Operations Research for Health Care 4 (2015) 5–14

Contents lists available at ScienceDirect

Operations Research for Health Care

journal homepage: www.elsevier.com/locate/orhc

An analysis of overlapping appointment scheduling model in an outpatient clinic

Kelsey Anderson [1], Bichen Zheng, Sang Won Yoon *, Mohammad T. Khasawneh

Department of Systems Science and Industrial Engineering, State University of New York at Binghamton, Binghamton, NY 13902, United States

ARTICLE INFO

Article history:
Received 4 October 2013
Accepted 2 December 2014
Available online 12 December 2014

Keywords:
Appointment scheduling
Overlapping scheduling
Outpatient clinics
Stochastic service time

ABSTRACT

This research addresses an overlapping appointment scheduling (OLAS) model to minimize patient waiting time and doctor idle time in an outpatient healthcare clinic when a stochastic service time is considered. In general, outpatient clinics should determine proper appointment schedules for their patients to maximize doctor utilization and patient satisfaction. As a result, the OLAS model has been proposed to find the optimal overlap period between patient appointment and allocated service times. A mathematical model is developed to minimize the total cost of patient waiting and doctor idle time, which has been analyzed with the assumption that the service time is followed by a uniform distribution. In addition, a Monte Carlo simulation model is developed to verify the optimal overlap period driven from the proposed OLAS model and to evaluate the effect of implementing an overlap period in clinics with different service distributions, overtime, and no-shows. The experimental results indicate that the optimal environment to apply an OLAS model is an outpatient clinic with a high no-show rate, long appointment lengths, and a high coefficient of variation. The results indicate that the utilization of overlapping scheduling can lead to a 40%–70% reduction in total costs.

1. Introduction

In recent years, healthcare systems have shifted care from hospital settings to outpatient facilities due to its improved technology and the need to reduce costs [1]. Once complicated procedures can now be completed in one day, including the recovery period. Moving patients from the hospital to the outpatient facility decreases their exposure to nosocomial infections. The benefits of outpatient treatment have led to a rise in patient volumes at the majority of outpatient clinics, where the ever increasing demands for service have now made efficiency to a concern [2]. Due to increased demands at outpatient clinics, efficient scheduling has become highly important for enabling the maximum number of patients to be seen while minimizing their wait time. A balance must be achieved between patient waiting time and number of patients scheduled daily. The wait time for the appointment needs to be reasonable to ensure that the patients continue to use healthcare services from the same clinic. At the same time, doctors should not remain idle while waiting for patients to arrive, given that they are an expensive resource.

Most of outpatient facilities consist of a queuing system. When the patient arrives at the clinic at his or her specified time, there is no guarantee that the provider will be available. Often the patient must wait until the previous patient's treatment is complete. With such a system, if one patient has an uncharacteristically long service time, or arrives extremely late, chaos is likely to occur. This can also lead to significant customer dissatisfaction because a majority of outpatient patients call ahead to set up an appointment time in advance and expect to be seen at the specified appointment time. Another complexity of the system is patient no-shows, which can affect the clinic's revenue because the doctor is idle and not fully utilized.

For the best efficiency, an outpatient clinic must understand its patient types and clinical environment. Each clinic has unique characteristics, such as patient no-show rate, service time variation, and the presence of walk-ins that require a different scheduling method to be used to optimize performance and minimize costs [3]. Therefore, the objective of this research is to develop an outpatient scheduling model that overlaps appointments by a designated time period and then determine the optimal overlapping period for a specific clinic's characteristics.

Overlapping appointments are defined as the appointments that are scheduled earlier than the expected end time of the previous appointment. It is also called an overlap period. When patients do not arrive for an appointment or doctors finish the treatments earlier, the overlap appointment strategy can reduce

* Corresponding author.
E-mail address: yoons@binghamton.edu (S.W. Yoon).
[1] Current affiliation: Christiana Care Health System, Newark, DE 19718, United States.

http://dx.doi.org/10.1016/j.orhc.2014.12.001
2211-6923/

doctor idle time. It also enables additional patients to be added to the schedule, i.e. walk-in patients without reservations. In these cases, additional profit for the outpatient clinic is gained.

This research aims to develop an overlapping appointment scheduling (OLAS) model that can be used to minimize the total costs (TC) of an outpatient clinic which consists of the doctor idle time (TD), total patient waiting time (TW), and overtime (O). The total cost is defined as

$$\text{minimize} \quad TC = \beta \cdot TD + \omega \cdot TW + \gamma \cdot O \tag{1}$$

where β, ω, and γ represent the doctor idle time, patient waiting time, and overtime cost coefficients, respectively. In this research, mathematical and simulation models are developed to determine the effects of various parameters, such as service time distribution, no-shows, cost coefficients, and service time variation on the performance of the OLAS model. To develop the OLAS model, the following assumptions are made:

- It is a single provider outpatient clinic.
- No walk-ins or baulking.
- The patient/provider is punctual.
- Service durations for each patient are identically and independently distributed.

The overview of this paper is summarized as follows. The background research of outpatient clinic appointment scheduling models is presented in Section 2. The details of the proposed OLAS model and solution methodologies are explained in Section 3. The experimental results and analyses of the OLAS model performance in various clinical environments are discussed in Section 4. Finally, the conclusion, recommendations, and future work are summarized in Section 5.

2. Background

Scheduling problems comprise a large research area in both the service and manufacturing industries. In general, scheduling problems are investigated to optimize task schedules based on time limitations and resource allocations in a production or manufacturing environment. Scheduling in healthcare is more difficult since human nature is involved in the process. There is a high degree of variability in the actual time depending on an array of variables, such as the patient being treated, time of day, or provider's treatment patterns.

Much research has been completed regarding appointment scheduling, specifically. Cayirli and Denton [1] and Gupta and Denton [2] completed thorough reviews of previously used methodologies found in literature and summarized the characteristics of scheduling models. The major concern for Cayirli and Denton [1] and many other researchers is the gap between research and applicability. Many models, even when generalized, are highly specified to one type of clinic's characteristics [4].

Outpatient scheduling models used today are appointment, open access, or walk-in. Open access is a newer appointment system and became prevalent in literature about 10 years ago [5]. Much planning goes into this type of scheduling to determine the proper percentage of appointments that should be scheduled in advance. Open access is optimal in clinics where the demand and supply of appointments are balanced. Various studies have been completed that analyze the effectiveness of an open access schedule [6–11]. To model open access in an appointment scheduling model, a probability distribution such as an exponential distribution can be selected to describe patient inter-arrival times based upon historical patterns [12–14].

The first scheduling method researched in history is not based on appointment. It requires that all patients arrive in one large block at the beginning of the clinic care session [15]. This old scheduling method creates extensive wait time for patients, however, staff is highly utilized. Appointment scheduling evolved and became more patient focused. The benefits of spreading the patients throughout the appointment session were proven in terms of cost and patient waiting time reduction [16]. Today, individual block appointment systems are most common [17]. LaGanga and Lawrence [18] captured the trade-off between the benefits and costs of using overbooking and found that overbooking can often lead to improved provider productivity, clinical performance, and patient access in the individual block appointment system.

The first studies on individual appointment scheduling assumed intervals were spaced evenly throughout the clinic session with appointment lengths of 15 or 30 min. Further research discovered that appointment intervals could be manipulated to help compensate for no-shows, walk-ins, and service length variability. The optimal interval size depends on the clinical environment. An early study [19], which focused on variable appointment intervals, used simulation to test a variety of variable interval appointment rules with traditional appointment rules.

The dome shape is the most versatile appointment rule [14]. The dome role has the appointment block length gradually getting longer throughout the day and then decreasing by the end of the day. Wang [20] started the research on the dome shape and discovered the distributions' positive characteristics based upon using exponential distributions for service duration. Gupta and Denton [2] extended the applicability of the dome shape for arbitrarily distributed patient service time. However, the dome shape was highly affected by the weights given to patient waiting, doctor idle time, and doctor overtime. Their research findings encouraged the development of a new "dome rule". The "dome" rule is found to be more flexible than traditional scheduling rules and the optimal choice under a wide range of environments [14,21].

Dove and Schneider [22] analyzed a large amount of variables to determine that the appointment interval, patient's age, and previous no-show record were most applicable in a sequential decision-tree model when predicting no-shows. The most important characteristic was the patients' previous appointment-keeping pattern [22]. Their studies expanded on these findings and found that costs could be reduced by having the proper sequence of patients based upon each specific patient no-show rate and expected service time [23–25]. Begen and Queyranne [26] developed a mathematical model that found the optimal sequence of jobs for a single processor when the service time was randomly given by a joint discrete probability distribution based upon minimizing overtime costs. They determined a method for optimizing appointment times in cases where the time a patient needs is uncertain. In reality, a patient can show up early, in which case there is a cost associated with customer wait time, or a patient could show up late, in which case there is a cost associated with idle resources (poor resource utilization). Denton and Gupta [24] used stochastic linear programming model with flexible cost structures, such as overtime costs, waiting, and idling costs. In their model, absent patients were considered by adjusting allocated job durations, and their results showed that the stochastic approach works well in practical clinic implementations, where waiting costs were high relative to overtime or idle resource costs. Begen and Queyranne [26] discussed the issues of finding an optimal appointment schedule for a sequence of jobs to reduce the expected costs when the duration of the job is uncertain. The authors developed a model that can find an optimal appointment schedule in polynomial time. The two scenarios were highlighted in their research: (1) processing time distribution is estimated by historical data and the required number of independent samples is identified to obtain a near optimal solution and (2) independent processing times are considered to find an optimal schedule.

Appointment schedule assessment methods can be classified into three categories. They are simulation-based, analytical, and case study. During the beginning of designing and evaluating scheduling rules, simulation was commonly used due to the complexity and randomness involved. The simulation model evaluates

the various scheduling rules, while algorithms are used to find the optimal schedules.

Simulation modeling allows more complexity of outpatient queuing systems to be included. Patient preference, patient punctuality, no-shows, doctor lateness, interruptions, and varying patient types can all be easily added to simulation models [27]. A summary of early applications of simulation in healthcare was written by Jun et al. [28]. A manual Monte-Carlo simulation technique is utilized to determine the optimal fixed appointment interval when the block size is initialized to be two [15]. This scheduling rule is still referenced in the literature and found optimal for certain clinical environments.

The clinical scheduling problem has also been studied by applying analytical models. Analytical methodologies can be classified into either deterministic or stochastic approaches, of which the deterministic approach can be further divided into linear programming, nonlinear programming, dynamic programming, etc. Deterministic programming approaches to finding the optimal clinic schedule often use heuristic algorithms. For example, a local search mechanism is developed to obtain the optimal outpatient schedules by considering time allowances as discrete intervals [3]. When time allowance is considered as a continuous variable, a closed-form heuristic for setting job allowances is proposed by Robinson and Chen [29].

Early analytical models under the stochastic process approach mainly focused on queuing theory, assumed that the system reached steady state, and consisted of a single server queue with constant inter-arrival times and an infinite number of identical customers [30]. Their results are not very accurate because actual appointment systems do not reach a steady state and do not involve an infinite number of patients.

Fries and Marthe [31] were one of the first researchers to use dynamic programming instead of simulation. They approximated the optimal block size for a given appointment slot. Before Robinson and Chen's [29] heuristics for scheduling, most of deterministic approaches were only compared with other published heuristics, so there was no way to determine if their performance was close to the actual optimal performance.

The optimal scheduling policy is typically chosen based on a multi-objective optimization formulation. The weighted combination of patient waiting time, doctor idle time, and doctor overtime are key characteristics to determine what scheduling rule is optimal. The cost of doctor idle time and overtime can usually be determined based upon the doctor's salary, where overtime is 1.5 times the standard salary.

3. The OLAS model

The objective of this research is to determine the effect of overlapping appointments in an outpatient clinic setting. A traditional appointment system (TAS) with an overlapping scheduling system is illustrated in Fig. 1 where I_n is the appointment length for an appointment slot n, L is the overlap period, and A_n is the scheduled appointment time for an appointment slot n. Fig. 1 illustrates that the overlap period can decrease both overtime and doctor idle time without drastically increasing patient waiting time. The details of the proposed OLAS models are addressed as below.

3.1. Mathematical model

For simplicity purposes, it was assumed that the appointment length is represented as a uniform distribution in the mathematical model. The model was developed to determine the overlap period, which could be utilized to reduce the total cost based upon patient waiting and doctor idle time. The expected total cost is composed of the expected doctor idle time and patient waiting time defined as

$$E[TC] = \beta \cdot E[TD] + \omega \cdot E[TW] \tag{2}$$

where β and ω represent the doctor idle and patient waiting time cost coefficients, respectively. Assuming that the doctor is idle given such that the service time ends before the next appointment is scheduled, $E[TD]$ can be defined as

$$E[TD] = \sum_{n=1}^{N} (\mu - L - I_n) \cdot P\{I_n < (\mu - L)\} \tag{3}$$

where $N = \lceil \frac{Z}{\mu} \rceil$ is the number of patients seen in a clinic session, Z is the clinical session length, n is the index of the appointment slot, L is the overlap period, μ is the mean appointment time, and I_n is a random service time. The actual appointment interval becomes $\mu - L$. If $I_n < (\mu - L)$, the doctor is idle. On the other hand, the patient will wait if $I_n > (\mu - L)$. Then, $E[TW]$ can be defined as

$$E[TW] = \sum_{n=1}^{N} (I_n - \mu + L) \cdot P\{I_n > (\mu - L)\}. \tag{4}$$

Assuming that I_n is uniformly distributed, $P\{I_n < (\mu - L)\}$ is based on the cumulative distribution function, which for the uniform distribution is $\frac{\mu - a}{b - a}$. Therefore, Eq. (2) can be revised as

$$E[TC] = \beta \sum_{n=1}^{N} \left[(\mu - L - I_n) \left(\frac{\mu - L - a}{b - a} \right) \right] + \omega \sum_{n=1}^{N} \left[(I_n - \mu + L) \left(1 - \frac{\mu - L - a}{b - a} \right) \right]. \tag{5}$$

By using the uniform distribution, the first and second derivatives of $E[TC]$ can be calculated as

$$\frac{\partial E[TC]}{\partial L} = -\frac{\beta(\mu - L - a)}{b - a} - \frac{\beta(\mu - L - I_n)}{b - a} + \omega \left(1 - \frac{\mu - L - a}{b - a} \right) + \frac{\omega(I_n - \mu + L)}{b - a} \tag{6}$$

$$\frac{\partial^2 E[TC]}{\partial^2 L} = \frac{2\beta}{b - a} + \frac{2\omega}{b - a}. \tag{7}$$

Since β, $\omega > 0$ and $a < b$, $\frac{\partial^2 E[TC]}{\partial^2 L} > 0$ and $E[TC]$ is a strictly convex function, which indicates that there is a minimum cost value for a specified overlap value. Therefore, the sufficient optimality condition to minimize $E[TC]$ is $\frac{\partial E[TC]}{\partial L} = 0$, which gives us

$$L = \frac{-\beta a - \beta I_n + 2\beta\mu - \omega b - \omega I_n + 2\omega\mu}{2(\beta + \omega)}. \tag{8}$$

Since a random variable for the service time is used, the expected value of the overlap period should be calculated by using the Riemann–Stieltjes integral where the integral is taken of the function multiplied by the probability distribution of I_n from the lowest limit (a) to the upper limit (b). The expected value of the overlap period is then written as follows

$$E(L) = \frac{\beta(4\mu - b - 3a) + \omega(4\mu - a - 3b)}{4(\beta + \omega)}. \tag{9}$$

By replacing μ with $\frac{a+b}{2}$, Eq. (9) can be revised as

$$E(L) = \frac{\beta(-b - a) + \omega(-a - b)}{4(\beta + \omega)}. \tag{10}$$

Eq. (9) shows an interesting relationship between the cost, service time variability (distribution parameters), and mean service time, which will be further discussed in Section 4.

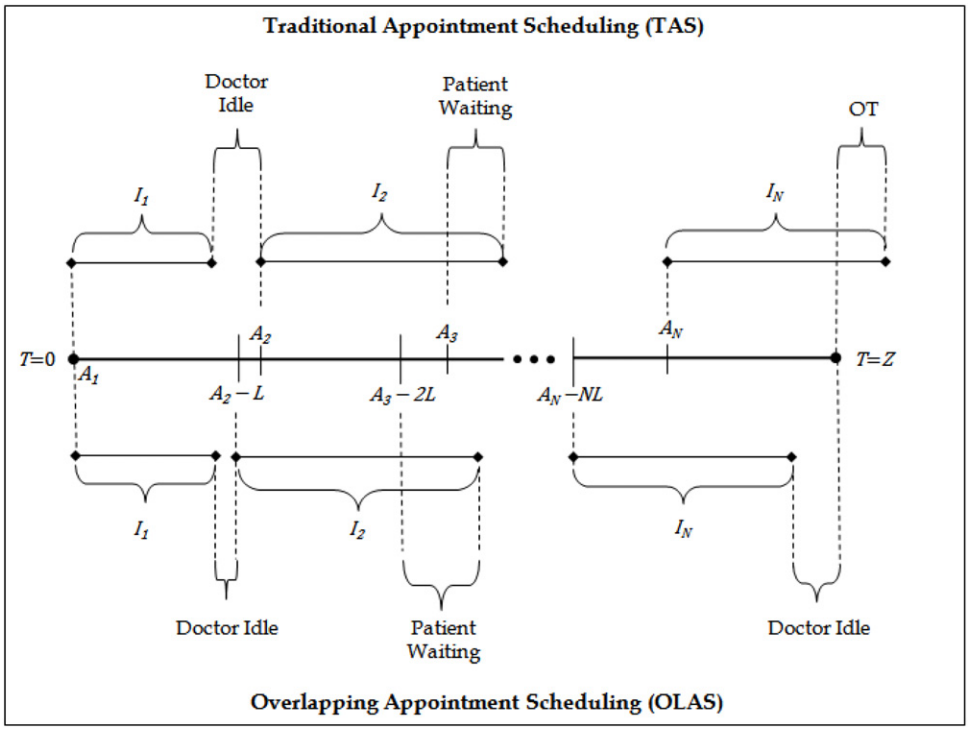


Fig. 1. An illustration between OLAS and TAS.

3.2. Monte-Carlo simulation model

When the doctor service time is uniformly distributed, the expected overlap period is calculated as Eq. (10). However, it would be extremely difficult to obtain the optimal overlap period by evaluating the first and second derivatives of $E[TC]$ if other distributions are considered. Due to this computational complexity, a Monte-Carlo simulation model is developed in Mathematica to determine the minimum cost of various overlap periods. The mathematical model is used to obtain the optimal overlap period with minimum cost when the doctor service time (I_n) is uniformly distributed, while the simulation model is used when the doctor service time follows other distributions. The main simulation procedure is illustrated in Fig. 2. The simulation model is replicated until the optimal overlap period is stable or the desired variation of overlap period is achieved.

To validate and test the solutions under different conditions, the simulation model also incorporates overtime costs and no-shows, which are common occurrences in outpatient clinics. If a no-show occurs, $I_n = 0$ and the doctor will be idle until the next patient's arrival, which is defined as

$$I_n = \begin{cases} 0 & \text{if no-show} \\ I_n & \text{otherwise.} \end{cases} \tag{11}$$

In addition, the amount of overtime can be defined as

$$O = -Z + \sum_{n=1}^{N} \begin{cases} I_n & \text{if } I_n > (\mu - L) \\ \mu - L & \text{if } I_n \le (\mu - L). \end{cases} \tag{12}$$

If $I_n \le (\mu - L)$, it can be seen that treatment time is the allotted time of $\mu - L$. However, if $I_n > (\mu - L)$, the treatment time changes to I_n. This was evaluated for each appointment slot, and then Z (clinical session length) was subtracted from the total time allotted for all appointments to determine the overtime.

4. Numerical results and analyses

In this section, numerical results in the OLAS models illustrate how various system parameters affect the performance of the OLAS models and in what clinical environment it will provide the largest benefit. The performance of the OLAS model is analyzed based upon the cost reduction and optimal overlap period. Two system parameters were examined in the simulation model, which were the coefficient of variation, CV, and the cost ratio, CR. Based on the literature, the coefficient of variation is commonly used to measure variability in treatment times, which is defined as

$$CV = \frac{\sigma}{\mu}. \tag{13}$$

The cost ratio for outpatient clinics is typically based on a ratio between costs of doctor idle time and patient waiting time, which is defined as:

$$CR = \frac{\beta}{\omega}. \tag{14}$$

In this model, the CR is modified to include overtime cost, assuming that the overtime rate is 1.5 times that of the doctor idle cost rate since a typical overtime rate for any employee is 150% of their normal salary. Then, the cost ratio is redefined as:

$$CR = \frac{\beta + \gamma}{\omega} \tag{15}$$

where γ is 1.5β. These parameter settings are summarized in Table 1. To explore the most beneficial clinic environment for the model, wider ranges of parameters are selected. Combinations of parameter values are tested and validated via the simulation modeling to evaluate the best environment in which to apply the optimal solutions obtained.

4.1. Uniformly distributed doctor service time

In this series of experiments, doctor service time is assumed to be uniformly distributed. The optimal overlap period is obtained

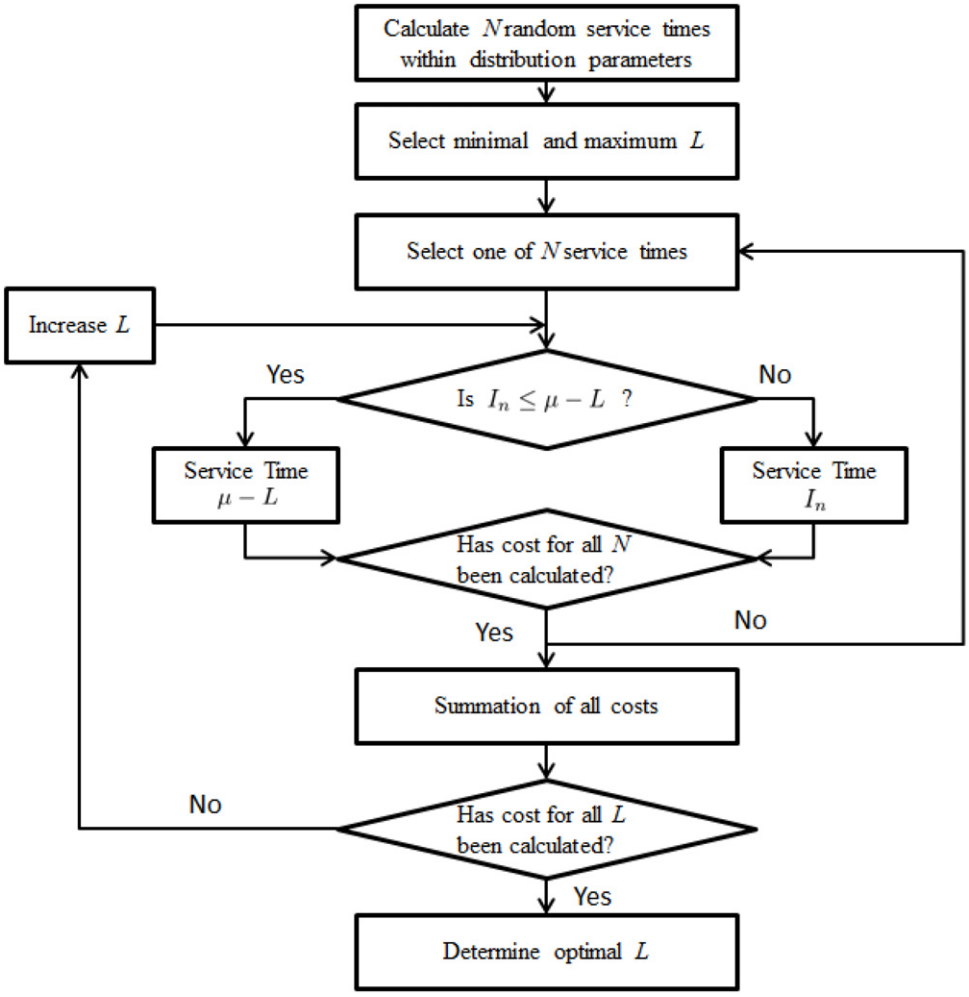


Fig. 2. Simulation model flow map for determining optimal overlap period.

by Eq. (8) in the mathematical model. Then, the cost reduction is analyzed based on the simulation model.

- Cost ratio analysis.
  The OLAS model was analyzed to examine the impact of the *CR* by using the mathematical model based on uniformly distributed doctor service time. The *CR* was analyzed for four uniform distributions with parameters that are typically found in an outpatient clinic environment. When the cost ratio is larger than 20, the optimal overlap period increases to a steady state value, which is not significantly affected by the cost ratio. When the cost ratio is small, meaning that, the doctor idle time and patient waiting time cost coefficients are close to each other, the cost ratio plays an important role in determining the overlap period. The numerical simulation results are shown in Fig. 3.
  According to the results of the simulation, Fig. 4 illustrates the total cost reduction vs. different cost ratios under the assumption of a 10% patient no-show rate. Although the total cost varies for different *CR*s, all minimum total costs occur at an overlap period above zero. In other words, the simulation results confirm that the overlap period scheduling strategy provides a total cost reduction as opposed to the traditional scheduling strategies without overlap. Compared to the schedule without the overlap period, the schedule with the overlap period provides a significant cost reduction, as seen in Fig. 5. The cost reduction is defined as the percentage reduction in total cost of L^* compared to $L = 0$. The cost reduction is compared within cases where the only difference is the overlap period length. All the remaining clinic settings are the same, such as doctor service time distribution, the coefficient of variation, etc. Fig. 5 shows the percentage of cost reduction based on various cost ratio (*CR*) values. This is a similar pattern to L^* compared to *CR*, which is illustrated in Fig. 3. Therefore, when $L^* > 20$, a further increase in the *CR* does not affect the cost reduction and the optimal overlap period significantly.
- Analysis of service time mean and coefficient of variation.
  Service time and coefficient of variation have been compared for L^* at various *CV* and μ values using the mathematical model when $CR = \frac{3}{2}$, as shown in Fig. 6. The restriction of the uniform distribution is that the maximum $CV = 0.58$. For $\mu = 30$, the L^* increases by 0.4 min for every 0.1 increment in *CV*. Also, for $\mu = 15$, L^* rises 0.2 min for every 0.1 increment in *CV*. In practice, a short overlap period may not have a significant impact on the overall schedule or operational cost. For example, an overlap period is less than 1 min, which could be difficult to measure and implement in the actual patient schedule. The operational cost reduction is also low due to the short overlap period. Thus, a longer overlap period is more desirable to clinics. Based on the previous analysis, outpatient clinics with a higher *CV* achieve a higher L^*, which is easier to implement in the model.
  In addition, Fig. 7 illustrates the relationship between μ and L^*. It also shows L^* is positively correlated to *CV*. Since a larger optimal overlap period is preferred in practice, outpatient clinics with a higher μ achieve a higher L^*, which allows the OLAS model to become more practical.

Table 1
Summary of parameter settings.

Parameter	Description	Values/units
CR	Cost ratio	[0, 20]
CV	Coefficient of variation	[0.36, 0.58] (min)
I_n	Stochastic doctor service time	$U(a, b)$, $\exp(\lambda_1)$, $Erl(\kappa, \lambda_2)$, $N(\mu, \sigma)$
L	Overlap period	[−30, 30] (min)
a	Lower parameter of uniform distribution	[0, 5, 10] (min)
b	Upper parameter of uniform distribution	[25, 30, 50, 60] (min)
μ	Mean of historical service time	[5, 60] (min)
σ	Standard deviation of historical service time	[7.5, 90] (min)
λ_1	Rate parameter of exponential distribution	$\frac{1}{15}$
λ_2	Rate parameter of Erlang distribution	$\frac{1}{30}$
κ	Shape parameter of Erlang distribution	2
ρ	No-show rate	[0, 40%]

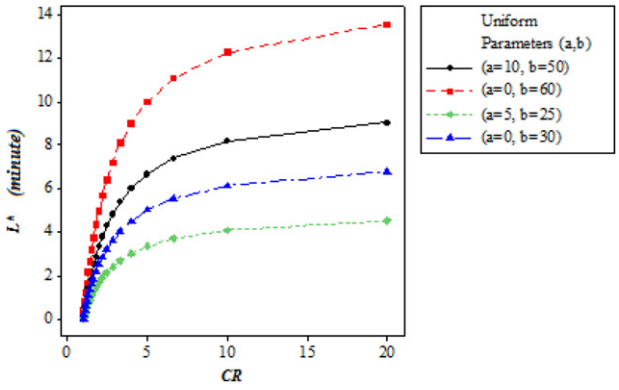


Fig. 3. Optimal overlap period vs. cost ratios.

- Analysis of total cost with overtime.
  To examine more complex environments, the total cost was analyzed using Monte Carlo simulation model while adding the consideration of doctor overtime cost and patient no-show rate. Fig. 8 illustrates the wait, idle, and overtime costs when all are equally weighted, or in other words have a cost ratio of 2. The no-show rate of 10% is shown to illustrate the effect of no-shows on costs. Total cost follows similar patterns when the no-show rate is changed, which is illustrated in Fig. 9. The total cost of the higher no-show rate is lower because it is assumed if the doctor finishes giving treatment to all patients earlier than the clinic session length, the doctor will not be idle for the remaining part of the session. Also, with a high no-show rate there is a lower probability that overtime will occur. In other words, with higher no-show rates, the doctor often finishes treatment early. According to Figs. 4, 8 and 9, the minimum total cost achieves a positive overlap period, which validates the hypothesis that the schedule with the overlap period may decrease the total cost.

The previous discussion shows that the overlapping appointment scheduling strategy can reduce operational costs for clinics. Using the analytical model, the optimal overlap period can be obtained, which results in reduced system costs. The simulation model provides a good test-bed for implementation. It confirms that the overlap period can decrease the total operational cost by considering the overtime cost and patient no-show rate. Based on the simulation results, it can be concluded that the most beneficial environment for an OLAS is when the CR is high within the range

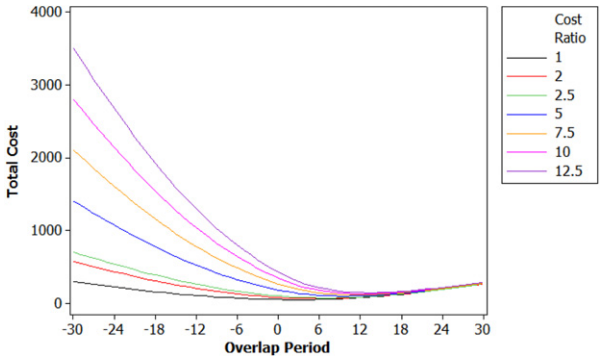


Fig. 4. Total cost for various overlap periods and cost ratios for a uniform distribution with 10% no-shows.

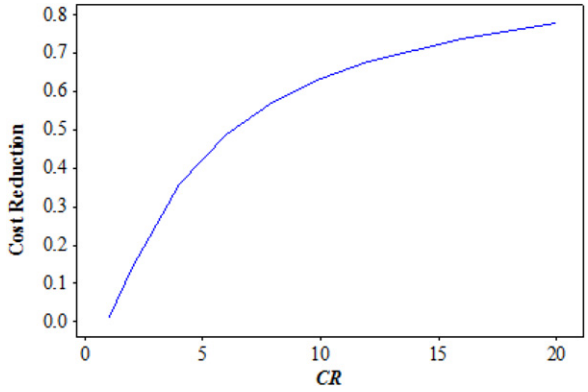


Fig. 5. Percentage cost reduction against $L = 0$ vs. cost ratio for uniform distribution.

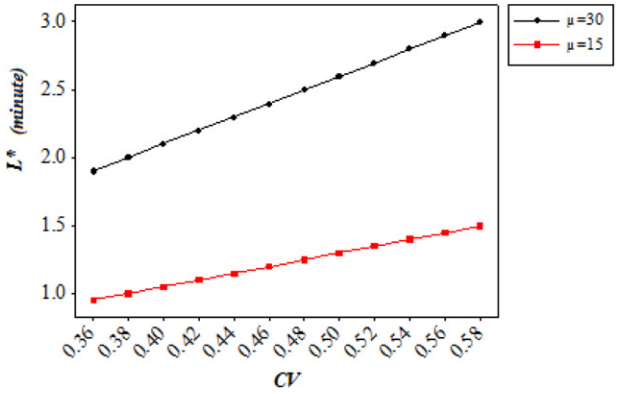


Fig. 6. Optimal overlap period vs. coefficient of variation.

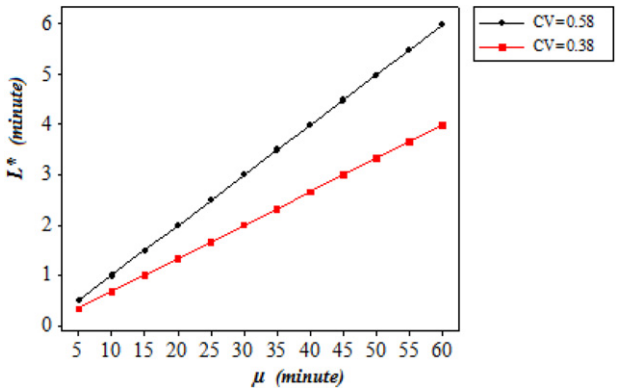


Fig. 7. Optimal overlap period vs. mean service time ($CR = 3/2$).

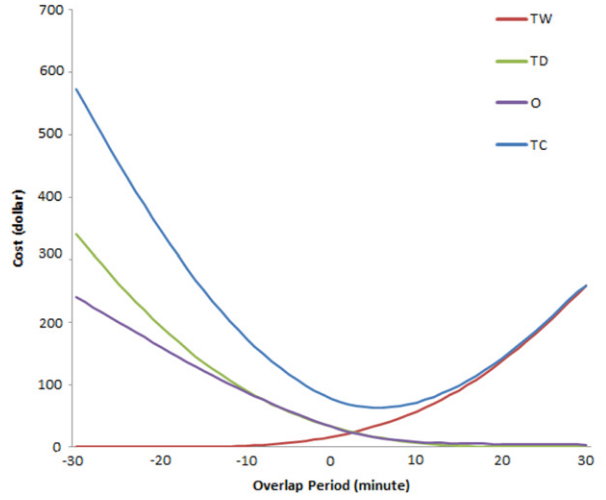


Fig. 8. Wait, idle, and overtime cost for various overlap periods for a uniform distribution with 10% no-shows ($a = 10$, $b = 50$, $CR = 2$, $CV = 0.38$, $\mu = 30$).

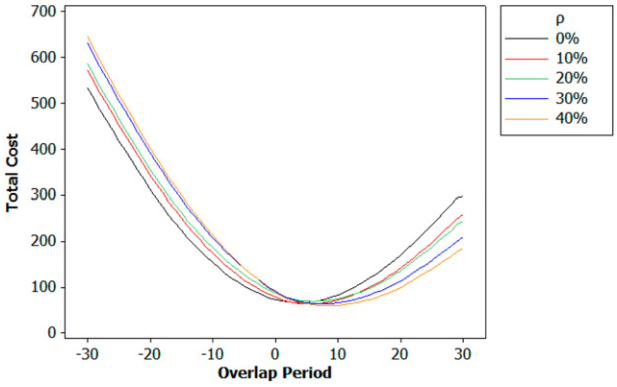


Fig. 9. Total cost for various overlap periods and NS rates for a uniform distribution (cost ratio = 2).

[0, 20]. A high historical average service time leads to a longer optimal overlap period. The longer optimal overlap period makes the model more easily applied in real life, since appointments are usually set at 5 min intervals. A small overlap period is difficult to implement because of difficulties in accurately tracking the overlap period, especially when L is less than 1 min.

4.2. Different doctor service time distributions

- Analysis of total cost with overtime.
  Given a patient no-show rate ($\rho = 10\%$) and a cost ratio ($CR = 2$, $\beta : \gamma : \omega = 1 : 1 : 1$), Fig. 10 illustrates the relationship of TW, TD, and O for various L values. The analysis was completed using the simulation model. TW for all distributions increases exponentially as L increases. As L increases, TD approaches 0, which indicates that a longer overlap period makes doctors much busier. The minimum cost for all distributions occurs when the overlap period is greater than 0, which validates the hypothesis that overlapping scheduling strategies reduce costs when compared to a traditional strategy under a wide range of service time distributions.
  A total of 100 simulation runs were used to validate the total cost savings for each doctor service time with an optimal overlap period. The procedure of the simulation follows Fig. 2.
- Analysis of cost ratio and no-show rate.
  Various CR and ρ values are compared with regard to L^* and cost reduction using the simulation model, where ρ represents the no-show rate of patients. Fig. 11 illustrates the optimal overlap period (L^*) when various CR and ρ values are applied. It shows that L^* is positively correlated to CR, but at a decreasing rate. ρ has the same pattern but does not have a significant effect on L^*. Take the uniform distributed doctor service time contour plot as an example: given a fixed cost ratio ($CR = 2$), by changing the patient no-show rate ($\rho = 0$–40%), the optimal overlap period does not significantly vary (stays in the range [5.0–7.5]). Fig. 12 shows cost reductions for various CR and ρ values, where the cost reduction percentage is obtained by comparing TC when L^* and $L = 0$. In other words, the cost is compared between overlapped and non-overlapped scheduling. It also reveals that for the four typical doctoral service time distributions, overlapped scheduling saves more than 40% total cost, when cost ratio is higher than 8 and patient no-show rate is above 30%. For the same patient failure appointment rates, a higher cost ratio leads to a higher cost reduction for the clinics. In summary, under a clinical environment of a high patient failure appointment rates and cost ratio, the proposed overlapped scheduling model outperforms over the traditional non-overlapped scheduling rules.

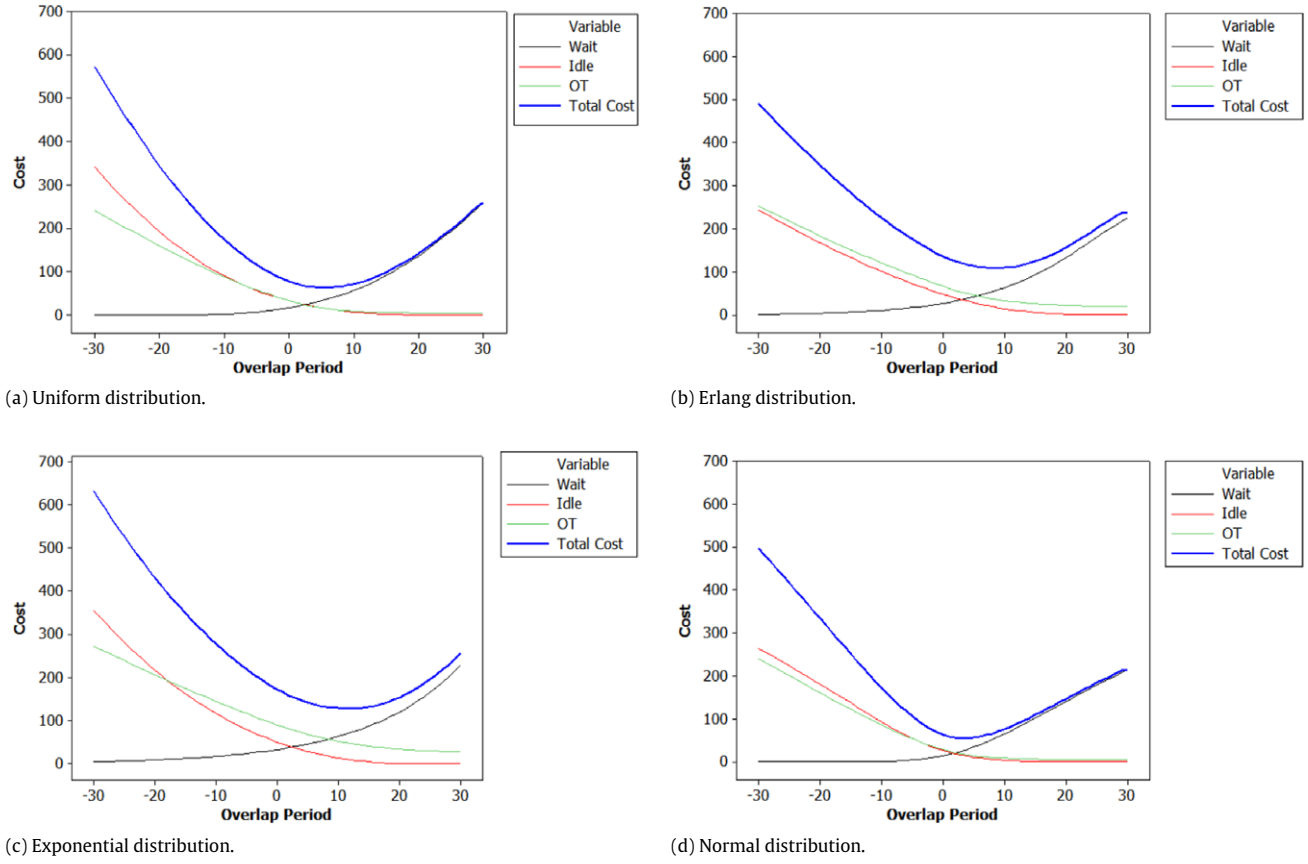
Table 2
Summary of optimal overlap and cost reduction rate.

Doctor service time distributions	ρ	CR	L^*	Cost reduction
Uniform distribution $I_n \sim U(a, b)$	0	2	3.8	9.8%
	0	10	10.5	47.8%
	15	2	5.8	19.7%
	15	10	13.2	63.2%
	30	2	7.5	30.3%
	30	10	15.3	70%
Exponential distribution $I_n \sim \exp(\lambda_1)$	0	2	9.5	19.7%
	0	10	18.9	49.6%
	15	2	12.0	30.9%
	15	10	19.9	64.2%
	30	2	13.5	43.4%
	30	10	20.2	74.5%
Erlang distribution $I_n \sim Erl(\kappa, \lambda_2)$	0	2	7.4	18.0%
	0	10	15.8	51.8%
	15	2	9.0	24.4%
	15	10	17.9	57.7%
	30	2	11.1	33.8%
	30	10	20	73.0%
Normal distribution $I_n \sim N(\mu, \sigma)$	0	2	2.3	5.7%
	0	10	7.9	41.2%
	15	2	4.0	15.9%
	15	10	10.6	58.2%
	30	2	5.6	26.4%
	30	10	12.7	68.7%

The analysis implies that the OLAS model can improve the appointment scheduling system under all the distributions tested when the cost ratio and no-show rate is high. Table 2 summarizes the results for each distribution by showing the L^* and cost reduction rate for a high and low CR, and the ρ determined from the simulation model. The benefit of the OLAS model is observed even when all patients arrive for their appointments, and $CR \geq 2$. In this case, there was at least a 20% decrease in cost for all service distributions and $L \geq 5$ for all service distributions except the normal distribution. Compared to CR, the no-show rate (ρ) had less impact on the variation of both cost reduction and L. For all distributions, a larger L leads to an exponentially rise in TW time. However, if $L \leq 10$ in all cases, the average patient waiting time never exceeds 15 min. The service distributions that benefited the most from the OLAS model were the exponential and Erlang distributions, while the least benefited model was for the normal distribution.

5. Conclusion and future work

In this research, OLAS models were developed for an outpatient clinic. Various parameter combinations were tested using mathematical and Monte Carlo simulation models. The optimal overlap



(a) Uniform distribution.

(b) Erlang distribution.

(c) Exponential distribution.

(d) Normal distribution.

Fig. 10. Wait, idle, overtime and total cost for various overlap periods with $\rho = 10\%$ no-shows and $CR = 2$ ($\beta : \gamma : \omega = 1 : 1 : 1$).

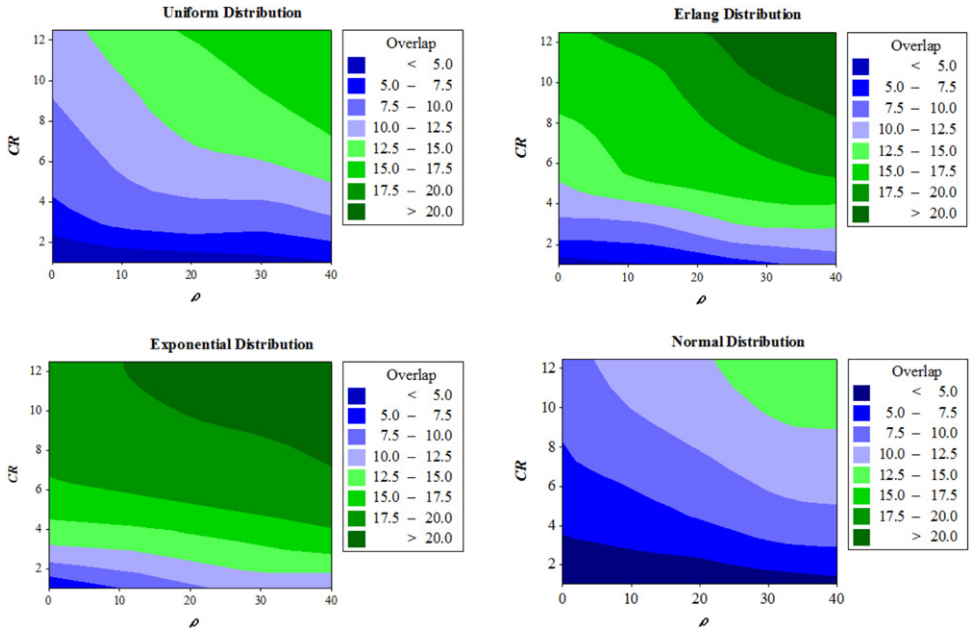


Fig. 11. Contour plot of optimal overlap period vs. cost ratio and no-show rate.

period was selected based upon its ability to minimize the cost of patient waiting time, doctor idle time, and overtime. The results were compared with a scheduling system that had appointments scheduled at even intervals, based upon the service mean throughout the clinical session. The parameters that were analyzed and compared were the mean service time, coefficient of variation for the service, cost ratio of patient waiting compared with total doctor idle time and overtime, and the no-show rate. The numerical results indicate that the OLAS models can significantly reduce the cost for most clinical environments and be effective when there are high no-shows and doctor costs.

In summary, the largest overlap period and cost reduction are achieved in an outpatient clinic with a high service time, coefficient of variation for service time, cost ratio, and no-show rate. The

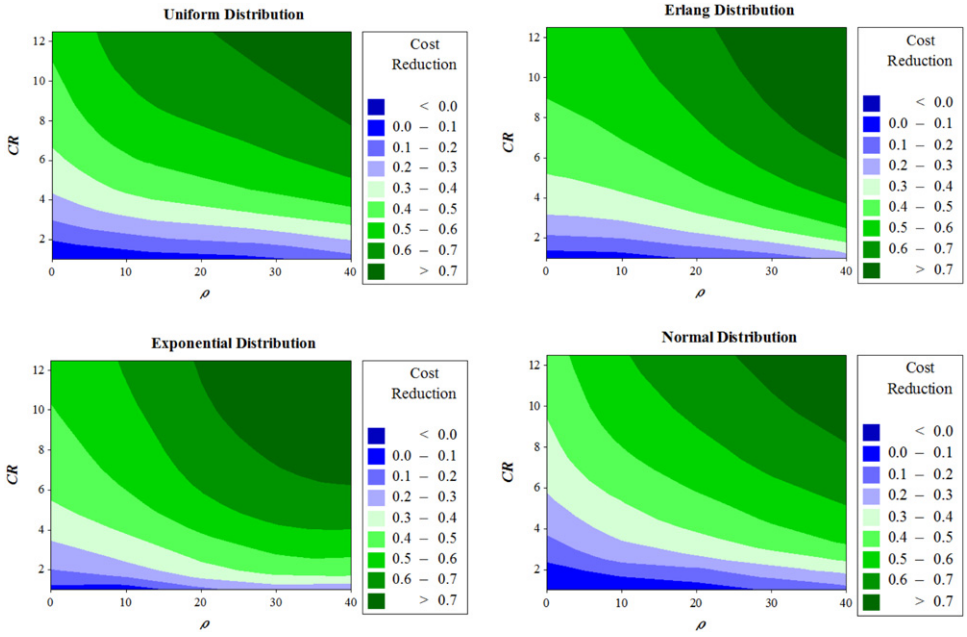


Fig. 12. Contour plot of cost reduction vs. cost ratio and no-show rate.

distribution of the service time varies the no-show patient and the cost ratio impacts on the optimal overlap period. The mean service time and coefficient of variation do not affect the cost reduction for the outpatient clinic.

The appointment scheduling model developed in this research could be applied to a real-life clinical environment in the future. The results and analyses would significantly benefit the clinics and patients. The model can be expanded into a variety of areas and can be improved to include more characteristics observed in an outpatient clinic. This research was to approach appointment scheduling by looking at overlapping appointments. The model was based upon characteristics of an outpatient clinic, but could easily be applied to other appointment systems both in the healthcare arena, and the general service industry. Moreover, the overlapping appointment scheduling model can be expanded to include multiple providers.

The mathematical model can be further developed to include more of the complexities of an outpatient clinic such as no-shows and overtime costs. In addition, the proposed OLAS models can be modified by using a profit function. This would allow the OLAS models to schedule walk-in patients in the buffer period at the end of the clinical session. The OLAS models can be further developed by adding techniques tried in previous research, such as double-booking, patient classification, and/or varying appointment intervals, and then, a hybrid model could be developed.

Acknowledgments

This research was supported by the Watson Institute for Systems Excellence at State University of New York at Binghamton. The authors also would like to thank Megan Smey and anonymous reviewers for the valuable suggestions to improve this paper.

References

[1] T. Cayirli, E. Veral, Outpatient scheduling in health care: a review of literature, Prod. Oper. Manage. 12 (4) (2003) 519–549.
[2] D. Gupta, B. Denton, Appointment scheduling in health care: challenges and opportunities, IIE Trans. 40 (9) (2008) 800–819.
[3] G.C. Kaandorp, G. Koole, Optimal outpatient appointment scheduling, Health Care Manage. Sci. 10 (3) (2007) 217–229.
[4] H.-S. Lau, A.H.-L. Lau, A fast procedure for computing the total system cost of an appointment schedule for medical and kindred facilities, IIE Trans. 32 (9) (2000) 833–839.
[5] X. Qu, J. Shi, Modeling the effect of patient choice on the performance of open access scheduling, Int. J. Prod. Econ. 129 (2) (2011) 314–327.
[6] S. Herriott, Reducing delays and waiting times with open-office scheduling, Fam. Pract. Manag. 6 (4) (1999) 38–43.
[7] D.G. Bundy, G.D. Randolph, M. Murray, J. Anderson, P.A. Margolis, Open access in primary care: results of a north Carolina pilot project, Pediatrics 116 (1) (2005) 82–87.
[8] L.W. Robinson, R.R. Chen, A comparison of traditional and open-access policies for appointment scheduling, Manuf. Serv. Oper. Manage. 12 (2) (2010) 330–346.
[9] S. Lee, Y. Yih, Analysis of an open access scheduling system in outpatient clinics: a simulation study, Simulation 86 (8–9) (2010) 503–518.
[10] X. Qu, R.L. Rardin, J.A.S. Williams, A mean–variance model to optimize the fixed versus open appointment percentages in open access scheduling systems, Decis. Support Syst. 53 (3) (2012) 554–564.
[11] X. Qu, R.L. Rardin, J.A.S. Williams, Single versus hybrid time horizons for open access scheduling, Comput. Ind. Eng. 60 (1) (2011) 56–65.
[12] J. Vissers, Selecting a suitable appointment system in an outpatient setting, Med. Care 17 (12) (1979) 1207–1220.
[13] J.R. Swisher, S.H. Jacobson, J.B. Jun, O. Balci, Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation, Comput. Oper. Res. 28 (2) (2001) 105–125.
[14] T. Cayirli, K.K. Yang, S.A. Quek, A universal appointment rule in the presence of no-shows and walk-ins, Prod. Oper. Manage. 21 (4) (2012) 682–697.
[15] N.T. Bailey, A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times, J. R. Stat. Soc. Ser. B Stat. Methodol. 14 (2) (1952) 185–199.
[16] W.L. Johnson, L.S. Rosenfeld, Factors affecting waiting time in ambulatory care services, Health Serv. Res. 3 (3) (1968) 286.
[17] T.R. Rohleder, K.J. Klassen, Using client-variance information to improve dynamic appointment scheduling performance, Omega 28 (3) (2000) 293–302.
[18] L.R. LaGanga, S.R. Lawrence, Clinic overbooking to improve patient access and increase provider productivity, Decis. Sci. 38 (2) (2007) 251–276.
[19] C.-J. Ho, H.-S. Lau, Minimizing total cost in scheduling outpatient appointments, Manage. Sci. 38 (2) (1992) 1750–1764.
[20] P.P. Wang, Static and dynamic scheduling of customer arrivals to a single-server system, Naval Res. Logist. (NRL) 40 (3) (1993) 345–360.
[21] K.J. Klassen, R. Yoogalingam, Improving performance in outpatient appointment services with a simulation optimization approach, Prod. Oper. Manage. 18 (4) (2009) 447–458.
[22] H.G. Dove, K.C. Schneider, The usefulness of patients' individual characteristics in predicting no-shows in outpatient clinics, Med. Care 19 (7) (1981) 734–740.
[23] P.M.V. Bosch, D.C. Dietz, Scheduling and sequencing arrivals to an appointment system, J. Serv. Res. 4 (1) (2001) 15–25.
[24] B. Denton, D. Gupta, A sequential bounding approach for optimal appointment scheduling, IIE Trans. 35 (11) (2003) 1003–1016.
[25] S. Chakraborty, K. Muthuraman, M. Lawley, Sequential clinical scheduling with patient no-shows and general service time distributions, IIE Trans. 42 (5) (2010) 354–366.

[26] M.A. Begen, M. Queyranne, Appointment scheduling with discrete random durations, Math. Oper. Res. 36 (2) (2011) 240–257.
[27] J. Luo, V.G. Kulkarni, S. Ziya, Appointment scheduling under patient no-shows and service interruptions, Manuf. Serv. Oper. Manage. 14 (4) (2012) 670–684.
[28] J.B. Jun, S.H. Jacobson, J.R. Swisher, Application of discrete-event simulation in health care clinics: a survey, J. Oper. Res. Soc. 50 (2) (1999) 109–123.
[29] L.W. Robinson, R.R. Chen, Scheduling doctors' appointments: optimal and empirically-based heuristic policies, IIE Trans. 35 (3) (2003) 295–307.
[30] A. Mercer, A queueing problem in which the arrival times of the customers are scheduled, J. R. Stat. Soc. Ser. B Stat. Methodol. 22 (1960) 108–113.
[31] B.E. Fries, V.P. Marathe, Determination of optimal variable-sized multiple-block appointment systems, Oper. Res. 29 (2) (1981) 324–345.