Socio-Economic Planning Sciences 47 (2013) 205–219

Contents lists available at SciVerse ScienceDirect

Socio-Economic Planning Sciences

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

# Sequential clinical scheduling with patient no-show: The impact of pre-defined slot structures

Santanu Chakraborty[a], Kumar Muthuraman[b], Mark Lawley[c,*]

[a] Enterprise Optimization, United Airlines, Chicago, IL, USA
[b] McCombs School of Business, University of Texas, Austin, TX, USA
[c] Weldon School of BioMedical Engineering, Purdue University, 206 S. Martin Jischke Drive, West Lafayette, IN 47907, USA

ARTICLE INFO

*Article history:*
Available online 19 December 2012

*Keywords:*
Overbooking
Health care
Clinical scheduling
No-shows
Primary care clinics

ABSTRACT

This paper develops a sequential scheduling algorithm for consultation periods not divided into slots. Patients call a scheduler and request appointments with a specified provider. The scheduler provides the patient with an appointment time before the call terminates. In making the appointment, the scheduler cannot alter the appointments of previously scheduled patients. Service times are random and each scheduled patient has a probability of "no-showing" for the appointment. Each arriving patient generates a given amount of revenue, and costs are incurred from patient waiting and provider overtime. The scheduling method selects the calling patient's appointment time by minimizing the total expected cost. We prove that total expected cost is a convex function of appointment times and that the expected profit of the schedule is unimodal, which provides a unique stopping criterion for the scheduling algorithm. Computational studies compare this approach with no-show based sequential scheduling methods for out-patient clinics where a predefined slot structure is assumed. The new method yields higher expected profit and less overtime than when service periods are pre-divided into slots. Because slot scheduling is ingrained in healthcare, we use the model to design slot structures that account for no-show and service time variation.

© 2012 Elsevier Ltd. All rights reserved.

## 1. Introduction

Recent years have seen a resurgence in appointment scheduling research. This is due in part to the urgent need for improved access to outpatient clinical care. Although significant research exists in clinical scheduling, it is not widely applied due to inappropriate simplifying assumptions and implementation barriers. For example, the majority of clinical scheduling is sequential in nature, meaning that schedules are built incrementally as customers call for appointments. Further, complex patient behaviors, such as no-show, cancellation, and "leaving without being seen" are rarely accounted for in analytical work. These types of environmental complexities render much of the published literature inapplicable. In recent work, we have addressed this problem by developing sequential scheduling techniques that account for patient no-show behavior and stochastic general service times [3,21,23,24]. These approaches attempt to maximize clinic profit by fully utilizing schedule capacity while controlling for patient waiting time and staff overtime. These techniques are shown to outperform both current practice and existing heuristics. They also provide optimal stopping rules which indicate when the schedule is "saturated".

These algorithms, as is the case in almost all other papers in literature, assume that the entire service period is pre-divided into a fixed number of slots. In most cases, the slot length is set arbitrarily, sometimes as a function of the process service time, but more often as some convenient interval. In this research we develop a sequential scheduling algorithm for the situation where the service period is continuous, i.e. not pre-divided into slots. Numerical experiments show that the expected profit obtained from this method is higher compared to a situation where the service period is divided into slots. Though a continuous service period model is more flexible, fixed time slots have several administrative conveniences. Hence, we also investigate how a fixed slot structure can be designed which yields comparable performance to our continuous scheduling method.

The paper is structured as follows. Section 2 gives a comprehensive review of related work. Section 3 presents our model formulation, the scheduling algorithm, and the derivation of various

* Corresponding author.
*E-mail addresses:* santanu.chakraborty02@gmail.com (S. Chakraborty), kumar.muthuraman@mccombs.utexas.edu (K. Muthuraman), malawley@purdue.edu, malawley09@yahoo.com (M. Lawley).

0038-0121/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.seps.2012.11.003

expressions that are necessary for the scheduling algorithm. Section 4 provides the theoretical guarantees that establish the convexity of the expected cost function and unimodality of the expected profit function. Section 5 presents a comparative study of the performance of the algorithm developed in this work to those using fixed slot structures. Furthermore, using numerical experiments we illustrate that a better slot structure can be designed by taking into account the properties of the service time and the distribution of no-show probabilities within the customer panel. Finally, Section 6 provides the concluding remarks.

## 2. Literature review

In this section, we provide brief reviews of work related to patient no-show and clinical scheduling. Patient no-show has been a problem in healthcare for many years (see Lee et al. [17]). No-show is defined as a scheduled appointment which is not canceled for which the patient does not appear. Such missed appointments mean a loss of clinic revenue and a loss of opportunity for some other patient to be seen. Pesata et al. [22] reports that in a family care clinic, a total of 14,000 appointments were missed in one year (1996), resulting in an estimated loss of over one million dollars.

Studies suggest that patient no-show probabilities can be estimated based on various factors such as, patient demographics, prevailing environmental conditions, and lead time to appointment (see Dervin et al. [6], Deyo and Inui [7], Goldman et al. [9], Gruzd et al. [11], Martinez et al. [20], Cashman et al. [1], Lee et al. [17]). Our own research shows that significant improvements in scheduling efficiency can be achieved when more accurate no-show modeling is in place [4].

Recent research incorporating patient no-show into scheduling decisions includes Erdogan and Denton [8], Gupta and Wang [12], Hutzschenreuter [13], Green and Savin [10], Kim and Giachetti [15], Kaandorp and Koole [14], Muthuraman and Lawley [21], Chakraborty et al. [3], Zeng et al. [24], Turckan et al. [23], and Luo et al. [19].

Erdogan and Denton [8] extend the work of Denton and Gupta [5], which studies the problem of determining the optimal appointment times for a sequence of jobs with stochastic duration. They express the model as a two-stage stochastic linear program that minimizes the expected cost of customer waiting, server idle time, and a cost of tardiness. Exploiting the problem structure, they derive upper bounds that are independent of the service time distribution. Finally, they employ the standard L-shaped algorithm in conjunction with these upper bounds to obtain optimal solutions. Erdogan and Denton [8] extend this work to account for no-show patients.

Gupta and Wang [12] develop a Markov Decision Process model for the appointment scheduling problem which incorporates patient choice. They show that for a single server scenario, the optimum policy is a threshold-type, under certain weak assumptions on the probability distribution. For a multiple server case they use several heuristics to solve the problem. They also investigate the effects of patient loyalty to primary care providers, the total clinic load, and the load imbalance among physicians on the clinic's optimal profit.

Hutzschenreuter [13] compares the performance of a selection of appointment-scheduling rules under D+ noise/M/1 models using simulation. Kaandorp and Koole [14] define a local search scheduling algorithm and prove that it converges to the optimal schedule with respect to a weighted sum of average expected waiting times of customers, idle time of server, and tardiness in the schedule. The algorithm is able to incorporate no-shows.

Green and Savin [10] use single server queuing models with both deterministic and exponential service time and patient no-show to capture the dynamics of appointment scheduling. They study two performance metrics, the expected patient backlog and the probability of getting a same-day appointment, and emphasize the usefulness of queuing approximations in providing guidelines for determining daily patient panel size.

Kim and Giachetti [15] also propose clinical overbooking to cope with patient no-show, again with a single no-show probability for all patients. Their model assumes deterministic service time and maximizes expected revenue minus overtime and penalty costs incurred when patients leave without being seen. They do not explicitly consider patient waiting time as part of the objective. Their numerical results indicate that overbooking can significantly improve the revenue as well as provide increased access to patients.

Muthuraman and Lawley [21] develop a sequential scheduling model with exponential service times and multiple patient no-show probabilities. The objective is to maximize the expected revenue for patients seen minus costs for patient waiting and staff overtime. The paper uses the memoryless property of the exponential distribution to show that expected profit evolution is unimodal. That is, the expected profit of a schedule is non-decreasing with the number of patients until some critical number is scheduled and then decreases monotonically as more patients are added. This provides an optimal stopping rule, after which patients are rejected for that particular schedule and can be considered for the schedules of other days.

Chakraborty et al. [3] generalizes the model proposed by Muthuraman and Lawley [21] and shows that the objective unimodality is preserved under general service time assumptions. However, scheduling for general service time requires numerical computation of several large multidimensional integrals. Hence the paper proposes an approximate technique using gamma service time. This technique reduces the computational needs significantly and the quality of the approximation is high when the gamma can be shaped and scaled to provide a good fit.

Zeng et al. [24] study the clinical scheduling problem with overbooking for heterogeneous patients (i.e. patients with the different no-show probabilities). The objective maximizes the expected profit, which includes revenue from patients and costs associated with patient waiting times and physician overtime. They show that the objective function with homogeneous patients is multimodular. However, multimodularity does not hold when patients are heterogeneous. Hence, for heterogeneous patients they propose a local search algorithm to find local optimal schedules. Finally, they extend the results to sequential scheduling and propose two sequential scheduling procedures.

Turkcan et al. [23] use stochastic multiobjective optimization approaches to model the clinical scheduling problem and present several sequential scheduling algorithms to obtain Pareto-optimal solutions. These methods are notable in that waiting times are handled in constraints rather than being costed in the objective. Furthermore, the paper introduces a new fairness measure and investigates its impact on clinic revenue and cost.

Luo et al. [19] develop scheduling methods for healthcare systems subject to interruptions due to high priority patients. Their objective is to balance patient waiting and provider utilization, and their models provide several interesing insights, such as that equally spaced slot-structures perform well when interruption rate is constant but poorly when interruption rate is time dependent.

All the papers reviewed above, as is the case for almost all the papers in literature, assume a slot structure in some form. In this paper, however, we assume that the service period is continuous, i.e. not divided into slots. We assume that the call-in patients are heterogeneous (in terms of their no-show probability) and the service time distribution is exponential with parameter $\lambda$. Note that exponential service times are commonly used in scheduling literature for primary care clinics [10,14,18]. The exponential is notable in that its memoryless property allows closed form results that are often not possible with general service time. This is useful for understanding the effect and possible impact of different scheduling schemes on performance. We do not claim that all clinics have exponential service, although we have seen this in our previous work (see Kopach et al. [16]). We note that the research in this paper can be formulated in a general service time setting (as in Chakraborty et al. [3]). But this requires an accumulation of multi-dimensional integrals, which would detract from our objective of studying the effect of slot structure on schedule performance.

## 3. Problem formulation

In this section, we develop the appointment scheduling model and present the algorithm for scheduling customers based on their call-in sequence. The total operation time of the service facility is predetermined (typically 8–9 h) and is denoted by $T$. Customers are scheduled during the call-in period, which precedes the day for which appointments are given. Any call-in customer is given an appointment in $[0,T]$ or rejected (possibly scheduled for another day). Each customer has a fixed no-show probability which can be determined from historical data. The no-show behavior of each customer is independent of any other. During the call-in session, the scheduler looks for the most suitable time to schedule the customer based on the customer's no-show probability and the state of the existing schedule (partial schedule). It is further assumed that scheduled appointments may not be readjusted later.

Let the ordered set $S^n = \{s_1, s_2, \ldots, s_n\}$ denote the partial schedule after $n$ call-ins, where $s_i$, $i = 1,2,\ldots,n$, denotes the appointment time of the $i$th "scheduled" customer (note $s_i \le s_{i+1}$, $\forall i$). It is important to note that the sequence of call-in customers is different from the sequence of "scheduled" customers, i.e. the $k$th call-in customer may be scheduled at $s_j$, where $k \ne j$. Let $p_i$ denote the probability that the $i$th scheduled customer will arrive for the appointment.

Before we describe the insertion of a new patient into the schedule $S^n$, we will derive the expressions for the expected waiting time and the expected overtime for the schedule $S^n$. Let $X^n(s_i)$ be the random variable denoting the number of customers waiting for service, including the one in service, *just prior* to time $s_i$. Note that $X^n(s_i)$ includes *only* those customers scheduled in $[0,s_i)$, who arrive at their appointed time, and who have not yet completed service by $s_i$. Thus, $X^n(s_i) \in \{0,1,\ldots,i-1\}$, as there are $(i-1)$ scheduled customers in $[0,s_i)$.

Let $Q_k^n(s_i)$ denote the probability that there are $k$ customers in the system, *just prior* to time $s_i$. Then

$$Q_k^n(s_i) = \Pr\{X^n(s_i) = k\}, \quad k = 0,1,\ldots,i-1 \tag{1}$$

Next, let $Y(t)$ be a Poisson random variable, denoting the number of service completions in time $t$, and set $f(n,\lambda t) = \Pr\{Y(t) = n\}$. Also, define $F(n,\lambda t) = \Pr\{Y(t) \ge n\} = \sum_{j=n}^{\infty} f(j,\lambda t)$. Let $\mathcal{A}_i$ denote the arrival of the $i$th scheduled customer. Then for $k > 0$

$$\begin{aligned} Q_k^n(s_i) &= \Pr\{X^n(s_i) = k\} = \sum_{j=k-1}^{i-2} \Pr\{X^n(s_{i-1}) \\ &= j, \mathcal{A}_{i-1}, Y(s_i - s_{i-1}) = j+1-k\} \\ &\quad + \sum_{j=k}^{i-2} \Pr\{X^n(s_{i-1}) = j, \mathcal{A}_{i-1}^c, Y(s_i - s_{i-1}) = j-k\} \end{aligned} \tag{2}$$

The first term in the last equation represents the joint probability that there are $j$ customers waiting at $s_{i-1}$, the $(i-1)^{st}$ customer arrives, and the number of service completions within the interval $[s_{i-1}, s_i)$ is $j+1-k$. Similarly, the second term represents the joint probability that there are $j$ customers waiting at $s_{i-1}$, the $(i-1)^{st}$ customer does not arrive, and the number of service completions within the interval $[s_{i-1}, s_i)$ is $j-k$. Note that $j$ can be at most $(i-2)$, since the total number of customers scheduled in the interval $[0, s_{i-1})$ is $(i-2)$. Conditioning on $\mathcal{A}_{i-1}$ and $X^n(s_{i-1})$ we have

$$\begin{aligned} Q_k^n(s_i) &= \sum_{j=k-1}^{i-2} \Pr\{Y(s_i - s_{i-1}) = j+1-k | X^n(s_{i-1}) \\ &= j, \mathcal{A}_{i-1}\}\Pr\{X^n(s_{i-1}) = j\}p_{i-1} + \sum_{j=k}^{i-2} \Pr\{Y(s_i - s_{i-1}) \\ &= j-k | X^n(s_{i-1}) = j, \mathcal{A}_{i-1}^c\}\Pr\{X^n(s_{i-1}) = j\}q_{i-1} \end{aligned} \tag{3}$$

where $q_i = 1-p_i$. Substituting $Q_j^n(s_{i-1})$ and using independence, we have

$$\begin{aligned} Q_k^n(s_i) &= \sum_{j=k-1}^{i-2} \Pr\{Y(s_i - s_{i-1}) = j+1-k\}Q_j^n(s_{i-1})p_{i-1} \\ &\quad + \sum_{j=k}^{i-2} \Pr\{Y(s_i - s_{i-1}) = j-k\}Q_j^n(s_{i-1})q_{i-1} \\ &= \sum_{j=k-1}^{i-2} Q_j^n(s_{i-1})\Big[p_{i-1}f(j+1-k, \lambda(s_i - s_{i-1})) \\ &\quad + 1_{(j \ge k)} q_{i-1} f(j-k, \lambda(s_i - s_{i-1}))\Big] \\ &\quad \text{for } k = 1,2,\ldots,i-1 \end{aligned} \tag{4}$$

where $1_{(\varepsilon)}$ is the indicator function for and event $\varepsilon$. Similarly for $k = 0$

$$\begin{aligned} Q_0^n(s_i) &= \Pr\{X^n(s_i) = 0\} \\ &= \sum_{j=1}^{i-2} Q_j^n(s_{i-1})[p_{i-1}F(j+1, \lambda(s_i - s_{i-1})) \\ &\quad + q_{i-1}F(j, \lambda(s_i - s_{i-1}))] \\ &\quad + Q_0^n(s_{i-1})[p_{i-1}F(1, \lambda(s_i - s_{i-1})) + q_{i-1}] \end{aligned} \tag{5}$$

Note that $Q_0^n(0) = 1$. Let $X^n(T)$ denote the number of customers waiting for service at the end of the day, i.e. at time $T$. Then $Q_k^n(T)$ denotes the probability that $k$ customers are waiting for service at time $T$. Thus we have

$$Q_k^n(T) = \Pr\{X^n(T) = k\} \quad k = 0,1,\ldots,n \tag{6}$$

The expected waiting time of the $j$th "scheduled" customer is given by

$$\mathrm{E}[W^n(s_j)] = \sum_{k=0}^{j-1} \frac{k}{\lambda} Q_k^n(s_j) = \frac{1}{\lambda}\mathrm{E}[X^n(s_j)] \quad j = 1,2\ldots n \tag{7}$$

Similarly, the expected overtime after scheduling $n$ customers is given by

$$\mathrm{E}[V^n] = \sum_{k=0}^{n} \frac{k}{\lambda} Q_k^n(T) = \frac{1}{\lambda}\mathrm{E}[X^n(T)] \tag{8}$$

Assume that $c_w$ is the waiting cost per unit time for each customer. Overtime cost is incurred if the service facility runs beyond normal operating hours. We use $c_o$ to denote the overtime cost per unit time. Let the function $C(\cdot)$ denote the total expected cost of a schedule. Then, after scheduling $n$ customers, the total expected cost due to customer waiting and staff overtime is given by

$$C(S^n) \;=\; \mathrm{E}\left[\sum_{j=1}^{n} c_w p_j W^n(s_j) + c_o V^n\right] \tag{9}$$

Let us now compute the total expected revenue and the total expected profit for a given schedule. Let the random variable $Z^n$ denote the number of scheduled customers arriving for service after $n$ customers are scheduled. Set

$$A_m^n \;=\; \Pr\{Z^n \;=\; m\} \tag{10}$$

Let $\tilde{p}$ denote the probability that the $n$th customer will show up for the appointment. Then conditioning on $A_m^{n-1}$, we can easily obtain

$$A_m^n = \begin{cases} A_m^{n-1}(1-\tilde{p}) + \left(A_{m-1}^{n-1}\right)\tilde{p} & \text{for } m = 1,2,\dots,n-1 \\ \left(A_{n-1}^{n-1}\right)\tilde{p} & \text{for } m = n \\ A_0^{n-1}(1-\tilde{p}) & \text{for } m = 0. \end{cases} \tag{11}$$

Let $r$ be the revenue per unit time of the server. Then the expected revenue after scheduling $n$ customers is given by

$$R(S^n) \;=\; \frac{r}{\lambda}\mathrm{E}[Z^n] \tag{12}$$

Thus, the total expected profit is given by

$$P(S^n) \;=\; R(S^n) - C(S^n) \tag{13}$$

$$= \frac{r}{\lambda}\mathrm{E}[Z^n] - \mathrm{E}\left[\sum_{j=2}^{n} c_w p_j W^n(s_j) + c_o V^n\right] \tag{14}$$

Next, suppose $n$ customers have been scheduled and the $(n+1)^{st}$ customer calls for an appointment. Define the set $\mathcal{U} = \{u_1,u_2,\dots,u_n\}$, where $u_i \equiv [s_i,s_{i+1})$, $i = 1,2,\dots,n-1$ and $u_n \equiv [s_n,T]$. To determine the best appointment time for the $(n+1)^{st}$ call-in customer, the scheduler assigns the customer sequentially in each of the time intervals, $u_i$, $i = 1,2,\dots,n$, and computes within each interval the appointment time that minimizes the total expected cost for the schedule due to customer waiting time and staff overtime. This step provides a set of $n$ candidate appointment times, each with a corresponding total expected cost. Computing these candidate appointment times requires solving $n$ nonlinear continuous optimization problems.

Consider one such problem, where the $(n+1)^{st}$ call-in customer is assigned at some time $s \in u_{i-1} \equiv [s_{i-1},s_i)$. This problem is described schematically in Fig. 1. The updated schedule vector is given by $S_{i-1}^{n+1} \;=\; \{s_1,\dots,s_{i-1},s,s_i,\dots,s_n\}$. Here we use the subscript $(i-1)$ to indicate that this is an intermediate schedule for the $(i-1)^{st}$ optimization problem (since the customer is being scheduled in the interval $u_{i-1}$).

Now let $s_{i-1}^*$ denote the best appointment time for the problem described in Fig. 1 (i.e. when the $(n+1)^{st}$ call-in customer is assigned at $s \in u_{i-1}$). As mentioned before, the best appointment time is obtained by minimizing the total expected cost function $C(S_{i-1}^{n+1})$ given in Equation (9). Hence,

$$s_{i-1}^* \;=\; \underset{s \in u_{i-1}}{\operatorname{argmin}}\, C\left(S_{i-1}^{n+1}\right) \tag{15}$$

$$C^*\left(S_{i-1}^{n+1}\right) \;=\; \min_{s \in u_{i-1}} C\left(S_{i-1}^{n+1}\right) \tag{16}$$

The best appointment time for the $(n+1)^{st}$ call-in customer will be the one which yields the minimum value of the total expected cost among all the above $n$ assignments. Thus,

$$j^* \;=\; \underset{1 \le j \le n}{\operatorname{argmin}}\, C^*\left(S_j^{n+1}\right) \tag{17}$$

$$s^* \;=\; s_{j^*}^* \tag{18}$$

Thus in this section we have developed the objective function and outlined the scheduling algorithm. Below, we shall describe the scheduling algorithm in detail.

### 3.1. Scheduling algorithm

Given a partial schedule $S^n$, the scheduling algorithm computes an appointment time for the new call-in customer for each assignment of the customer in $u_i \in \mathcal{U}$. As mentioned in the previous section, the best appointment time is obtained by minimizing the total expected cost function given in Equation (9). Among all these *best appointments*, the one which yields the minimum value of the objective is chosen as the appointment time of the current customer. Note here that the algorithm schedules customers in the sequence they call-in. However, it does not consider future arrivals when making the assignment. Thus the algorithm is:

1. Set $s_1 = 0$ and $Q_0(0) = 1$.
2. Wait for the $(n+1)^{st}$ call.
3. For each $u_i \in \mathcal{U}$, $i = 1,2,\dots,n$
   - (a) Assign the $(n+1)^{st}$ call-in customer at $s \in u_i$.
   - (b) Compute $Q_k^{n+1}(s)$ using Equations (4) and (5).
   - (c) Compute $C^*(S_i^{n+1})$ using Equations (9) and (16). Determine $s_i^*$ using Equation (15).
4. Set $i^* = \underset{1 \le i \le n}{\operatorname{argmin}}\, C^*(S_i^{n+1})$ and $C^*(S^{n+1}) = \min_{1 \le i \le n} C^*(S_i^{n+1})$. Set $s^* = s_{i^*}^*$.
5. Compute the expected profit from Equation (14) using $C^*(S^{n+1})$.
6. If $P(S^{n+1}) \ge P(S^n)$, then schedule the call-in customer at $s^*$ and goto Step 2.
7. Else reject the call-in customer and stop.

Next we shall discuss some of the characteristics of the expected cost and the expected profit function followed by some numerical examples.

## 4. Characterization of objective function

In this section, we investigate the characteristics of the objective function and compute closed form expressions for the best appointment times for the first and the second call-in customer. We also establish the convexity of the total expected cost function. Recall that, when a new customer calls-in, the scheduling algorithm

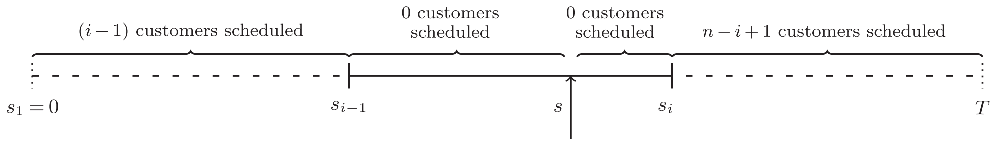

**Fig. 1.** The appointment scheduling model after scheduling $n$ customers.

iteratively assigns the customer to $u_i$, for all $u_i \in \mathcal{U}$, and for each $u_i$ computes the best appointment time within $u_i$ by minimizing the objective function. Here we show that the objective function is convex for each of these assignments. Obviously, this guarantees the existence of the global minimum in each $u_i$ interval.

**Proposition 1**. *The first call-in customer should be scheduled at time* 0.

*Proof*: See Proof of Proposition 1 in the appendix.

**Proposition 2**. *The best appointment time for the second call-in customer is given by*

$$s^* = \max\left(0, \frac{1}{\lambda}\ln\left[\frac{p_1 + \sqrt{p_1^2 + 4p_1\beta e^{\lambda T}}}{2}\right]\right) \tag{19}$$

*where* $\beta = c_w/c_o$.

*Proof*: See Proof of Proposition 2 in the appendix.

Now we shall establish the convexity of the objective function. Let us consider the optimization problem described in Fig. 1. Recall that in this problem, the $(n+1)^{st}$ call-in customer is assigned at $s \in u_{i-1} \equiv [s_{i-1}, s_i)$ and the new partial schedule is given by the vector $S_{i-1}^{n+1} = \{s_1, \ldots, s_{i-1}, s, s_i, \ldots, s_n\}$. We will show that the total expected cost given in Equation (9) is a convex function of s. The proof uses three results which are established in Lemmas 1, 2 and 3. Lemma 1 and 2, respectively, show that the expected waiting time of the "new" $i$th and the "new" $(i+1)^{st}$ "scheduled" customer are a convex functions of s. Next, Lemma 3 shows that the expected waiting time of the "new" $(i+m)^{th}$, $m \geq 2$, "scheduled" customers and the expected overtime are also convex. For clarification see Fig. 2. Proposition 3, 4, 5, and 6 establish the results necessary to prove these lemmas. Below we set $Q_j^{n+1}(\cdot) \equiv Q_j(\cdot)$, and $W^{n+1}(s_j) = W_j$, $\forall j$ for convenience. Also we use $\dot{f}(\cdot) = \mathrm{d}f/\mathrm{d}s$.

**Proposition 3**. *Suppose the* $(n+1)^{st}$ *call-in customer is scheduled as the "new"* $i$*th customer at* $s \in u_{i-1}$. *Then, for* $k = 1,2,\ldots,i-1$ *we have*

$$\dot{Q}_k(s) = -\lambda Q_k(s) + \lambda Q_{k+1}(s) \tag{20}$$

and

$$\dot{Q}_0(s) = \lambda Q_1(s) \tag{21}$$

*Proof*: See Proof of Proposition 3 in the appendix.

**Lemma 1**. *Suppose the* $(n+1)^{st}$ *call-in customer is scheduled as the "new"* $i$*th customer at* $s \in u_{i-1}$. *Then the expected waiting time of the "new"* $i$*th customer,* $\mathrm{E}[W_i]$, *is a non-increasing, convex function of s.*

*Proof*: From Equation (7) we get

$$\mathrm{E}[W_i] = \frac{1}{\lambda}\mathrm{E}[X(s)] \tag{22}$$

$$= \frac{1}{\lambda}\sum_{k=1}^{i-1} kQ_k(s) \tag{23}$$

Differentiating with respect to $s$ we get

$$\begin{aligned}
\dot{\mathrm{E}}[W_i] &= \frac{1}{\lambda}\sum_{k=1}^{i-1} k\dot{Q}_k(s) \\
&= \frac{1}{\lambda}\sum_{k=1}^{i-1} k[-\lambda Q_k(s) + \lambda Q_{k+1}(s)] \quad \text{(from Proposition 3)} \\
&= -\sum_{k=1}^{i-1} kQ_k(s) + \sum_{k=1}^{i-1} kQ_{k+1}(s)
\end{aligned}$$

Note that the maximum queue length for the $i$th scheduled customer is $(i-1)$. Hence $Q_k(s) = 0$, for $k \geq i$. Thus

$$\begin{aligned}
\dot{\mathrm{E}}[W_i] &= -\sum_{k=1}^{i-1} kQ_k(s) + \sum_{k=2}^{i-1} (k-1)Q_k(s) + Q_1(s) - Q_1(s) \\
&= -\sum_{k=1}^{i-1} Q_k(s) \leq 0 \quad (\because Q_k(s) = 0, \forall k \geq i,)
\end{aligned}$$

Thus $\mathrm{E}[W_i]$ is non-increasing. Further differentiating with respect to $s$,

$$\begin{aligned}
\ddot{\mathrm{E}}[W_i] &= -\sum_{k=1}^{i-1} \dot{Q}_k(s) = -\lambda\sum_{k=1}^{i-1}[-Q_k(s) + Q_{k+1}(s)] \\
&= \lambda\sum_{k=1}^{i-1}[Q_k(s) - Q_{k+1}(s)] = \lambda Q_1(s) \geq 0 \quad (\because Q_k(s) = 0, \forall k \geq i)
\end{aligned}$$

Thus establishing the convexity of $\mathrm{E}[W_i]$.

**Proposition 4**. *Suppose the* $(n+1)^{st}$ *call-in customer is scheduled as the "new"* $i$*th customer at* $s \in u_{i-1}$ *and let* $\hat{p}$ *denote the probability that the customer will attend the appointment. Then*

$$\dot{Q}_k(s_i) = 0, \ \text{for } 2 \leq k \leq i \tag{24}$$

*where* $s_i$ *is the scheduled appointment time of the "new"* $(i+1)^{st}$ *customer.*

*Proof*: See Proof of Proposition 4 in the appendix.

The following corollary follows easily from Proposition 4.

**Corollary 1**. *Suppose the* $(n+1)^{st}$ *call-in customer is scheduled as the "new"* $i$*th customer at* $s \in u_{i-1}$. *Then*

$$\dot{Q}_j(s_{i+k-1}) = 0, \ \text{for } j \geq k+1, \ \text{for } k \geq 1 \tag{25}$$

*Proof*: The result can be easily derived by differentiating $Q_j(s_{i+k-1})$ with respect to s and using Proposition 4.

**Proposition 5**. *Suppose the* $(n+1)^{st}$ *call-in customer is scheduled as the "new"* $i$*th customer at* $s \in u_{i-1}$ *and let* $\hat{p}$ *denote the probability that the customer will attend the appointment. Then*

$$\dot{Q}_1(s_i) = Q_0(s)\hat{p}\lambda f(0, \lambda(s_i - s)) \tag{26}$$

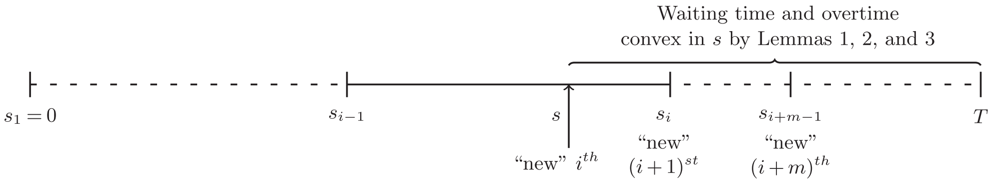

**Fig. 2.** Convexity of expected cost function.

*and*

$$\dot{Q}_0(s_i) = -\dot{Q}_1(s_i) \tag{27}$$

*Proof*: See Proof of Proposition 5 in the appendix.

**Lemma 2**. *Suppose the* $(n+1)^{st}$ *call-in customer is scheduled as the "new"* $i$th *customer at* $s \in u_{i-1}$. *Then the expected waiting time of the "new"* $(i+1)^{st}$ *customer,* $\mathrm{E}[W_{i+1}]$, *is a non-decreasing, convex function of s.*

*Proof*: By definition,

$$\mathrm{E}[W_{i+1}] = \frac{1}{\lambda}\sum_{k=1}^{i} kQ_k(s_i) \tag{28}$$

Differentiating with respect to $s$

$$\begin{aligned}\dot{\mathrm{E}}[W_{i+1}] &= \frac{1}{\lambda}\sum_{k=1}^{i} k\dot{Q}_k(s_i)\\ &= \frac{1}{\lambda}\dot{Q}_1(s_i) \quad (\text{from Proposition 4, Equation (24)})\\ &= Q_0(s)\widehat{p}f(0,\lambda\Delta_i) \geq 0\\ &\quad (\text{from Proposition 5, Equation (26)})\end{aligned}$$

Thus $\mathrm{E}[W_{i+1}]$ is non-decreasing. Differentiating again with respect to $s$

$$\begin{aligned}\ddot{\mathrm{E}}[W_{i+1}] &= \dot{Q}_0(s)\widehat{p}f(0,\lambda\Delta_i) + \lambda Q_0(s)\widehat{p}f(0,\lambda\Delta_i)\\ &= \lambda[Q_1(s) + Q_0(s)]\widehat{p}f(0,\lambda\Delta_i) \geq 0\\ &\quad (\text{from Proposition 3, Equation (21)})\end{aligned}$$

This establishes the convexity of $E[W_{i+1}]$ in $s$.

Up to now we have shown that the expected waiting time of the "new" $i$th and the "new" $(i+1)^{st}$ scheduled customer in the schedule $S_{i-1}^{n+1}$ are convex in s. Next we want to show that the expected waiting time of the $(i+m)^{th}$ scheduled customer, for $m = 2,\ldots,n-i$, and the expected overtime of the schedule $S_{i-1}^{n+1}$ are also convex in s.

**Proposition 6**. *Suppose the* $(n+1)^{st}$ *call-in customer is scheduled as the "new"* $i$th *customer at* $s \in u_{i-1}$. *Then*

$$\sum_{l=k}^{m} \dot{Q}_l(s_{i+m-1}) \geq 0, \;\; \text{for } 1 \leq k \leq m, 2 \leq m \leq n-i \tag{29}$$

*and*

$$\sum_{l=k}^{m} \ddot{Q}_l(s_{i+m-1}) \geq 0, \;\; \text{for } 1 \leq k \leq m, 2 \leq m \leq n-i \tag{30}$$

*Proof*: See Proof of Proposition 6 in the appendix.

**Corollary 2**. *Suppose the* $(n+1)^{st}$ *call-in customer is scheduled as the "new"* $i$th *customer at* $s \in u_{i-1}$. *Then*

$$\sum_{l=k}^{n+1} \dot{Q}_l(T) \geq 0, \;\; \text{for } 1 \leq k \leq n+1 \tag{31}$$

*and*

$$\sum_{l=k}^{n+1} \ddot{Q}_l(T) \geq 0, \;\; \text{for } 1 \leq k \leq n+1 \tag{32}$$

*Proof*: The result directly follows from Proposition 6.

**Lemma 3**. *Suppose the* $(n+1)^{st}$ *call-in customer is scheduled as the "new"* $i$th *customer at* $s \in u_{i-1}$. *Then*

1. *The expected waiting time of the "new"* $(i+m)^{th}$ *customer, i.e the* $(i+m-1)^{st}$ *customer in the schedule* $S^n$, $\mathrm{E}[W_{i+m}]$, *is a non-decreasing, convex function of s, for* $2 \leq m \leq n-i$.
2. *The expected overtime of the schedule* $S_{i-1}^{n+1}$, *i.e.* $\mathrm{E}[V]$, *is a non-decreasing, convex function of s.*

*Proof*: Recall that the $(i+m)^{th}$ scheduled customer in the schedule $S_{i-1}^{n+1}$ is scheduled at $s_{i+m-1}$. Thus from Equation (7) we have

$$\begin{aligned}\mathrm{E}[W_{i+m}] &= \frac{1}{\lambda}[X(s_{i+m-1})] = \frac{1}{\lambda}\sum_{k=1}^{i+m-1} kQ_k(s_{i+m-1})\\ &= \frac{1}{\lambda}\sum_{k=1}^{i+m-1}\sum_{l=k}^{i+m-1} Q_k(s_{i+m-1}) \quad (\text{rewriting the sum})\end{aligned} \tag{33}$$

Thus first part of the result follows easily by differentiating Equation (33) and applying Proposition 6, Equation (30) and Corollary 1.

Similarly, the second can also be derived by applying Corollary 2, Equation (32) and Corollary 1.

**Theorem 1**. *Suppose the* $(n+1)^{st}$ *call-in customer is scheduled as the "new"* $i$th *customer at* $s \in u_{i-1}$. *Then the total expected cost of the schedule* $S_{i-1}^{n+1}$ *is a convex function of s.*

*Proof: From* Equation (9) *we have*

$$C\left(S_{i-1}^{n+1}\right) = \mathrm{E}\left[\sum_{k=1}^{n+1} c_w p_k W_k + c_o V\right] \tag{34}$$

Thus the result easily follows by differentiating Equation (34) with respect to s, and using Lemma 1, 2 and 3.

The next result shows that the total expected profit given in Equation (12) is unimodal. Thus, if the best assignment for the current call-in customer results in an objective decrease, then all subsequent assignments will lead to additional decreases in the objective. As mentioned before, this provides a unique stopping criterion for the algorithm. This result can be easily proved by following an argument similar to the one given in Theorem 1 in Chakraborty et. al. [3]. Below we provide an outline of the proof.

**Theorem 2**. *The total expected profit is unimodal. Thus if* $P(S^{n+1}) < P(S^n)$, *for some n, then* $P(S^{n+2}) \leq P(S^{n+1})$.

*Outline of the proof*: Let $\mathcal{A}_{n+1}$ denote the event that the $(n+1)^{st}$ call-in customer arrives for the appointment. Also suppose, $\widehat{p}$ is the probability that the $(n+1)^{st}$ call-in customer will arrive. Thus

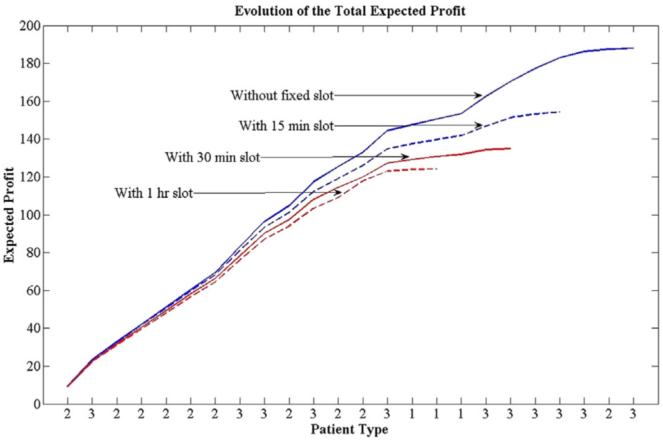

**Fig. 3.** Evolution of the expected profit for the proposed algorithm and for slot models with slot lengths 1 h, 30 min and 15 min.

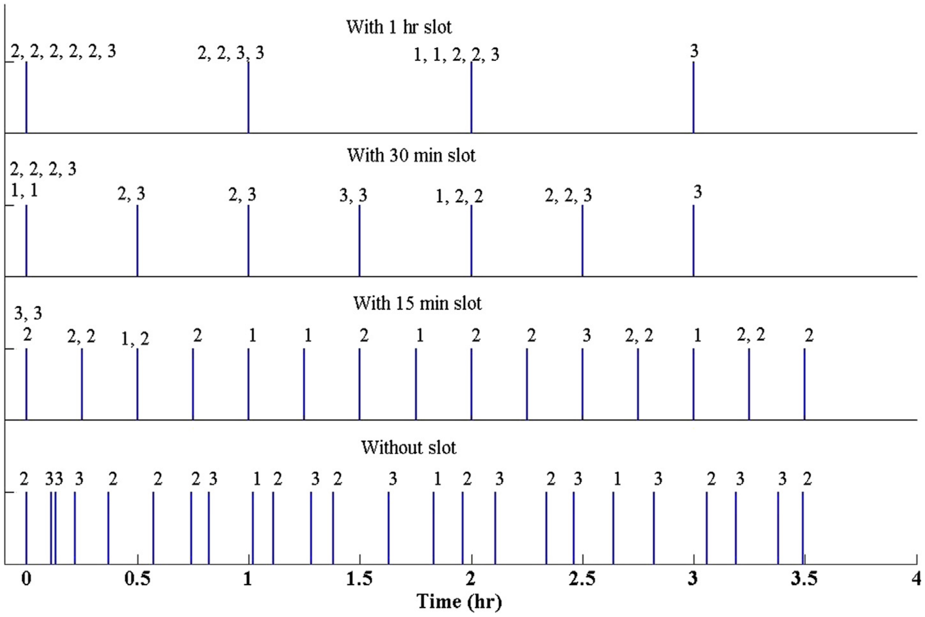

**Fig. 4.** Gantt chart for the proposed algorithm and for slot models with slot lengths 1 h, 30 min and 15 min.

$$P\left(S^{n+1}\right) < P(S^n) \Rightarrow R\left(S^{n+1}\right) - C\left(S^{n+1}\right) < R(S^n) - C(S^n) \Rightarrow \frac{r}{\lambda}\mathrm{E}\left[Z^{n+1} - Z^n\right] < C\left(S^{n+1}\right) - C(S^n) \Rightarrow \frac{r}{\lambda} < \left[C\left(S^{n+1}\middle|\mathcal{A}_{n+1}\right) - C(S^n)\right] \tag{35}$$

Now to prove the result, we have to show that

$$C\left(S^{n+2}\middle|\mathcal{A}_{n+2}\right) \; - C\left(S^{n+1}\right) > \frac{r}{\lambda} \tag{36}$$

given that

$$C\left(S^{n+1}\middle|\mathcal{A}_{n+1}\right) \; - C(S^n) > \frac{r}{\lambda} \tag{37}$$

Now for the schedules $S^{n+2}$ and $S^{n+1}$ it is clear that

$$\left[C\left(S^{n+2}\middle|\mathcal{A}_{n+2}, \mathcal{A}_{n+1}\right) - C\left(S^{n+1}\middle|\mathcal{A}_{n+1}\right)\right] \geq C\left(S^{n+2}\middle|\mathcal{A}_{n+2}\right) - C\left(S^{n+1}\right) \tag{38}$$

Hence

$$\left[C\left(S^{n+2}\middle|\mathcal{A}_{n+2}\right) - C\left(S^{n+1}\right)\right] \geq \left[C\left(S^{n+2}\middle|\mathcal{A}^c_{n+1}, \mathcal{A}_{n+2}\right) - C(S^n)\right] \tag{39}$$

Now suppose the $(n+1)^{st}$ call-in customer is scheduled at $\hat{s}$ and the $(n+2)^{nd}$ call-in customer is scheduled at $\hat{\hat{s}}$. Then $C(S^{n+2}|\mathcal{A}^c_{n+1}, \mathcal{A}_{n+2})$ is equal to the expected cost of a schedule obtained from $S^n$, by assigning the $(n+1)^{st}$ call-in customer at $\hat{\hat{s}}$ instead of $\hat{s}$, conditioned on the arrival of the customer. Hence from our scheduling algorithm we have

$$C(S^n) < C\left(S^{n+1}\middle|\mathcal{A}_{n+1}\right) \leq C\left(S^{n+2}\middle|\mathcal{A}^c_{n+1}, \mathcal{A}_{n+2}\right) \tag{40}$$

Hence, combining Equations (39) and (40) with Equation (37) we get the result.

## 5. Computational results

The purpose of this section is to gain insights into the behavior of the proposed scheduling algorithm. For this purpose, we first compare the expected profit obtained from the algorithm proposed in this paper with the one obtained from Muthuraman and Lawley [21]. Next, we evaluate the performance of our scheduling methodology by analyzing the expected customer waiting time and staff overtime. We also investigate how the performance changes with change in arrival rates of calling customers and with different cost coefficients for waiting time and overtime. Finally, we examine how a fixed slot model, which yields performance similar to the proposed method can be established.

In the experiments, we take the service time distribution as exponential with mean 15 min. We assume that there are three types of customers, and their show-up probabilities are $p = (0.45, 0.55, 0.65)$. Total service period is 4 h. The reward is set to $r$ = \$75 per hour and the overtime cost is $c_o$ = \$150 per hour. Customer waiting time costs are $c_w$ = \$0.33 per minute.

**Table 1**
Percentage improvement statistics for the proposed algorithm over the slot models.

| | Percentage improvement in maximum expected profit | | Percentage improvement in number of patients scheduled | |
|---|---|---|---|---|
| | Average | Standard deviation | Average | Standard deviation |
| With 1 h slot | 37.3 | 2.8 | 31.2 | 3.2 |
| With 30 min slot | 26.4 | 3.1 | 20.9 | 2.9 |
| With 15 min slot | 17.4 | 2.3 | 13.2 | 3.8 |

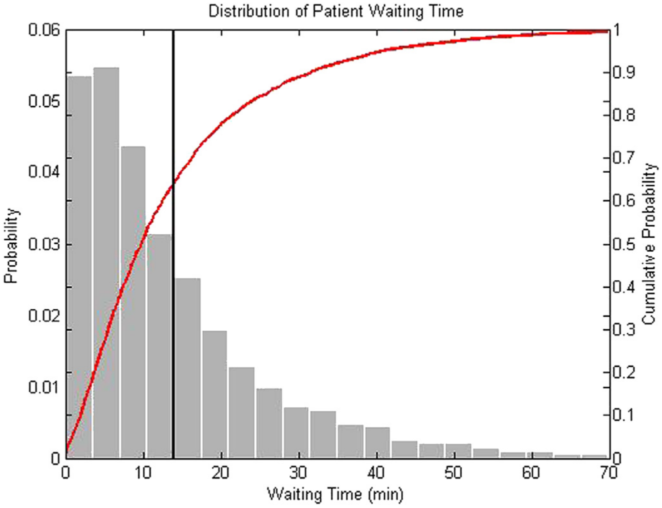

**Fig. 5.** Distribution of customer waiting time.

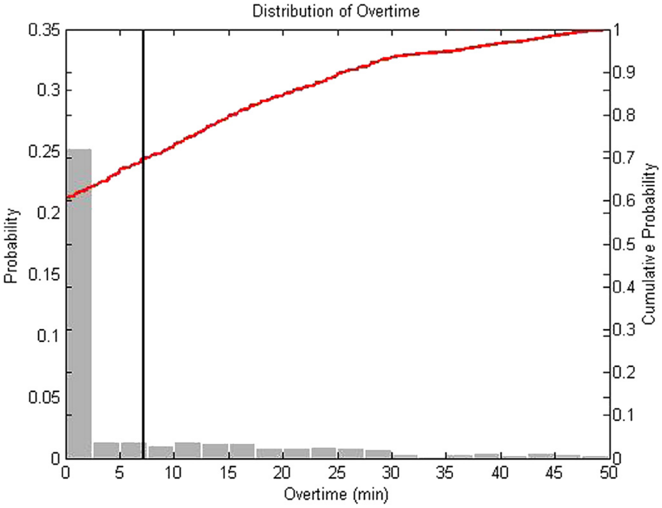

**Fig. 6.** Distribution of overtime.

### 5.1. *Comparison with slot model*

In this section, we compare the expected profit obtained from the proposed algorithm with the case where the service period is divided into slots (see Muthuraman and Lawley [21]). For the latter, we assume that the total hours of operation of the service facility is divided into slots of length 1 h, 30 min, and 15 min. From our experience in working with our clinical partners we have observed that the standard slot length is often close to 15 min. Also, the heuristic algorithms proposed in the literature suggest that the slot length should be approximately equal to the mean of the service time distribution (see Cayirli et al. [2]). However, in this experiment we still consider slots longer than 15 min to investigate the effects of slot length. Figs. 3 and 4 respectively show one example of the evolution of the expected profit and the Gantt chart for the above three slot models and the proposed algorithm.

Fig. 3 illustrates that the algorithm proposed in this paper schedules more customers and also has higher expected profit. Furthermore, the number of patients scheduled and the total expected profit increases with decrease in slot length. We performed a simulation experiment with 500 call-in sequences for each of the above scenarios. Table 1 summarizes the statistics on performance improvement obtained from the proposed algorithm over the slot model. In the above experiments percentage improvement for number of patients scheduled is defined as

$$\frac{\text{no. of scheduled by proposed algorithm} - \text{no. of scheduled by slot model}}{\text{no. of scheduled by proposed algorithm}} * 100$$

Percentage improvement for maximum expected profit is defined similarly.

### 5.2. *Distribution of waiting time and overtime*

Below we present the results of a simulation experiment conducted to obtain the distribution of waiting time and overtime. This experiment uses schedules generated from 500 random call-in sequences. Each schedule was simulated 10 times. Fig. 5 shows both the cumulative distribution and the probability distribution of customer waiting time. From this figure it is clear that on any day, the probability of having 30 min or more waiting time is only 10%. The average daily waiting time is 14 min and is represented by the solid vertical line in the figure.

Similarly, Fig. 6 shows the cumulative distribution and the probability distribution of staff overtime. From this figure we can say that on less that 10% of days, overtime goes above 20 min. Also the average overtime per day is only 8 min. Next, we compute physician's utilization on each of the above 5000 days. Utilization ($\rho$) for a particular day is defined as, $\rho = 1-$(total idle time/total service hours). Fig. 7 shows the distribution of physician's utilization. The average utilization over 5000 days is 84% and is indicated by the solid vertical line in the figure. Fig. 8 shows the distribution of the number of customers served over these 5000 days.

### 5.3. *Sensitivity to customer mix*

In this subsection we investigate the effect of customer mix (in terms of no-show probability) on different performance measures such as maximum expected profit and number of customers scheduled. For this purpose, we consider three scenarios, with three different call-in rates for the customers with show-up probability $p = (0.45,0.55,0.65)$. In the first case we assume that the call-in rate for customers in each of the three groups are respectively (1/6,2/6,3/6). In the other two scenarios, the call-in rates were modified to respectively (1/3,1/3,1/3) and to (3/6,2/6,1/6). In Table 2 we summarize the results for 500 simulation runs.

From Table 2 we see that the average maximum profit decreases with increasing "no-show" probabilities. Also the total number customers scheduled increases with increasing no-show probabilities, as expected.

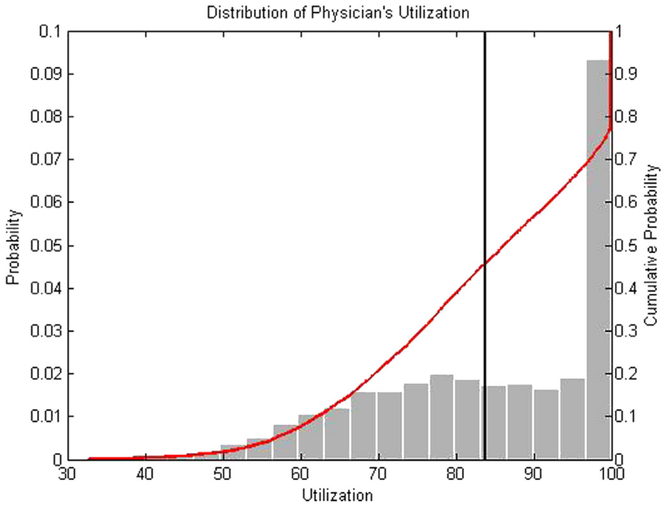

**Fig. 7.** Distribution of physician's utilization.

### 5.4. Sensitivity analysis for cost coefficients

It is obvious that decreasing cost (either waiting cost or overtime cost) will increase both the maximum expected profit and the total number of customers scheduled. But a lower waiting cost may result in higher customer waiting time, since more customers are scheduled. In this subsection we investigate how sensitive these performance measures are to the cost coefficients. This will provide managerial insight into determining suitable cost parameters to be used with the scheduling model.

As before, we assume that there are three types of customers with show-up probability $p = (0.45,0.55,0.65)$ and each type of customer has equal probability to call-in. Then we generate 1000 call-in sequences and use our policy to schedule the customers for various cost coefficients. Figs. 9 and 10 show the effect of waiting time cost and overtime cost on the maximum expected profit and the number of customers scheduled.

From Fig. 9 we see that, when the expected waiting time cost is doubled (from \$0.33/min to \$0.66/min), the average expected profit falls by 38% (from \$170 to \$105), whereas the average number of patients scheduled falls by 36% (from 22 to 14). Fig. 10 indicates that when the overtime cost is increased by 27% (from \$150 to \$190), the average expected profit falls by 4.6% (from \$170 to \$162.2), whereas the average number of patients scheduled falls by 8.2% (from 22 to 20.16). Hence, Figs. 9 and 10 indicate that both the performance measures have higher sensitivity to waiting time cost ($c_w$) than to overtime cost ($c_o$).

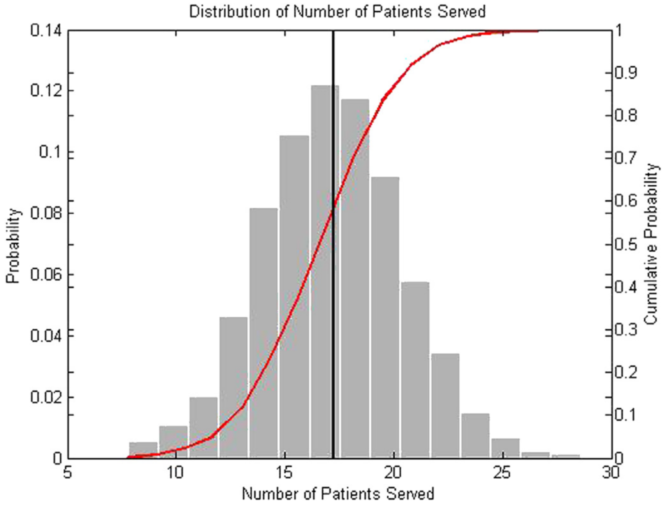

**Fig. 8.** Distribution of the number of customers served.

**Table 2**
Comparison of performance measures for different customer mix.

| | Scenario 1 arrival rate $\left(\frac{1}{6},\frac{2}{6},\frac{3}{6}\right)$ | Scenario 2 arrival rate $\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$ | Scenario 3 arrival rate $\left(\frac{3}{6},\frac{2}{6},\frac{1}{6}\right)$ |
|---|---|---|---|
| Avg. maximum profit (\$) | 183.24 | 164.64 | 155.05 |
| Avg. number of customers scheduled | 21.66 | 25.17 | 29.75 |

### 5.5. Analysis for slot structure

In many service facilities, particularly in healthcare, appointments are given at fixed intervals or "slots" between scheduled arrivals. But as mentioned earlier, the optimal slot size is very difficult to compute under sequential scheduling setup. In this section we investigate this problem and try to gain insight into designing the slot structure.

Recall that the ordered set $S^n = \{s_1,\ldots,s_n\}$ denotes the schedule obtained by minimizing the total expected cost after $n$ call-ins, according to the algorithm described in section 3. Let $\Delta_i = s_{i+1}-s_i$. We shall refer to $\Delta_i$ as the $i$th *"appointment interval"*. Also let $\overline{\Delta} = 1/(n-1)\sum_{i=1}^{n-1}\Delta_i$ denote the average appointment interval. Note that $\overline{\Delta}$ will vary with call-in sequence. Now suppose for a given call-in sequence $P1$, we obtain $\overline{\Delta}$ using our algorithm. Then we create a slot structure by dividing the entire time horizon $[0,T]$, into fixed intervals of length $\overline{\Delta}$. Next, we take the call-in sequence $P1$ and schedule patients sequentially using the algorithm described in [3]. We repeated this experiment for 100 different call-in sequences for each of the three arrival rates (1/3,1/3,1/3), (1/6,2/6,3/6) and (3/6,2/6,1/6) for patients with show-up probability $p = (0.45,0.55,0.65)$. The maximum difference in expected profit obtained from these two algorithms in the above 300 experiments was not more than 3.3%.

Figs. 11 and 12 show two examples of expected profit evolution for two different call-in sequences obtained by using arrival rate

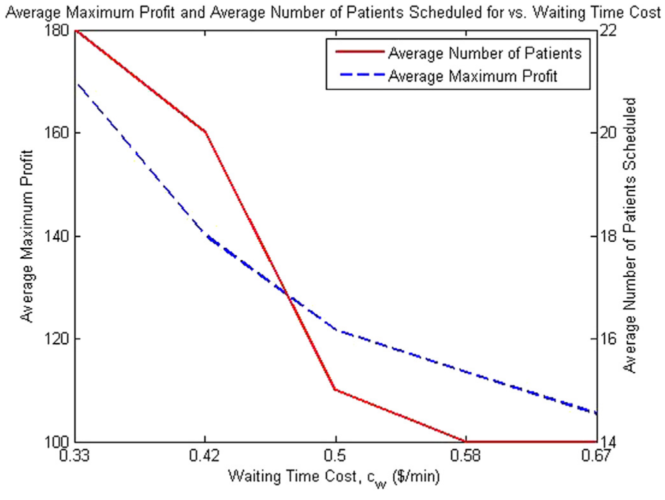

**Fig. 9.** Average maximum profit and average number of customers scheduled for various $c_w$, $r$ = \$75/hr, and $c_o$ = \$150/hr.

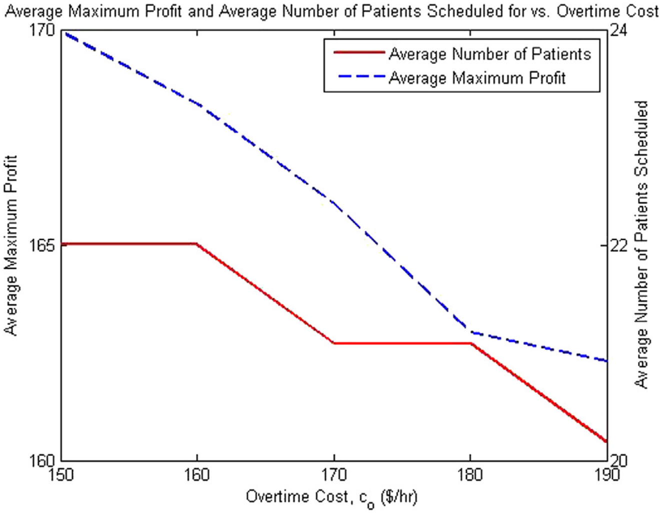

**Fig. 10.** Average maximum profit and average number of customers scheduled for various $c_o$, $r$ = \$75/hr and $c_o$ = \$0.33/min.

(1/3,1/3,1/3) and (1/6,2/6,3/6), respectively. The slot lengths for the above two cases are respectively 8.2 min and 9.5 min.

### 5.5.1. *Effect of patient mix on mean appointment interval*

It is obvious that the mean *"appointment interval"* used to create the slot structure depends on the composition (in terms of no-show probability) of the call-in patient sequence. In this subsection, we try to determine the distribution of mean appointment interval with various patient mixes. As before, we assume that the show-up probability of the call-in patients are given by $p = (0.45,0.55,0.65)$. Then, we reuse the data obtained from the experiments described in section 5.3 to compute the distribution of average appointment interval for three different call-in rates (1/3,1/3,1/3), (1/6,2/6,3/6) and (3/6,2/6,1/6). Figs. 13 and 14 show the cumulative distribution and the probability distribution of mean appointment interval for arrival rate (1/6,2/6,3/6) and (3/6,2/6,1/6), respectively. It is clear from Figs. 13 and 14 that the mean appointment interval is shorter if there are more patients with higher no-show probability within the population. This implies that in clinics with significant no-show

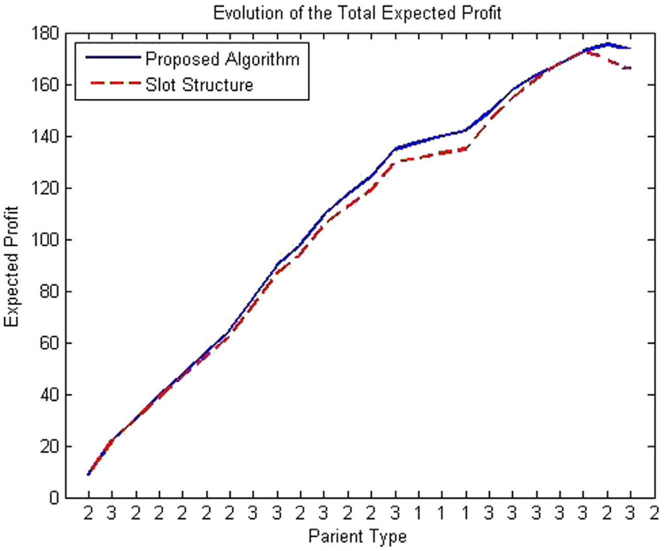

**Fig. 11.** Evolution of expected profit for arrival rate (1/3,1/3,1/3), and slot length = 8.2 min.

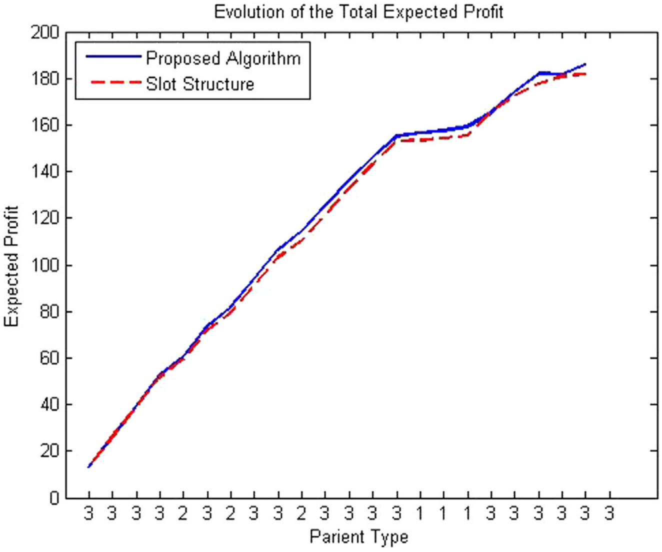

**Fig. 12.** Evolution of expected profit for arrival rate (1/6,2/6,3/6), and slot length = 9.5 min.

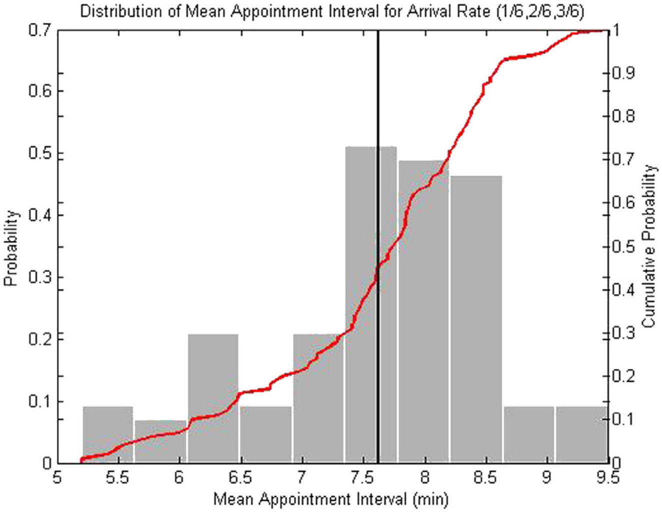

**Fig. 13.** Distribution of mean appointment interval for arrival rate (1/6,2/6,3/6).

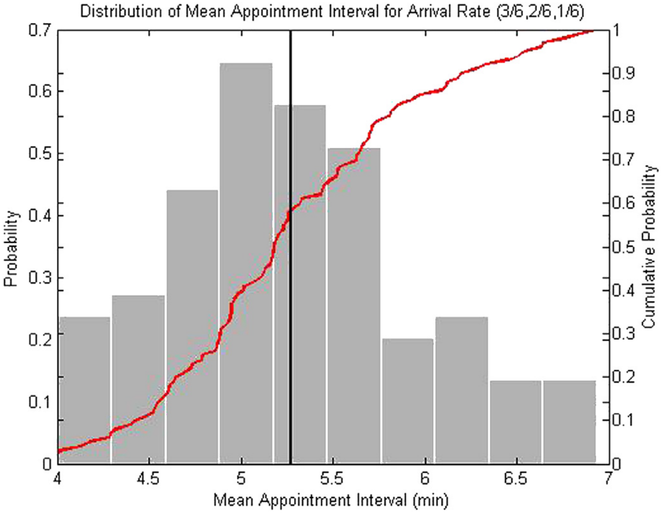

**Fig. 14.** Distribution of mean appointment interval (3/6,2/6,1/6).

problems, the slot length should be significantly shorter than the average service time of the arriving patients.

## 6. Conclusion

In this paper, we developed a model for sequential scheduling under patient no-show, when the service period is not divided in to slots. The schedules generated by the proposed algorithm depend on the state of the existing schedule and the patient' call-in sequence. The best appointment time for a call-in patient is determined by minimizing the expected cost due to customer waiting time and staff overtime. We prove that the objective function for this minimization problem is a convex function of appointment time. This guarantees the existence of global minimum in each optimization problem. Furthermore, we establish that the expected profit is a unimodal function, which provides a unique stopping rule for the scheduling algorithm. To study the behavior of the proposed algorithm, we investigate different performance measures such as expected waiting time and staff overtime. We also compare the performance of the proposed algorithm with the one suggested by Muthuraman and Lawley [21], where the service period is divided into a fixed number of slots. We show that the method proposed in this work yields higher expected profit and less overtime. We also examine the performance of the algorithm for changes in patient mix, and changes in cost coefficients. The work illustrates that slot structures need to be carefully designed to account for both variation in service time and patient no-show behavior. Further, that shorter slot length leads to better overall performance since shorter slot length implies more slots which implies greater scheduling flexibility for the clinic. When coupled with appropriate scheduling algorithms, we believe this additional scheduling flexibility can lead to greater patient access, less patient wait time, and better workload balance for clinic staff.

## Appendix

**Proof of Proposition 1**. Recall that according to Proposition 1, the first call-in customer should be scheduled at time 0.

Suppose the first "call-in" customer is scheduled at $s\in[0,T]$. Since the system is empty before the first customer is scheduled, there is only overtime. For notational simplicity, set $Q_j^1(\cdot)\equiv Q_j(\cdot)$, $\forall j$. Thus, $Q_0(s)=1$ and, from Equation (4)

$$\begin{aligned} Q_1(T) &= p_1 f(0,\lambda(T-s)) \\ &= p_1 e^{-\lambda(T-s)} \end{aligned} \tag{41}$$

Then using Equation (9), the expected cost is given by

$$C\left(S^1\right) = \frac{1}{\lambda}c_o \mathrm{E}[Y(T)] = \frac{1}{\lambda}c_o p_1 e^{-\lambda(T-s)}$$

which is minimized at $s=0$.

**Proof of Proposition 2**. Recall that according to Proposition 2, the best appointment time for the second call-in customer is given by

$$s^* = \frac{1}{\lambda}\ln\left[\frac{p_1+\sqrt{p_1^2+4p_1\beta e^{\lambda T}}}{2}\right] \tag{42}$$

where $\beta = c_w/c_o$.

Suppose the second call-in customer is scheduled at $s\in(s_1,T]$, and let $\tilde{p}$ denote the probability that the customer will attend. For notational simplicity, set $Q_j^2(\cdot)\equiv Q_j(\cdot)$, $\forall j$. Note that the maximum queue length for the second call-in customer can be one. Hence, for the waiting time of the second customer, we get from Equation (4)

$$\begin{aligned} Q_1(s) &= \Pr\{X(s)=1\} \\ &= p_1 f(0,\lambda s) \\ &= p_1 e^{-\lambda s} \end{aligned} \tag{43}$$

Similarly, from Equation (5) we get

$$\begin{aligned} Q_0(s) &= \Pr\{X(s)=0\} \\ &= p_1 F(1,\lambda s)+q_1 \\ &= p_1\left(1-e^{-\lambda s}\right)+q_1 \\ &= 1-p_1 e^{-\lambda s} \end{aligned} \tag{44}$$

Next, for overtime, we get from Equation (4)

$$\begin{aligned} Q_1(T) &= \Pr\{Y(T)=1\} \\ &= Q_0(s)\tilde{p}f(0,\lambda(T-s))+Q_1(s)[\tilde{p}f(1,\lambda(T-s)) \\ &\quad +\tilde{q}f(0,\lambda(T-s))] \\ &= \left[1-p_1e^{-\lambda s}\right]\tilde{p}e^{-\lambda(T-s)}+p_1e^{-\lambda s}\left[\tilde{p}\lambda(T-s)e^{-\lambda(T-s)}\right. \\ &\quad \left.+\tilde{q}e^{-\lambda(T-s)}\right] \\ &= \tilde{p}e^{-\lambda(T-s)}-p_1\tilde{p}e^{-\lambda T}+p_1\tilde{p}(T-s)\lambda e^{-\lambda T}+p_1\tilde{q}e^{-\lambda T} \\ &= \left[\tilde{p}e^{\lambda s}-p_1\tilde{p}s\lambda\right]e^{-\lambda T}+[p_1\tilde{q}+p_1\tilde{p}\lambda T-p_1\tilde{p}]e^{-\lambda T} \end{aligned} \tag{45}$$

Similarly

$$\begin{aligned} Q_2(T) &= \Pr\{Y(T)=2\} \\ &= Q_1(s)\tilde{p}f(0,\lambda(T-s)) \\ &= p_1e^{-\lambda s}\tilde{p}e^{-\lambda(T-s)} \\ &= p_1\tilde{p}e^{-\lambda T} \end{aligned} \tag{46}$$

Now, the total expected cost is given by (see Equations (7)–(9))

$$\begin{aligned} C\left(S_1^2\right) &= \mathrm{E}\left[\sum_{j=1}^{2}c_w p_j W_j^2 + c_o V^2\right] \\ &= \frac{1}{\lambda}\sum_{j=1}^{2}p_j c_w \mathrm{E}\left[X^j(s_j)\right] + \frac{1}{\lambda}c_o\mathrm{E}[Y(T)] \\ &= \frac{1}{\lambda}c_w\tilde{p}p_1e^{-\lambda s}+\frac{1}{\lambda}c_o\left[\left(\tilde{p}e^{\lambda s}-p_1\tilde{p}s\lambda\right)e^{-\lambda T}\right. \\ &\quad \left.+[p_1\tilde{q}+p_1\tilde{p}\lambda T-p_1\tilde{p}]e^{-\lambda T}+2p_1\tilde{p}e^{-\lambda T}\right] \end{aligned} \tag{47}$$

Differentiating w.r.t $s$

$$\frac{\mathrm{d}C\left(S_1^2\right)}{\mathrm{d}s} = -c_w p_1\tilde{p}e^{-\lambda s}+c_o\left(\tilde{p}e^{\lambda s}-p_1\tilde{p}\lambda\right)e^{-\lambda T} \tag{48}$$

Setting $\mathrm{d}C(S^2)/\mathrm{d}s = 0$ and $\beta = c_w/c_o$

$$\begin{aligned}
0 &= -c_w p_1 \tilde{p} e^{-\lambda s} + c_o\left(\tilde{p}e^{\lambda s} - p_1\tilde{p}\right)\, e^{-\lambda T}\\
0 &= e^{2\lambda s} - p_1 e^{\lambda s} - \beta p_1 e^{\lambda T}\\
e^{\lambda s} &= \frac{p_1 \pm \sqrt{p_1^2 + 4p_1\beta e^{\lambda T}}}{2}\\
s^* &= \frac{1}{\lambda}\ln\left[\frac{p_1 + \sqrt{p_1^2 + 4p_1\beta e^{\lambda T}}}{2}\right]
\end{aligned}$$

**Proof of Proposition 3**. Recall that according to Proposition 3, if the $(n+1)^{st}$ call-in customer is scheduled as the "new" $i$th customer at $s \in u_{i-1}$, then, for $k = 1,2,\ldots,i-1$ we have

$$\dot{Q}_k(s) = -\lambda Q_k(s) + \lambda Q_{k+1}(s)$$

and

$$\dot{Q}_0(s) = \lambda Q_1(s)$$

Let's denote $\Delta_{i-1} = s - s_{i-1}$ for notational simplicity. Now for $k = 1,2,\ldots,i-1$ we get from Equation (4)

$$\begin{aligned}
Q_k(s) &= \sum_{j=k-1}^{i-2} Q_j(s_{i-1})\Big[p_{i-1}f(j+1-k, \lambda\Delta_{i-1})\\
&\quad + 1_{j\geq k}q_{i-1}f(j-k, \lambda\Delta_{i-1})\Big]\\
&= \sum_{j=k}^{i-2} Q_j(s_{i-1})[p_{i-1}f(j+1-k, \lambda\Delta_{i-1})\\
&\quad + q_{i-1}f(j-k, \lambda\Delta_{i-1})] + Q_{k-1}(s_{i-1})p_{i-1}f(0, \lambda\Delta_{i-1})
\end{aligned}$$

Differentiating with respect to $s$ we get

$$\begin{aligned}
\dot{Q}_k(s) &= -\lambda\Bigg[\sum_{j=k}^{i-2} Q_j(s_{i-1})[p_{i-1}f(j+1-k, \lambda\Delta_{i-1})\\
&\quad + q_{i-1}f(j-k, \lambda\Delta_{i-1})] + Q_{k-1}(s_{i-1})p_{i-1}f(0, \lambda\Delta_{i-1})\Bigg]\\
&\quad + \lambda\Bigg[\sum_{j=k+1}^{i-2} Q_j(s_{i-1})[p_{i-1}f(j-k, \lambda\Delta_{i-1})\\
&\quad + q_{i-1}f(j-k-1, \lambda\Delta_{i-1})] + Q_k(s_{i-1})p_{i-1}f(0, \lambda\Delta_{i-1})\Bigg]\\
&= -\lambda Q_k(s) + \lambda Q_{k+1}(s) \quad \text{(from Equation (4))} \qquad (49)
\end{aligned}$$

which establishes Equation (20). Now we know, for $s \in u_{i-1}$

$$\sum_{k=0}^{i-1} Q_k(s) = 1 \qquad (50)$$

Differentiating both sides with respect to $s$, and using Equation (49) we get

$$\begin{aligned}
\sum_{k=0}^{i-1}\dot{Q}_k(s) &= 0\\
Q_0(s) &= -\sum_{k=1}^{i-1}\dot{Q}_k(s)\\
&= \lambda\sum_{k=1}^{i-1}[Q_k(s) - Q_{k+1}(s)]\\
&= \lambda Q_1(s)
\end{aligned}$$

which establishes Equation (21).

**Proof of Proposition 4**. Recall that according to Propostiton 4, if the $(n+1)^{st}$ call-in customer is scheduled as the "new" $i$th customer at $s \in u_{i-1}$ and if $\widehat{p}$ denotes the probability that the customer will attend the appointment, then

$$\dot{Q}_k(s_i) = 0, \text{ for } 2 \leq k \leq i$$

Let's denote $\Delta_i = s_i - s$ and $\widehat{q} = 1 - \widehat{p}$. Then for $k = 2,3,\ldots,i$ we have from Equation (4)

$$\begin{aligned}
Q_k(s_i) &= \sum_{j=k}^{i-1} Q_j(s)\Big[\widehat{p}f(j+1-k, \lambda\Delta_i) + \widehat{q}f(j-k, \lambda\Delta_i)\Big]\\
&\quad + Q_{k-1}(s)\widehat{p}f(0, \lambda\Delta_i)
\end{aligned}$$

Differentiating with respect to $s$, and using Proposition 3 we get

$$\begin{aligned}
\dot{Q}_k(s_i) &= \lambda\sum_{j=k}^{i-1}[-Q_j(s) + Q_{j+1}(s)]\Big[\widehat{p}f(j+1-k, \lambda\Delta_i)\\
&\quad + \widehat{q}f(j-k, \lambda\Delta_i)\Big] + \lambda[-Q_{k-1}(s) + Q_k(s)]\widehat{p}f(0, \lambda\Delta_i)\\
&\quad + \lambda\Bigg[\sum_{j=k+1}^{i-1} Q_j(s)\widehat{p}[f(j+1-k, \lambda\Delta_i) - f(j-k, \lambda\Delta_i)]\Bigg]\\
&\quad + \lambda\Bigg[\sum_{j=k+1}^{i-1} Q_j(s)\widehat{q}[f(j-k, \lambda\Delta_i) - f(j-k-1, \lambda\Delta_i)]\Bigg]\\
&\quad + Q_k(s)\widehat{p}\lambda[f(1, \lambda\Delta_i) - f(0, \lambda\Delta_i)] + \lambda Q_k\widehat{q}f(0, \lambda\Delta_i)\\
&\quad + \lambda Q_{k-1}\widehat{p}f(0, \lambda\Delta_i) = \lambda\sum_{j=k}^{i-1}[-Q_j(s)\\
&\quad + Q_{j+1}(s)]\Big[\widehat{p}f(j+1-k, \lambda\Delta_i) + \widehat{q}f(j-k, \lambda\Delta_i)\Big]\\
&\quad + \lambda[-Q_{k-1}(s) + Q_k(s)]\widehat{p}f(0, \lambda\Delta_i)\\
&\quad + \lambda\Bigg[\sum_{j=k}^{i-1} Q_j(s)\Big[\widehat{p}f(j+1-k, \lambda\Delta_i) + \widehat{q}f(j-k, \lambda\Delta_i)\Big]\\
&\quad + Q_{k-1}(s)\widehat{p}f(0, \lambda\Delta_i)\Bigg] - \lambda\Bigg[\sum_{j=k+1}^{i-1} Q_j(s)\Big[\widehat{p}f(j-k, \lambda\Delta_i)\\
&\quad + \widehat{q}f(j-k-1, \lambda\Delta_i)\Big] + Q_k(s)\widehat{p}f(0, \lambda\Delta_i)\Bigg]
\end{aligned}$$

Now using Equation (4) we get

$$\dot{Q}_k(s_i) = \lambda[-Q_k(s_i) + Q_{k+1}(s_i) + Q_k(s_i) - Q_{k+1}(s_i)] = 0$$

**Proof of Proposition 5**. Recall that according to Proposition 5, if the $(n+1)^{st}$ call-in customer is scheduled as the "new" $i$th customer at $s\in u_{i-1}$ and if $\widehat{p}$ denotes the probability that the customer will attend the appointment, then

$$\dot{Q}_1(s_i) = Q_0(s)\widehat{p}\lambda f(0,\lambda(s_i-s))$$

and

$$\dot{Q}_0(s_i) = -\dot{Q}_1(s_i)$$

Letting $\Delta_i = s_i{-}s$ and $\widehat{q} = 1-\widehat{p}$, we get from Equation (4) for $k=1$

$$Q_1(s_i) = \sum_{j=1}^{i-1} Q_j(s)\left[\widehat{p}f(j,\lambda\Delta_i)+\widehat{q}f(j-1,\lambda\Delta_i)\right] + Q_0(s)\widehat{p}f(0,\lambda\Delta_i)$$

Differentiating with respect to $s$

$$\begin{aligned}
\dot{Q}_1(s_i) &= \lambda\sum_{j=1}^{i-1}[-Q_j(s)+Q_{j+1}(s)]\left[\widehat{p}f(j,\lambda\Delta_i)+\widehat{q}f(j-1,\lambda\Delta_i)\right]\\
&\quad+\lambda Q_1(s)\widehat{p}f(0,\lambda\Delta_i)+\lambda\sum_{j=1}^{i-1}Q_j(s)\widehat{p}[f(j,\lambda\Delta_i)\\
&\quad-f(j-1,\lambda\Delta_i)]+\lambda\sum_{j=2}^{i-1}Q_j(s)\widehat{q}[f(j-1,\lambda\Delta_i)\\
&\quad-f(j-2,\lambda\Delta_i)]+Q_1\widehat{q}\lambda f(0,\lambda\Delta_i)\\
&\quad+\lambda Q_0(s)\widehat{p}f(0,\lambda\Delta_i)\quad(\text{from Proposition }3)\\
&= -\lambda\left[\sum_{j=1}^{i-1}Q_j(s)\left[\widehat{p}f(j,\lambda\Delta_i)+\widehat{q}f(j-1,\lambda\Delta_i)\right]\right]\\
&\quad+\lambda\left[\sum_{j=1}^{i-1}Q_j(s)\left[\widehat{p}f(j,\lambda\Delta_i)+\widehat{q}f(j-1,\lambda\Delta_i)\right]\right]\\
&\quad+\lambda\left[\sum_{j=2}^{i-1}Q_j(s)\left[\widehat{p}f(j-1,\lambda\Delta_i)+\widehat{q}f(j-2,\lambda\Delta_i)\right]\right]\\
&\quad-\lambda\left[\sum_{j=2}^{i-1}Q_j(s)\left[\widehat{p}f(j-1,\lambda\Delta_i)+\widehat{q}f(j-2,\lambda\Delta_i)\right]\right]\\
&\quad+\lambda\left[Q_1(s)\widehat{p}f(0,\lambda\Delta_i)-Q_1(s)\widehat{p}f(0,\lambda\Delta_i)\right]\\
&\quad+\lambda Q_0(s)\widehat{p}f(0,\lambda\Delta_i)
\end{aligned}$$

which establishes Equation (26). The other result follows easily by differentiating the expression $\sum_{j=0}^{i}Q_j(s_i) = 1$ with respect to $s$.

**Proof of Proposition 6**. Recall that according to Proposition 6, if the $(n+1)^{st}$ call-in customer is scheduled as the "new" $i$th customer at $s\in u_{i-1}$, then

$$\sum_{l=k}^{m}\dot{Q}_l(s_{i+m-1})\geq 0,\ \text{for}\ \ 1\leq k\leq m, 2\leq m\leq n-i$$

and

$$\sum_{l=k}^{m}\ddot{Q}_l(s_{i+m-1})\geq 0,\ \text{for}\ 1\leq k\leq m, 2\leq m\leq n-i$$

The result is derived using induction. Note that the $(i+m)^{th}$ scheduled customer of the schedule $S_{i-1}^{n+1}$ is scheduled at $s_{i+m-1}$ (see Fig. 2). Denote $\Delta_m = s_{i+m-1}{-}s_{i+m-2}$. Now to establish the base case for induction, we will let $m=2$. There are two cases of $k$ to consider, $k=1$ and $k=2$. We will consider the case for $k=2$ first since it is easier to follow.

Case 1: $k=2$

Applying Equation (4) for $m=2$ and $k=2$ in Equation (29) we get

$$\begin{aligned}
Q_2(s_{i+1}) &= \sum_{j=2}^{i}Q_j(s_i)[p_if(j-1,\lambda\Delta_2)+q_if(j-2,\lambda\Delta_2)]\\
&\quad+Q_1(s_i)p_if(0,\lambda\Delta_2)
\end{aligned}$$

Differentiating with respect to $s$ and using Corollary 1

$$\begin{aligned}
\dot{Q}_2(s_{i+1}) &= \dot{Q}_1(s_i)p_if(0,\lambda\Delta_2)\\
&= Q_0(s)\widehat{p}p_i\lambda f(0,\lambda(s_i-s))f(0,\lambda\Delta_2)\geq 0\\
&\quad(\text{from Proposition }\ 5,\ \text{Equation }(26))
\end{aligned}$$

Further differentiation with respect to $s$ gives

$$\begin{aligned}
\ddot{Q}_2(s_{i+1}) &= \lambda^2[Q_1(s)+Q_0(s)]\widehat{p}f(0,\lambda(s_i-s))p_if(0,\lambda\Delta_2)\geq 0\\
&\quad(\text{from Proposition }\ 5,\ \text{Equation }(21))
\end{aligned}$$

Case 2: $k=1$

Next, for $k=1$ in Equation (29) we have from Equation (4)

$$\begin{aligned}
Q_1(s_{i+1}) &= \sum_{j=1}^{i}Q_j(s_i)[p_if(j,\lambda\Delta_2)+q_if(j-1,\lambda\Delta_2)]\\
&\quad+Q_0(s_i)p_if(0,\lambda\Delta_2)
\end{aligned}$$

Differentiating with respect to $s$

$$\begin{aligned}
\dot{Q}_1(s_{i+1}) &= \dot{Q}_1(s_i)[p_if(1,\lambda\Delta_2)+q_if(0,\lambda\Delta_2)]\\
&\quad+\dot{Q}_0(s_i)p_if(0,\lambda\Delta_2)(\text{from Corollary }\ 1)\\
&= \dot{Q}_1(s_i)[p_if(1,\lambda\Delta_2)+q_if(0,\lambda\Delta_2)]\\
&\quad-\dot{Q}_1(s_i)p_if(0,\lambda\Delta_2)(\text{from }\ \text{Proposition }\ 5)
\end{aligned}$$

Thus for $k=1$ we have,

$$\begin{aligned}
\dot{Q}_1(s_{i+1})+\dot{Q}_2(s_{i+1}) &= \dot{Q}_1(s_i)[p_if(1,\lambda\Delta_2)+q_if(0,\lambda\Delta_2)]\\
&= Q_0(s)\widehat{p}\lambda f(0,\lambda(s_i-s))[p_if(1,\lambda\Delta_2)\\
&\quad+q_if(0,\lambda\Delta_2)]\geq 0(\text{from Proposition }\ 5,\\
&\quad\text{Equation }(26))
\end{aligned}$$

Further differentiating with respect to $s$ and using Proposition 3 and 5

$$\begin{aligned}
\ddot{Q}_1(s_{i+1}){+}\ddot{Q}_2(s_{i+1}) &= \lambda^2[Q_1(s)+Q_0(s)]\widehat{p}f(0,\lambda(s_i-s))[p_if(1,\lambda\Delta_2)\\
&\quad+q_if(0,\lambda\Delta_2)]\geq 0(\text{from Proposition }\ 5,\ \text{Equation }(21))
\end{aligned}$$

Hence Equations (29) and (30) hold for $m=2$. Assume, as the induction hypothesis, that Equations (29) and (30), hold for all integers 3,4,…,$v{-}1$. We now show that the result holds for $v$. Set $\Delta_v = s_{i+v-1}{-}s_{i+v-2}$. We break the proof into three cases for $k$, $k=v$, $k=2,3,\ldots,v{-}1$ and $k=1$.

Case 1: for $k = v$

From Equation (4)

$$Q_v(s_{i+v-1}) = \sum_{j=v}^{i+v-2} Q_j(s_{i+v-2})[p_{i+v-2}f(j+1-v,\lambda\Delta_v) + q_{i+v-2}f(j-v,\lambda\Delta_v)] + Q_{v-1}(s_{i+v-2})p_{i+v-2}f(0,\lambda\Delta_v)$$

Differentiating with respect to $s$ and using Corollary 1 and the induction hypothesis we get

$$\dot{Q}_v(s_{i+v-1}) = \dot{Q}_{v-1}(s_{i+v-2})p_{i+v-2}f(0,\lambda\Delta_v) \geq 0$$

Further differentiating with respect to $s$ and using the induction hypothesis

$$\ddot{Q}_v(s_{i+v-1}) = \ddot{Q}_{v-1}(s_{i+v-2})p_{i+v-2}f(0,\lambda\Delta_v) \geq 0$$

Case 2: for $k = 2,3,\ldots,v-1$

Differentiating Equation (4) with respect to $s$ and using Corollary 1 we get

$$\dot{Q}_k(s_{i+v-1}) = \sum_{j=k}^{v-1} \dot{Q}_j(s_{i+v-2})[p_{i+v-2}f(j+1-k,\lambda\Delta_v) + q_{i+v-2}f(j-k,\lambda\Delta_v)] + \dot{Q}_{k-1}(s_{i+v-2})p_{i+v-2}f(0,\lambda\Delta_v)$$

Thus, using the induction hypothesis

$$\begin{aligned}\sum_{l=k}^{v} \dot{Q}_l(s_{i+v-1}) &= \sum_{l=k}^{v}\sum_{j=l}^{v-1} \dot{Q}_j(s_{i+v-2})[p_{i+v-2}f(j+1-l,\lambda\Delta_v) \\ &\quad + q_{i+v-2}f(j-l,\lambda\Delta_v)] \\ &\quad + \sum_{l=k}^{v} \dot{Q}_{l-1}(s_{i+v-2})p_{i+v-2}f(0,\lambda\Delta_v) \\ &= p_{i+v-2}\sum_{j=0}^{v-k} f(j,\lambda\Delta_v)\left[\sum_{l=j+k-1}^{v-1} \dot{Q}_l(s_{i+v-2})\right] \\ &\quad + q_{i+v-2}\sum_{j=0}^{v-k-1} f(j,\lambda\Delta_v)\left[\sum_{l=j+k}^{v-1} \dot{Q}_l(s_{i+v-2})\right] \geq 0\end{aligned}$$

Further differentiating with respect to $s$, and using the induction hypothesis it follows that $\sum_{l=k}^{v}\ddot{Q}_l(s_{i+v-1}) \geq 0$, for $2 \leq k \leq v$.

Case 3: $k = 1$

Next, for $k = 1$, we have from Equation (4) and Corollary 1

$$\begin{aligned}\dot{Q}_1(s_{i+v-1}) &= \sum_{j=1}^{v-1} \dot{Q}_j(s_{i+v-2})[p_{i+v-2}f(j,\lambda\Delta_v) \\ &\quad + q_{i+v-2}f(j-1,\lambda\Delta_v)] + \dot{Q}_0(s_{i+v-2})p_{i+v-2}f(0,\lambda\Delta_v) \\ &= \sum_{j=1}^{v-1} \dot{Q}_j(s_{i+v-2})[p_{i+v-2}f(j,\lambda\Delta_v) + q_{i+v-2}f(j-1,\lambda\Delta_v)] \\ &\quad - \sum_{j=1}^{v-1} \dot{Q}_j(s_{i+v-2})p_{i+v-2}f(0,\lambda\Delta_v) \\ &\quad \times \left(\because \sum_{j=0}^{i+v-2} Q_j(s_{i+v-2}) = 1\right)\end{aligned}$$

Thus

$$\begin{aligned}\sum_{k=1}^{v} \dot{Q}_k(s_{i+v-1}) &= \dot{Q}_1(s_{i+v-1}) + \sum_{k=2}^{v} \dot{Q}_k(s_{i+v-1}) \\ &= p_{i+v-2}\sum_{j=1}^{v-1} f(j,\lambda\Delta_v)\left[\sum_{l=j}^{v-1} \dot{Q}_l(s_{i+v-2})\right] \\ &\quad + q_{i+v-2}\sum_{j=0}^{v-2} f(j,\lambda\Delta_v)\left[\sum_{l=j+1}^{v-1} \dot{Q}_l(s_{i+v-2})\right] \\ &\quad + p_{i+v-2}f(0,\lambda\Delta_v)\left[\sum_{l=1}^{v-1} \dot{Q}_l(s_{i+v-2}) - \sum_{l=1}^{v-1} \dot{Q}_l(s_{i+v-2})\right] \\ &= p_{i+v-2}\sum_{j=1}^{v-1} f(j,\lambda\Delta_v)\left[\sum_{l=j}^{v-1} \dot{Q}_l(s_{i+v-2})\right] \\ &\quad + q_{i+v-2}\sum_{j=0}^{v-2} f(j,\lambda\Delta_v)\left[\sum_{l=j+1}^{v-1} \dot{Q}_l(s_{i+v-2})\right] \geq 0\end{aligned}$$

Similarly, $\sum_{k=1}^{v}\ddot{Q}_l(s_{i+v-1}) \geq 0$. Hence the result follows by induction.

## References

[1] Cashman SB, Savageau JA, Lemay CA, Ferguson W. Patient health status and appointment keeping in an urban community health center. Journal of Health Care Poor Underserved 2004;15:474488.

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

[3] Chakraborty S, Muthuraman K, Lawley M. Sequential clinical scheduling with patient no-shows and general service time distributions. IIE Transactions 2010;42:1–13.

[4] Daggy J, Lawley M, Willis D, Thayer D, Suelzer D, Sands L, et al. Using no-show modeling to improve clinic performance. Health Informatics Journal 2010; 16(4):246–59.

[5] Denton B, Gupta D. A sequential bounding approach for optimal appointment scheduling. IIE Transactions 2003;35(11):1003–16.

[6] Dervin JV, Stone DL, Beck CH. The no-show patient in the model family practice unit. Journal of Family Practice 1978;7(6):1177–80.

[7] Deyo RA, Inui TS. Dropouts and broken appointments. Medical Care 1980; 18(11):1146–57.

[8] Erdogan SA, Denton B. Dynamic appointment scheduling with uncertain demand. INFORMS Journal of Computing, in press.

[9] Goldman L, Freidin R, Cook EF, Eigner J, Grich P. A multivariate approach to the prediction of no-show behavior in a primary care center. Archives of Internal Medicine 1982;142(3):563–7.

[10] Green LV, Savin S. Providing timely access to medical care: a queueing model. Operations Research 2008;56(3):1526–38.

[11] Gruzd DC, Shear CL, Rodney W. Determinants of no-show appointment behavior: the utility of multivariate analysis. Family Medicine 1986;18: 217–20.

[12] Gupta D, Wang L. Revenue management for a primary care clinic in the presence of patient choice. Operations Research 2007;56(3):576–92.

[13] Hutzschenreuter A. Queueing models for outpatient appointment scheduling. Master's thesis, Ulm University; Germany: 2005.

[14] Kaandorp Guido, Koole Ger. Optimal outpatient appointment scheduling. Health Care Management Science September 2007;10(3):217–29.

[15] Kim S, Giachetti R. 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: 1211–9.

[16] Kopach R, DeLaurentis P, Lawley M, Muthuraman K, Ozsen L, Rardin R, et al. Effects of clinical characteristics on successful open access scheduling. Health Care Management Science 2007;10(2):111–24.

[17] Lee VJ, Earnest A, Chen MI, Krishnan B. Predictors of failed attendances in a multi-specialty outpatient center using electronic databases. BMC Health Services Research 2005;5:51.

[18] Liu L, Liu X. Block appointment systems for outpatient clinics with multiple doctors. Journal of the Operational Research Society 1998;29(12):1254–9.

[19] Luo J, Kulkarni V, Ziya S. Appointment scheduling under patient no-shows and service interruptions. MSOM 2012;14:670–84.

[20] Martinez PE, Algarin H, Beauchamp VE, Lugo C, Ortiz C, Vega M, et al. How do elderly veterans who fail to keep outpatient clinic appointments differ from those who do not. Puerto Rico Health Science Journal 1987;6:141–6.

[21] Muthuraman K, Lawley M. A stochastic overbooking model for outpatient clinical scheduling with no-shows. IIE Transactions 2007;40(9):820–37.

[22] Pesata V, Palliga V, Webb A. A descriptive study of missed appointments: families' perceptions of barriers to care. Journal of Pediatric Health Care 1999; 13(4):178–82.

[23] Turkcan A, Zeng B, Lawley M, Muthuraman K. Sequential clinical scheduling with service criteria. European Journal of Operational Research 2011;214: 780–95.

[24] Zeng B, Lin J, Turkcan A, Lawley M. Clinic scheduling models with overbooking for patients with heterogeneous no-shows probabilities. Annals of Operations Research 2010;178(1):121–44.

**Santanu Chakraborty** works in Enterprise Optimization at United Airlines. He obtained his doctoral degree in 2010 from Purdue University. Before joining United Airlines, Santanu worked at BNSF Railway and as a Visiting Research Scientist at University of Illinois Urbana-Champaign. His research interests include stochastic control, health care operations and revenue management.

**Kumar Muthuraman** is an associate professor with the McCombs School of Business at the University of Texas at Austin. After obtaining his doctoral degree in 2003 from Stanford University, he worked as an assistant professor of Industrial Engineering at Purdue University. In 2007, he joined the University of Texas. His research interests include quantitative finance, health care operations and stochastic control.

**Mark Lawley** is Professor of Biomedical Engineering in the Weldon School of Biomedical Engineering at Purdue University. Before joining Biomedical Engineering in 2007, he served nine years as Assistant and Associate Professor of Industrial Engineering, also at Purdue, two years as Assistant Professor of Industrial Engineering at the University of Alabama, and he has held engineering positions with Westinghouse Electric Corporation, Emerson Electric Company, and the Bevill Center for Advanced Manufacturing Technology. As a researcher in academics, he has authored over 100 technical papers including book chapters, conference papers, and refereed journal articles, and has won four best paper awards for his work in systems optimization and control. He received the PhD in Mechanical Engineering from the University of Illinois at Urbana Champaign in 1995 and passed the Professional Engineers exam in the State of Alabama.