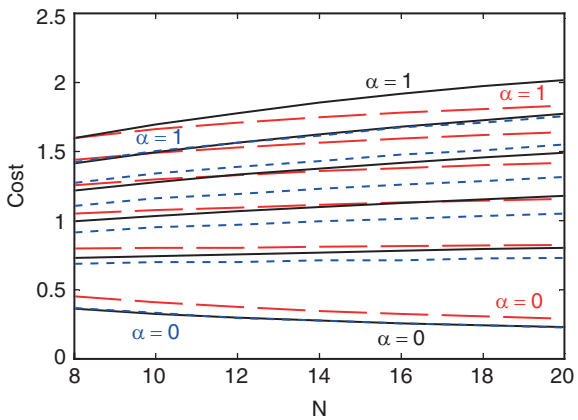

Figure 21. Average operational costs of AS systems under different N values.

3.7. Effects of the Relative Overtime Cost

This subsection illustrates the effects of the overtime ratio on the system performances. β is varied from 1 to 4 with an increment of 0.5, while other parameters remain the same as those in the basic problem. Figures 22 to 24 present the system performance under different values of β by a black solid, red dashed and blue dotted line representing the average waiting time, idle time and overtime, respectively. As with the previous session, the curves are plotted only with α values of 0, 0.2, 0.4, 0.6, 0.8, 1. As shown in the figures, higher overtime ratios require more efficient use of the working hours in order to reduce the overtime as much as possible, leading to shorter idle time and longer waiting time. In other words, higher overtime ratios decrease the relative weights given to the patients in

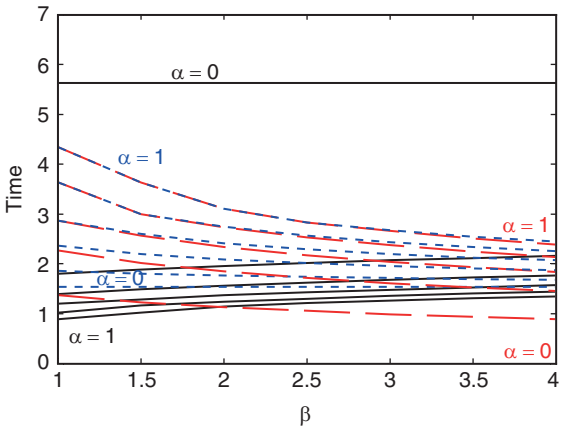

Figure 22. System performance of traditional policy under different β values.

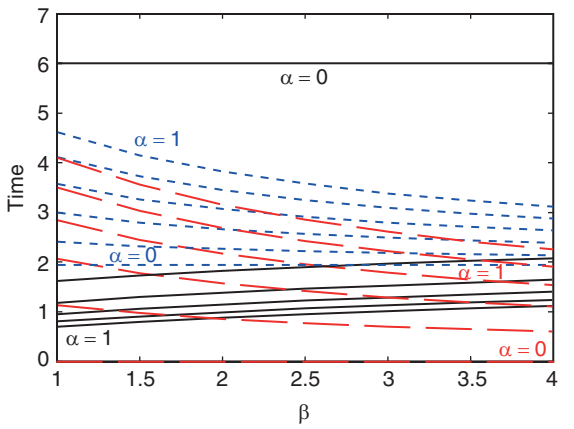

Figure 23. System performance of same-session policy under different β values.

the total costs. Consequently, the optimal schedule alleviates the relatively high overtime cost by increasing patients waiting time. Figure 25 presents the total operational costs for all policies under different values of β. The same-or-next-session policy (represented by the blue dotted line) reduces the total cost except when α is small, whereas the same-session policy does not outperform the traditional policy beyond certain thresholds of β for all cases. The same-session policy has a significant increase in total costs when β is large owing to the higher overtime than that of other policies.

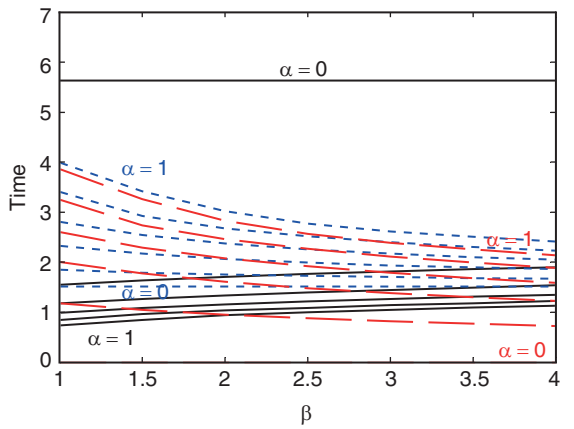

Figure 24. System performance of same-or-next-session policy under different β values.

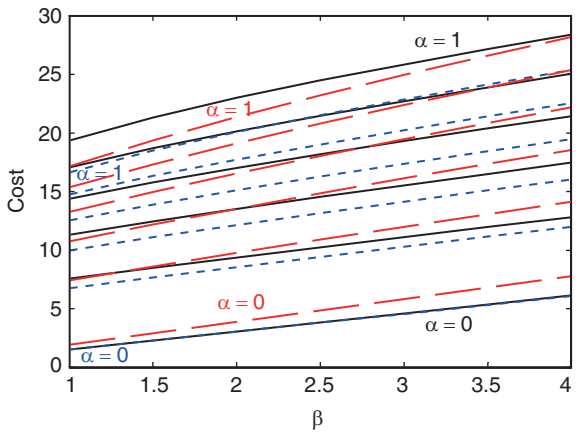

Figure 25. Total operational costs of AS systems under different β values.

4. DISCUSSION

Sequential quadratic programming methods are employed to search for the numerical optimal solutions for different AS systems. Session length, no-show probabilities, expected workload and relative overtime cost are examined separately to investigate their influences on the comparison. It is noted that the same-session AS system has a threshold of relative waiting cost, beyond which it outperforms the traditional system; and the same-or-next-session system is always preferable except when the no-show

probability or the weight of the patients is low. According to the numerical results, the no-golf policy has significant impacts on the length of the last intervals, and the expected waiting time is reduced especially when relative waiting cost is small. This indicates that when patients' satisfaction is not valued highly, enlarging the last intervals will improve the efficiency of the clinic, and thus reduce the total operational costs. Increasing the session length certainly results in less average waiting time, less overtime and more idle time. Increasing the penalty coefficient for overtime forces the AS system to offset the overtime by more patients waiting. Consistent with previous studies, no-show is identified as a key factor influencing the system performance. Although traditional overbooking policies of compressing the interval length balances the negative effects of no-show to some extent, the open-access policy is a more effective alternative when the no-show probability is high.

Although the current comparisons are focused on the impacts of exponentially distributed service time on the differences of two widely used appointment scheduling policies, some assumptions may restrict the generality of the results. First, the service time may follow general distribution, such as lognormal or Gamma distribution [28]. It remains unknown to what extent this will change the results of the comparisons. Second, the real appointment processes are more complicated than the scenarios studied here. Patients usually have preferences of the appointment time based on their convenience, and undesirable time intervals obviously increase the possibilities of no-shows. Taking patients' choice into consideration is definitely an effective way to reduce no-shows [45]. Third, late cancellations are not entirely the same as no-shows. If the confirmation phone calls or emails reveal potential no-shows, the time intervals concerned can be released to other urgent patients who hope to see the doctor in the calling session [8]. Fourth, all patients are assumed to be homogenous in this research. Cayirli et al. [26] evaluated the effects of patients' classification on the efficiency of the AS system, and their simulation results indicate that appropriate sequencing and interval adjustment according to patient types significantly reduce patients' waiting time, physician's idle time and overtime. Further optimization methods need to be developed to accommodate the heterogeneous characteristics of service time and walk-in patients [46]. Finally, physicians can cooperate and share medical appointments in case one physician is overloaded in a particular session while other physicians have free intervals [47]. Furthermore, when patients arrive at the clinic, they usually go through registration, examination by the doctor, X-ray, laboratory and checkout [2]. These imply that multiple-server and multiple-stage models are needed to study the transition probabilities between different stages.

5. CONCLUSIONS

In this paper, two types of appointment scheduling policies are compared under exponentially distributed service times. The sequential quadratic programming method is employed to search for a solution to minimize the weighted sum of the expected waiting time, idle time and overtime under the traditional policy. An open-access policy is proposed as an alternative approach to mitigate the negative effects of no-shows. Numerical experiments show that an open-access policy can reduce the operational cost

of a clinic when the no-show probability is high, and that the same-or-next-session policy outperforms the same-session policy due to less fluctuation. However, there are no apparent small thresholds of relative waiting cost beyond which open-access policy is superior to the traditional policy. The results indicate that the same-session policy performs better when the relative waiting cost is higher than a certain threshold in most cases, depending on the concrete setting of the parameters. The same-or-next-session policy is preferable except when the relative waiting cost is low. Future studies can consider modeling other realistic situations, such as patients' preferences, changing the number of doctors, classification of patients and multiple-stage queueing networks.

ACKNOWLEDGMENTS

This work is financially supported by the National Natural Science Foundation of China (71021061, 70625001) and a grant from the HKSAR RGC (Project No. T32-102/14-N). We would like to thank the peer reviewers for their insightful and constructive suggestions.

CONFLICT OF INTEREST

The authors indicate no potential conflicts of interest.

NOMENCLATURE

c_i: Doctor's idle cost per unit time (dollar/min).
c_o: Doctor's overtime cost per unit time (dollar/min).
c_w: Patients' waiting cost per unit time (dollar/min).
d: The maximum number of patients that can be postponed to the next session.
I: Expected idle time (min) that a doctor wastes during a standard working session.
$I(m)$: The idle time (min) of the doctor if m patients are scheduled in a session.
M: The maximum number of patients to be scheduled in a session.
$\overline{n}$: The expected workload in a session, $\overline{n} = N(1-p)$.
N: Number of patients to be scheduled in a standard working session.
N_i: Number of patients in the queue just before the arrival time of the ith scheduled patient.
O: Expected overtime (min) of the doctor in a standard working session.
$O(m)$: Expected overtime (min) of the doctor if m patients are scheduled in a session.
p: No-show probability of each patient.
q_i: The probability of starting a session with i patients delayed from the previous session.
T: The predetermined length of a standard working session.
t_i: Time (min) of the ith scheduled arrival, $t_i = \sum_{j=1}^{i-1} x_j$.
w_i: Expected waiting time of the ith patient if he/she shows up, $w_1 = 0$.
$w_i(m)$: Expected waiting time (min) of the ith patient if he/she shows up on the condition that m patients are seen in a session.

$w(m)$: Expected waiting time (min) if m patients are scheduled in a session.

W: Expected waiting time (min) of all patients scheduled in a session if a patient shows up, $W = \sum_{i=1}^{N} w_i$.

x: A vector of scheduled intervals, $x = (x_1, x_2, \ldots, x_{N-1})$.

x_i: Time interval (min) between the ith and $(i + 1)$th scheduled patient.

$\alpha\ c_w/c_i$: The relative cost of patients' waiting, as a fraction of the idle time cost per unit time(min).

$\beta\ c_o/c_i$: The relative cost of doctor's overtime, as a fraction of the idle time cost per unit time(min).

$1/\mu$: Mean of exponentially distributed service time.

$\varphi(m)$: The probability mass function that m patients call for appointments before the session begins.

$\Phi(m)$: The cumulative distribution function of $\varphi(m)$.

$\phi(m)$: The probability distribution that m patients are seen in a session.

REFERENCES

[1] Weerawat W, Pichitlamken J, Subsombat P. A generic discrete-event simulation model for outpatient clinics in a large public hospital. *Journal of Healthcare Engineering*, 2013, 4(2): 285–305.

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

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

[4] LaGanga L. Lean service operations: Reflections and new directions for capacity expansion in outpatient clinics. *Journal of Operations Management*, 2011, 29(5): 422–433.

[5] Rinder MM, Weckman G, Schwerha D, Snow A, Dreher PA, Park N, Paschold H, Young W. Healthcare scheduling by data mining: Literature review and future directions. *Journal of Healthcare Engineering*, 2012, 3(3): 477–502.

[6] Maglogiannis I. Towards the adoption of open source and open access electronic health record systems. *Journal of Healthcare Engineering*, 2012, 3(1): 141–161.

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

[8] Huang Y, Zuniga P. Effective cancellation policy to reduce the negative impact of patient no-show. *Journal of the Operational Research Society*, 2014, 65(5): 605–615.

[9] Murray M, Tantau C. Same-day appointments: Exploding the access paradigm. *Family Practice Management*, 2000, 7(8): 45–50.

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

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

[12] Tang JF, Yan Cj, Fung R. Optimal appointment scheduling with no-shows and exponential service time considering overtime work. *Journal of Management Analytics*, 2014, 1(2): 99–129.

[13] Creemers S, Lambrecht M. An advanced queueing model to analyze appointment-driven service systems. *Computers and Operations Research*, 2009, 36(10): 2773–2785.

[14] Creemers S, Lambrecht M. Queuing models for appointment-driven systems. *Annals of Operations Research*, 2010, 178(1): 155–172.

[15] Klassen KJ, Yoogalingam R. Appointment system design with interruptions and physician lateness. *International Journal of Operations and Production Management*, 2013, 33(4): 394–414.

[16] Kong Q, Lee C, Teo C, Zheng Z. Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operations Research*, 2013, 6(3): 711–726.

[17] De Vuyst S, Bruneel H, Fiems D. Computationally efficient evaluation of appointment schedules in health care. *European Journal of Operational Research*, 2014, 237(3): 1142–1154.

[18] Mak H, Rong Y, Zhang J. Sequencing appointments for service systems using inventory approximations. *Manufacturing and Service Operations Management*, 2014, 16(2): 251–262.

[19] Qu X, Shi J. Effects of two-level provider capacities on the performance of open access clinics. *Health care Management Science*, 2009, 12(1): 99–114.

[20] Wang WY, Gupta D. Adaptive appointment systems with patient preferences. *Manufacturing and Service Operations Management*, 2011, 13(3): 373–389.

[21] Huang Y, Zuniga P, Marcak J. A cost-effective urgent care policy to improve patient access in a dynamic scheduled clinic setting. *Journal of the Operational Research Society*, 2014, 65(5): 763–776.

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

[23] Mercer A. A queueing problem in which the arrival times of the customers are scheduled. *Journal of the Royal Statistical Society. Series B*, 1960, 22(1): 108–113.

[24] Kim S, Giachetti RE. A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. *IEEE Transactions on Systems, Man and Cybernetics Part A-systems and humans*, 2006, 36(6): 1211–1219.

[25] Green LV, Savin S. Reducing delays for medical appointments: A queuing approach. *Operations Research*, 2008, 56(6): 1526–1538.

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

[27] Erdogan SA, Denton B. Dynamic appointment scheduling of a stochastic server with uncertain demand. *Informs Journal on Computing*, 2013, 25(1): 116–132.

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

[29] Chakraborty S, Muthuraman K, Lawley M. Sequential clinical scheduling with patient no-show: The Impact of Pre-defined Slot Structures. *Socio-Economic Planning Sciences*, 2013, 47(3): 205–219.

[30] Berg B, Denton B, Erdogan SA, Rohleder T. Optimal booking and scheduling in outpatient procedure centers. *Computers and Operations Research*, 2014, 50(1): 24–37.

[31] Zeng B, Turkcan A, Lin J, Lawley M. Clinic scheduling models with overbooking for patients with heterogenous no-show probabilities, *Annals of Operations Research*, 2010, 178(1): 121–144.

[32] Begen MA, Queyranne M. Appointment scheduling with discrete random durations. *Mathematics of Operations Research*, 2011, 36(2): 240–257.

[33] Ho C, Lau H. Minimizing total cost in scheduling outpatient appointments. *Management Science*, 1992, 38(12): 1750–1764.

[34] Wang PP. Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Research Logistics*, 1993, 40(3): 345–360.

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

[36] Klassen KJ, Yoogalingam R. Improving performance in outpatient appointment services with a simulation optimization approach. *Production and Operations Management*, 2009, 18(4): 447–458.

[37] Cayirli T, Yang KK, Quek SA. A universal appointment rule in the presence of no-shows and walk-ins. *Production and Operation Management*, 2012, 21(4): 682–697.

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

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

[40] Dobson G, Hasija S, Pinker E. Reserving capacity for urgent patients in primary care. *Production and Operation Management*, 2011, 20(3): 456–473.

[41] Qu X, Rardin RL, Williams JAS. A mean-variance model to optimize the fixed versus open appointment percentages in open access scheduling systems. *Decision Support System*, 2012, 53(3): 554–564.

[42] Patrick J. A Markov decision model for determining optimal outpatient scheduling. *Health Care Management Science*, 2012, 15(2): 91–102.

[43] Lee, S., Min, D., Ryu, J. H., & Yih, Y. A simulation study of appointment scheduling in outpatient clinics: Open access and overbooking. *Simulation*, 2013, 89(12), 1459–1473.

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

[45] Feldman J, Liu N, Topaloglu H, Ziya S. Appointment scheduling under patient preference and no-show behavior. *Operations Research*, 2014, 62(4): 794–811.

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

[47] Balasubramanian H, Muriel A, Wang L. The impact of provider flexibility and capacity allocation on the performance of primary care practices. *Flexible Services and Manufacturing Journal*, 2012, 24(4): 422–447.