Stochastic Overbooking Model for Outpatient Clinical Scheduling with No-shows

Article *in* IIE Transactions · July 2008

DOI: 10.1080/07408170802165823 · Source: OAI

2 authors:

Mark Lawley
Texas A&M University
170 PUBLICATIONS **6,365** CITATIONS
SEE PROFILE

Kumar Muthuraman
University of Texas at Austin
36 PUBLICATIONS **1,253** CITATIONS
SEE PROFILE

IIE Transactions (2008) **40**, 820–837

ISSN: 0740-817X print / 1545-8830 online
DOI: 10.1080/07408170802165823

A stochastic overbooking model for outpatient clinical scheduling with no-shows

KUMAR MUTHURAMAN[1] and MARK LAWLEY[2,*]

[1] *McCombs School of Business, University of Texas, Austin, TX 78712, USA*
E-mail: kumar.muthuraman@mccombs.utexas.edu
[2] *Weldon School of BioMedical Engineering, Purdue University, West Lafayette, IN 47907, USA*
E-mail: malawley@purdue.edu

Received May 2006 and accepted September 2007

In this paper a stochastic overbooking model is formulated and an appointment scheduling policy is developed for outpatient clinics. The schedule is constructed for a single service period partitioned into time slots of equal length. A clinic scheduler assigns patients to slots through a sequential patient call-in process where the scheduler must provide each calling patient with an appointment time before the patient's call terminates. Once an appointment is added to the schedule, it cannot be changed. Each calling patient has a no-show probability, and overbooking is used to compensate for patient no-shows. The scheduling objective captures patient waiting time, staff overtime and patient revenue. Conditions under which the objective evolution is unimodal are derived and the behavior of the scheduling policy is investigated under a variety of conditions. Practical observations on the performance of the policy are presented.

Keywords: Open access, appointment scheduling, patient no-shows, outpatient clinic operations

1. Introduction

Healthcare currently consumes 15% of the US Gross Domestic Product and is expected to reach 19% within the coming decade (Centers for Medicare and Office of the Actuary Medicaid Services, 2007). These costs are due to factors such as new advances in expensive treatment technologies and pharmaceuticals, unfavorable trends in population demographics such as aging, obesity and chronic disease, and legal expenses resulting from medical errors and malpractice. Faced with this environment of increasing costs, limited capacity and burgeoning demand, many hospitals are emphasizing shorter lengths of stay and are shifting care from inpatient to outpatient facilities. This in turn is forcing outpatient clinical facilities to reassess their operations and capacities, with the dual objectives of stabilizing revenue streams and improving healthcare access.

Access to outpatient facilities is controlled through appointment scheduling. In traditional appointment scheduling, a patient seeking an appointment calls the clinic and is immediately booked for a future appointment time. When the clinic is working close to capacity, the near-term schedule tends to be fully utilized and appointment slots might not be available for many weeks or months. This long lead time is usually unacceptable for ill patients, who must either go without care or seek expensive emergency services. Furthermore, when the appointed time does arrive, the patient's needs could have changed significantly; the patient could have recovered, moved, forgot or died; leading to the problem of patient no-show. In some clinics, up to 42% of scheduled patients fail to show up for pre-booked appointments (Deyo and Inui, 1980). This behavior wastes clinic resources, decreases the quality of care, escalates costs and impacts accessibility. Many factors have been cited as indicators of patient no-show including patient demographics and medical conditions, physician characteristics and patient–physician interactions (Deyo and Inui, 1980).

Because of these problems and trends, many outpatient clinics are experimenting with open access scheduling, where patients get an appointment time within a day or two of when they call (see Forjuoh *et al.* (2001), Kennedy and Hsu (2003), Murray *et al.* (2003) and O'Hare and Corlett (2004) for representative discussions). In essence, there is little long-term pre-booking, clinics book only for a very short time horizon. The hope is that this short horizon will help more patients see their physician when they have a need, not at some distant time in the future. Operationally, as in any forecasting situation, short-term no-show predictions are more reliable, and hence, under open access, can play a more influential role in optimizing clinical patient scheduling. But, as a close reading of the appointment scheduling

*Corresponding author

literature shows, appointment scheduling methods do not fully integrate or exploit patient no-show models. This is unfortunate since patient no-show modeling is an active area of research with many fruitful results (please see Dervin *et al.* (1978), Goldman *et al.* (1982) and Bean and Talaga (1995)).

Overbooking is an important strategy for improving patient access and stabilizing revenue when there is a significant chance that some scheduled patients will not show up. Overbooking has been used in the airline industry for many years where the objective is to book passenger reservations to maximize flight revenue. Typically, the airline booking problem consists of a single-leg scheduled flight with a fixed cost, capacity limits and fares on different class seats, and a low marginal cost of carrying additional passengers. Reservation requests for each of the classes arrive according to a random process for some period of time prior to takeoff. Passengers with reservations may cancel or no-show, in which case some type of refund, possibly not full, is given. Because empty seats at flight time represent lost revenue, overbooking may occur. If overbooked passengers are denied boarding, the airline incurs a bumping penalty. Rothstein (1985) provides an engaging review of the evolution of airline overbooking as an acceptable practice, while McGill and Van Ryzin (1999) provide an informative literature review. Other representative research in airline overbooking includes Robinson (1995), Chatwin (1998, 1999), Coughlan (1999), Subramanian *et al.* (1999), Feng and Xiao (2001) and Karaesmen and Van Ryzin (2004).

Unfortunately, clinical booking has little in common with the airline problem. Although both have significant no-show probabilities, clinical booking has a stochastic service element resulting in critical patient waiting time and staff overtime features absent in the airline problem. Moreover, the decision in airline booking is binary, that is, the agent either reserves or refuses to accommodate the booking request of a potential passenger. In clinical scheduling, apart from the binary decision, the scheduler must search for an optimal appointment time. Furthermore, while airlines incur an explicit financial penalty for overshow situations, system dynamics are not affected. However, overshows in clinics not only result in excessive workload, but also substantially change system dynamics, resulting in longer patient waiting times. Thus, in this paper, we develop an overbooking process that accommodates the detailed requirements and dynamics of the clinical scheduling environment and leverages on no-show patient prediction.

In this work, clinical scheduling and overbooking are essentially problems of assigning appointment seeking patients to time slots. An operational or service period (called a "day", typically 4 or 8 hours), is divided into time periods (called "slots", typically 15, 20 or 30 minutes). When a patient calls for an appointment (typically before the service period begins), the appointment scheduler uses an estimate of the patient's no-show probability (obtained from the patient's attributes and the clinic's no-show model) to choose an appointment slot, which is communicated to the patient before the call ends. During the service period, two types of patients enter any given slot, those who were unserviced in the previous slot and those who arrive for the current slot. A random number of waiting patients are serviced in each slot and the remaining overflow into the next slot. Since patients usually request consultation with a particular physician, we can treat each physician's schedule independently, and thus we can assume a single server.

The objectives are to minimize patient wait times, maximize resource utilization and minimize the number of patients waiting at the end of the day. Patients waiting at the end of the day cannot be dismissed and have to be served during overtime. Because of no-shows, the clinic capacity will usually be underutilized without some overbooking. But, overbooking incurs the risk of overloading the clinic if too many patients show-up. Excess patient arrivals directly increase patient wait times and the number of patient overflows at the end of the day. Thus, an optimal policy must balance the risks of patient waiting, staff overtime and clinic under-utilization. Clearly, this balance is affected by the weights applied to each of the risks. A reasonable approach is to maximize a profit objective where attending patients provide a reward and costs are associated with patient waiting and physician/staff overtime. This is a multi-objective optimization problem with associated costs and rewards serving as weighting coefficients. It will provide the right balance between utilization, waiting time and overtime if the coefficients are properly chosen. While staff/physician-related costs and patient revenues can be explicitly estimated, the cost of patient waiting needs to be estimated based on local clinic conditions and patient demographics.

The contributions of this research are as follows. First, it formulates a model of the call-in scheduling problem and develops a myopic, sequential policy for scheduling call-ins (Section 3). Next, it presents and proves the necessary and sufficient conditions for the objective evolution to be unimodal (Section 4). By unimodal, we mean that the objective is non-decreasing up to a particular call-in patient and then is monotone decreasing thereafter, which guarantees an optimal stopping criterion. That is, once we encounter a decrease in the expected profit objective, we know that continuing to schedule patients will only result in further decreases. This implies that the costs associated with patient waiting times are outweighing the marginal revenues generated, and it is time to terminate the scheduling process for the given service period. Thus, even though our policy is myopic, its ability to generate this unimodal objective evolution is very important. Finally, by using an exhaustive set of numerical examples, the paper develops several insights into the practical characteristics of the policy (Section 5). In particular, we investigate the effect of cost coefficients on slot assignments and objective values and provide a tentative, experimental characterization of how much we lose due to the myopic and sequential nature of our method.

2. Literature review

Cayiril and Veral (2003) provided an extensive review of the appointment scheduling literature, covering 80 papers that span 50 years. They categorize the appointment scheduling literature by the following attributes: (i) static versus dynamic; (ii) performance measures; (iii) system design; and (iv) methodology. They also provide a good discussion of future research directions. In the following, we will briefly discuss (i)–(iv) and then categorize our own work with respect to these attributes. For a detailed discussion and listing of papers, we refer the reader to Cayiril and Veral (2003).

The first classification attribute is static versus dynamic appointment scheduling. In the static case, all decisions about appointment times are made prior to the start of a session, whereas in the dynamic case, appointment times can be adjusted as the system state evolves. The dynamic case is most applicable in situations where patients are already admitted to a hospital and scheduling is being done for some hospital laboratory operation. It has limited application to outpatient settings since, in outpatient scheduling, the schedule for a session tends to be completed before the session begins. Thus, most of the literature focuses on the static case, which typically involves a given set of N punctual patients with independent and identically distributed service times, who are to be scheduled for a single session (day) with a single physician (single server). Complications to the static problem include environmental factors such as physician lateness and interruptions; non-punctual, emergency, walk-in and no-show patients; and multi-stage check-in, service and check-out procedures, all of which are either addressed or at least discussed to some degree in the literature. A representative set of recent static papers includes Ho and Lau (1992), Vanden Bosch *et al*. (1999), Vanden Bosch and Dietz (2000, 2001), Lau and Lau (2000), Denton and Gupta (2003) and Robinson and Chen (2003).

Performance measures dictate how a given schedule is to be evaluated. These are categorized as time, congestion or "fairness" based. Time-based measures typically have some weighted function of patient waiting time, physician idle time and staff overtime. Congestion-based measures capture features such as queue length, utilization of waiting room resources and so forth, where it is important to consider the presence of patient companions. Fairness-based measures try to distribute patient waiting time evenly over the day (in many systems, average waiting time increases throughout the session period so that patients scheduled later in the day experience greater expected waiting). For a detailed review of performance and objective functions, the reader is referred to Mondschein and Weintraub (2003).

The design of an appointment scheduling system is typically specified by three parameters: the "block", the number of patients arriving at the beginning of an appointment period; the "initial block", the number of patients arriving for the initial appointment; and the "interval", the length of the appointment interval which is either fixed or variable. Typical designs include the Individual-block/Fixed-interval in which one patient is scheduled to arrive at the beginning of each appointment interval and each interval is of the same length; another design is the Multiple-block/Fixed-interval, and so forth. Also, the appointment system can be designed to make use of various types of patient classification systems, which tend to classify patients so that better estimations of service times can be attained and adjustments can be made for walk-ins, no-shows, and urgent and emergency patients. For detailed system design studies in complex environments, the reader is referred to Ho and Lau (1992), Liu and Liu (1998), Rohleder and Klassen (2002), Harper and Gamlin (2003), Klassen and Rohleder (2004) and Cayirli *et al*. (2006).

Finally, there are two broad classes of methodology: analytical studies and simulation. Analytical papers use queuing theory, math programming and dynamic programming and tend to focus on the basic appointment scheduling problem with limited consideration of patient-based environmental factors such as no-shows and walk-ins. The simulation studies focus on comparing detailed appointment scheduling systems in complex environments. Representative analytical papers include Vanden Bosch *et al.* (1999), Vanden Bosch and Dietz (2000, 2001), Denton and Gupta (2003), and Robinson and Chen (2003). Simulation studies include Babes and Sarma (1991), Ho and Lau (1992), Ho *et al.* (1995), Rohleder and Klassen (2002), Harper and Gamlin (2003), Klassen and Rohleder (2004) and Cayirli *et al.* (2006). Further, Jun *et al.* (1999) provides a review of simulation studies in health care clinics up to 1999.

Our work can be classified as static, since we do not adjust future scheduled appointment times as patients arrive. However, we note that our problem differs significantly from the typical static problem since we do not assume the complete set of patients to be scheduled is known when scheduling decisions are being made. Rather, our approach builds the schedule sequentially through a call-in process where we assume that each patient must be given an appointment before their call terminates. Furthermore, our patients are classified according to no-show probability that affects how the schedule is built and how many patients are eventually scheduled. Thus, our problem has many dynamic features not found in the typical static problem. Our performance measure is based on a weighted combination of patient waiting time, physician overtime costs and revenues generated for each patient served. We note that physician idle time is not explicitly captured since we consider the scheduling of a physician for a clinic session to be largely a fixed cost. With respect to system design, our work can be classified as Multiple-block/Fixed-interval where the block size can be variable due to overbooking. Finally, our work is based on probabilistic modeling and is therefore analytical.

We close this section by quoting Cayiril and Veral (2003), who say "No rigorous research exists which investigates possible approaches to adjusting the assignment schedule in

order to minimize the disruptive effects of no-shows, walk-ins, and/or emergencies." We view our research as helping to fill this gap.

3. The clinical booking model and scheduling policy

Let the period of interest (typically a day) be divided into I intervals each called a "slot." Each slot $i = 1, 2, \ldots, I$ is of length Δt_i. We assume that patients needing an appointment call in to the scheduler before the beginning of slot 1. These "call-ins" can be scheduled to one of the I slots or rejected, that is, not assigned to any slot. Patients scheduled for each slot have a no-show probability and arrive independently of other patients. Arriving patients join a queue and if they are not serviced in their scheduled slot, they overflow to the next slot. For now we will assume that service times are exponentially distributed (we discuss justifications and relaxations for this assumption later in this section).

At some point during the call-in period, suppose n patients have been scheduled. Let the random variable X_i^n denote the number of patients arriving for slot i and Y_i^n be the number of patients waiting for the completion of service at the end of slot i. That is, the number of patients overflowing from slot i into slot $i+1$ (see Fig. 1). Note that Y_i^n includes the patient that is in service at the end of slot i. Because service times are exponential, the number of service completions in a slot i is the minimum of a Poisson random variable and the number of patients in the slot. If L_i is Poisson with mean, $\lambda \Delta t_i$, then the overflow from slot i is given as

$$Y_i^n = \max\left(Y_{i-1}^n + X_i^n - L_i^n, 0\right). \tag{1}$$

Here, L_i can be interpreted as the number of services that would have been completed provided the queue does not empty, while $\min(L_i,\ Y_{i-1}^n + X_i^n)$ represents the actual number of services completed.

We assume that each scheduled patient has an estimated no-show probability. This probability can be estimated based on patient attributes and the historical data for the patient or for the group of patients with similar attributes. We will categorize the set of patients into J groups depending on their attributes. A patient belonging to group j has a probability $p_j > 0$ of showing up and a probability $1 - p_j$ of not showing up.

The state of the next day's schedule after n call-ins is represented by the matrix $\mathbf{S}^n \in \mathbf{R}^{I \times J}$, whose i,jth element $S_{i,j}^n$ denotes the number of patients of type j scheduled for slot i. The total number of scheduled patients in slot i will be represented by $N_i^n = \sum_j S_{ij}^n$. When the context is clear in the following we often suppress the superscript (as in Fig. 1). We also define the following matrices for further analysis. An assignment matrix $\mathbf{\Delta}^{ij}$ is of size $I \times J$ with a one at the i,jth position and zeros elsewhere. The function $Q(.)$ takes as argument the state matrix $\mathbf{S}$ and gives the *arrival probability matrix*, $\mathbf{Q}(\mathbf{S})$. The i,mth element of $\mathbf{Q}(\mathbf{S})$ denotes the probability of m patients arriving in slot i given the current state $\mathbf{S}$. For notational convenience we also take the matrix $\mathbf{Q}^n \equiv \mathbf{Q}(\mathbf{S}^n)$.

The function $R(\cdot)$ will represent the *overflow probability matrix*, that is, the i,kth element of $\mathbf{R}(\mathbf{S})$ represents the probability of k patients overflowing from slot i. Similarly as in $\mathbf{Q}^n$, $\mathbf{R}^n \equiv \mathbf{R}(\mathbf{S}^n)$. Obviously, $\mathbf{Q}^n$, $\mathbf{R}^n \in \mathbf{R}^{I \times \hat{N}^n}$ where $\hat{N}^n = \max_i N_i^n$. By definition, given $\mathbf{S}$:

$$Q_{im}^n = \Pr\left\{X_i^n = m\right\}, \tag{2}$$

$$R_{ik}^n = \Pr\left\{Y_i^n = k\right\}. \tag{3}$$

Suppose the nth patient calls for an appointment and is of type j. Letting U be the set of slots (that is, integers from one to I), our problem is to choose a slot $i \in U$ for the patient so as to maximize an objective. That is, at each call-in instance, we choose a decision that maximizes $f(\mathbf{Q}^n, \mathbf{R}^n)$, that is, we assign patient n to slot i^* where

$$i^* = \arg\max_{i \in U} f(Q(\mathbf{S}^{n-1} + \mathbf{\Delta}^{ij}), R(\mathbf{S}^{n-1} + \mathbf{\Delta}^{ij})), \text{ and} \tag{4}$$

$$\mathbf{S}^n = \mathbf{S}^{n-1} + \mathbf{\Delta}^{i^*j}. \tag{5}$$

While $\mathbf{S}^n$ will denote the state after an optimal assignment i^*, that is $\mathbf{S}^{n-1} + \mathbf{\Delta}^{i^*j}$, we will use $\mathbf{S}_i^n$ to denote the state where the last assignment is to slot i, which is not necessarily the best assignment, that is $\mathbf{S}_i^n = \mathbf{S}^{n-1} + \mathbf{\Delta}^{ij}$. Similarly $\mathbf{Q}_i^n$ and $\mathbf{R}_i^n$ are the arrival probability matrix and the overflow probability matrix associated with $\mathbf{S}_i^n$.

We take r as the reward for each patient served and let c_i represent the cost or penalty we charge ourselves for making a patient overflow from slot i to slot $i+1$. This provides sufficient flexibility to model the cost of physician and staff overtime by assigning an appropriate overflow cost to the end of the consulting period (assuming that a physician will see all patients before leaving for the day). Hence, our objective will be

$$\begin{aligned} f(\mathbf{Q}^n, \mathbf{R}^n) &= r \sum_i \sum_m m Q_{i,m}^n - \sum_i c_i \sum_k k R_{i,k}^n \\ &= E\left[r \sum_{i=1}^I X_i^n - \sum_{i=1}^I c_i Y_i^n\right]. \end{aligned} \tag{6}$$

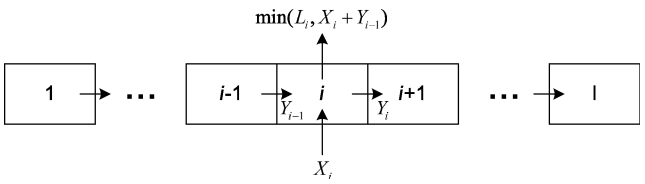

Fig. 1. The system.

3.1. *Calculating* $\mathbf{Q}^n$ *and* $\mathbf{R}^n$

Consider the ith row of a given $\mathbf{S}^n$. We are interested in the probability that m patients arrive given $S_{i,1}^n, S_{i,2}^n, \ldots, S_{i,J}^n$. Let Π be the set of all non-negative, integer, J-vectors

$\boldsymbol{\pi} \equiv (\pi_1, \pi_2, \ldots, \pi_J)$ such that $\sum_{j=1}^{J} \pi_j = m$ and $\pi_j \le S^n_{ij}$ for all j. Then conditioning on the event that π_j number of type j patients show up:

$$\begin{aligned} Q^n_{i,m} &= \Pr\{X^n_i = m\} \\ &= \sum_{\pi\in\Pi} \Pr\{X^n_i = m \,|\, (\pi_1, \ldots, \pi_J)\} \Pr\{(\pi_1, \ldots, \pi_J)\} \\ &= \sum_{\pi\in\Pi} \Pr\{(\pi_1, \ldots, \pi_J)\} \\ &= \sum_{\pi\in\Pi} \prod_j \frac{S^n_{i,j}!}{\pi_j!(S^n_{i,j} - \pi_j)!} p_j^{\pi_j} (1-p_j)^{S^n_{ij} - \pi_j}. \end{aligned} \tag{7}$$

Next consider $R^n_{i,k}$, that is, the probability of k patients overflowing into slot $i+1$ from slot i:

$$\begin{aligned} R^n_{i,k} &= \Pr\{Y_i = k\} \\ &= \Pr\{\max(X_i + Y_{i-1} - L_i, 0) = k\} \\ &= \begin{cases} \Pr\{X_i + Y_{i-1} - L_i = k\} & k > 0 \\ \Pr\{X_i + Y_{i-1} - L_i \le 0\} & k = 0. \end{cases} \end{aligned} \tag{8}$$

Further conditioning yields:

$$\begin{aligned} R^n_{i,0} &= \sum_m \sum_{\tilde{k}} \Pr\{m + \tilde{k} \le L_i\} Q^n_{i,m} R^n_{i-1,\tilde{k}} \\ &= \sum_m \sum_{\tilde{k}} (1 - F_{L_i}(m + \tilde{k})) Q^n_{i,m} R^n_{i-1,\tilde{k}}, \end{aligned} \tag{9}$$

and similarly for $k > 0$,

$$\begin{aligned} R^n_{i,k} &= \sum_m \sum_{\tilde{k}} \Pr\{m + \tilde{k} - k = L_i\} Q^n_{i,m} R^n_{i-1,\tilde{k}} \\ &= \sum_m \sum_{\tilde{k}} f_{L_i}(m + \tilde{k} - k) Q^n_{i,m} R^n_{i-1,\tilde{k}}, \end{aligned} \tag{10}$$

where $F_{L_i}(m) = \Pr\{L_i < m\}$ and $f_{L_i}(m) = \Pr\{L_i = m\}$ are directly obtained from the distribution of service times. Since our service times are taken to be exponentially distributed with mean $1/\lambda\Delta t_i$:

$$f_{L_i}(m) = e^{-\lambda\Delta t_i} \frac{(\lambda\Delta t_i)^m}{m!}, \tag{11}$$

$$F_{L_i}(m) = \sum_{\tilde{m}=0}^{m-1} f_{L_i}(\tilde{m}). \tag{12}$$

It is possible to relax the exponential service time distribution and replace it with general service times. All our calculations above would remain the same. However, relaxing the exponential distribution in our analysis implicitly brings in an approximation. The memoryless property of the exponential distribution helps us ignore the amount of time the person in service at the beginning of a slot has already spent in service. Hence, under a general distribution this ignorance would be an approximation whose quality depends on the service time distribution used. Moreover, observed data suggest exponential service time distributions (DeLaurentis *et al.*, 2006). For these reasons we detail our analysis and results for the exponential service times and simply note that this can be relaxed under the same analysis but that would implicitly mean an approximation. Alternatively, the restriction can be eliminated by including another state variable that records the amount of time the patient in service has spent in service and conditioning all our expectations on this variable.

Equations (7),(9) and (10) enable the calculation of $\mathbf{Q}^n$ and $\mathbf{R}^n$ for a given $\mathbf{S}^n$. For efficient real time application we would obviously like to calculate $Q(\mathbf{S}^{n-1} + \boldsymbol{\Delta}^{ij})$ and $R(\mathbf{S}^{n-1} + \boldsymbol{\Delta}^{ij})$ when $\mathbf{Q}^{n-1}$ and $\mathbf{R}^{n-1}$ are known. That is, given $\mathbf{S}^{n-1}$, $\boldsymbol{\Delta}^{ij}$, $\mathbf{Q}^{n-1}$, $\mathbf{R}^{n-1}$ we are interested in calculating $\mathbf{Q}^n_i$ and $\mathbf{R}^n_i$. When the addition to the schedule is at the ith slot, the arrival probabilities for the other slots are not altered. Hence, $\mathbf{Q}^n_{\tilde{i},m} = \mathbf{Q}^{n-1}_{\tilde{i},m}$ for all $\tilde{i} \ne i$ and for all m. The arrival probabilities for the ith slot, $\mathbf{Q}^n_{i,m}$ can be calculated by conditioning on the arrival probabilities of type j:

$$Q^n_{i,m} = \begin{cases} Q^{n-1}_{i,m}(1-p_j) + \left(Q^{n-1}_{i,m-1}\right)p_j & \text{when } m \ge 1 \text{ and} \\ \\ Q^{n-1}_{i,0}(1-p_j) & \text{when } m = 0. \end{cases} \tag{13}$$

The above equation establishes a recurrence relation that can be used efficiently, not only for the incremental calculation but also for the direct calculation of $\mathbf{Q}^n$ given $\mathbf{S}^n$. For the incremental calculation of the overflow probability matrix $\mathbf{R}^n$, note that $R^n_{\tilde{i},k} = R^{n-1}_{\tilde{i},k}$ for all $\tilde{i} < i$ and for all k and the calculation of $R^n_{\tilde{i},k}$ for $\tilde{i} \ge i$ is best achieved using Equations (9) and (10).

3.2. *The scheduling policy*

The scheduling policy is described below as an algorithm. Note that it enumerates all possible assignments for the current patient and selects the assignment that maximizes the objective function. It is sequential in the sense that it assigns patients as they call and myopic in the sense that it does not consider future arrivals when making the assignment. In Section 5, we investigate the effects of these features on solution quality. Furthermore, the algorithm will reject the patient and terminate when there is no way to schedule the patient without hurting the objective.

Step 1. Set $S_{i,j} = 0$ for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$
$Q_{i,0} = R_{i,0} = 1$ for all $i = 1, \ldots, I$, and $n = 1$.
Step 2. Wait for nth call.
Step 3. nth call occurs and is of type j.
Step 4. For each $i \in U$
4.1 Set $\mathbf{S}^n_i = \mathbf{S}^{n-1} + \boldsymbol{\Delta}^{ij}$.
4.2 Compute $\mathbf{Q}^n_i$ and $\mathbf{R}^n_i$ from $\mathbf{Q}^{n-1}$ and $\mathbf{R}^{n-1}$ using Equations (7),(9) and (10).
4.3 Compute $f^n_i = f(\mathbf{Q}^n_i, \mathbf{R}^n_i)$.
Step 5. If $\max f^n_i \ge f^{n-1}$
5.1 Then $i^* = \arg\max f^n_i$, $\mathbf{S}^n = \mathbf{S}^{n-1} + \boldsymbol{\Delta}^{i^*j}$, $\mathbf{Q}^n = \mathbf{Q}^n_{i^*}$, $\mathbf{R}^n = \mathbf{R}^n_{i^*}$. Set $n = n+1$. Goto Step 2.
5.2 Else Stop.

Note that it is likely that the calling in patient might be given a slot that conflicts with his new personal schedule. To accommodate such possibilities a more flexible version of the above algorithm is required. One of the advantages offered by such myopic scheduling policies is the ease in which such accommodations can be made. To this extent we define $U_n \subset U$ as the set of slots that the nth calling patient prefers scheduling. Then by simply replacing the search for the maximum f_i^n over all i in Step 5 with a search over $i \in U_n$, we accommodate patient preferences in scheduling. However, note that a sequence of U_n's each with one element, would then eliminate any flexibility that is available to the scheduler. While such sequences of U_n's are unlikely in actual practice, theoretically they would be feasible. Any arbitrary objective evolution can be constructed by using such sequences and a stopping criteria as in Step 5.1 necessarily does not guarantee maxima. Hence, in the following section in which we seek theoretical guarantees on the behavior of objective evolutions, we will always restrict our attention to $U_n = U$ for all n. However, we will provide insight into scheduling behaviors for cases where patient preferences $U_n \neq U$, in Section 5.

4. Objective formulation and characterization

This section establishes that our sequential booking policy is unimodal. By unimodal, we mean that the objective is non-decreasing until a particular call-in patient n and then is monotone decreasing after. Thus, if the best assignment for the current call-in patient results in an objective decrease, then all subsequent assignments will lead to additional decreases in the objective. This provides a natural stopping criterion. Theorem 1 and Corollary 1 establish the unimodality of the expected profit. Furthermore, Proposition 2 establishes that $r < c_I$ is both a necessary and sufficient condition for n being finite. Propositions 3 and 4 establish the sufficient and necessary conditions for n being greater than zero, respectively.

First we define some notation. The event $\mathcal{A}_n$ denotes that the nth call-in patient actually shows up for the assigned slot. P_i^n will denote the conditional probability that the assignment of the nth patient increases the overflow from slot i by 1 conditioned on the event $\mathcal{A}_n$. Each patient that shows up is identical from the system perspective. Hence, rearranging their precedence in the waiting queue would not affect any performance parameter. Therefore, to facilitate our analysis, after service completions we will always process the patient who called-in the earliest amongst the waiting patients.

Proposition 1. *$E[Y_i^n]$ is non-decreasing in n. Moreover, if the nth patient is of type j, then $E[Y_i^n] - E[Y_i^{n-1}] = p_j P_i^n$.*

Proof. Say the nth patient is scheduled to slot i_n. If $i_n > i$ then obviously, $E[Y_i^n] = E[Y_i^{n-1}]$. On the other hand if $i_n \leq i$, then we show that $E[Y_i^n - Y_i^{n-1}] \geq 0$. Since for any realization, $Y_i^n \geq Y_i^{n-1}$,

$$E\left[Y_i^n - Y_i^{n-1}\right] = p_j E\left[Y_i^n - Y_i^{n-1} \middle| \mathcal{A}_n\right] > 0. \tag{14}$$

Moreover, the random variable $Y_i^n - Y_i^{n-1}$ indicates the additional number of people showing up in slot i due to the assignment of the nth patient. The worst-case behavior of the system can result in $Y_i^n - Y_i^{n-1} = 1$ and in the best case behavior $Y_i^n - Y_i^{n-1} = 0$. Hence,

$$\begin{aligned} E\left[Y_i^n - Y_i^{n-1}\right] &= p_j E\left[Y_i^n - Y_i^{n-1} \middle| \mathcal{A}_n\right] \\ &= p_j \sum_{y=0,1} y \Pr\left\{Y_i^n - Y_i^{n-1} = y | \mathcal{A}_n\right\} \\ &= p_j \Pr\left\{Y_i^n - Y_i^{n-1} = 1 \middle| \mathcal{A}_n\right\} \\ &= p_j P_i^n. \end{aligned} \tag{15}$$

■

Theorem 1. *If n is such that $f(\mathbf{Q}^n, \mathbf{R}^n) < f(\mathbf{Q}^{n-1}, \mathbf{R}^{n-1})$ then for all $m \geq n$, $f(\mathbf{Q}^m, \mathbf{R}^m) < f(\mathbf{Q}^{m-1}, \mathbf{R}^{m-1})$.*

Proof. Since $f(\mathbf{Q}^n, \mathbf{R}^n) < f(\mathbf{Q}^{n-1}, \mathbf{R}^{n-1})$:

$$\begin{aligned} &E\left[r \sum_{i=1}^{I} X_i^n - \sum_{i=1}^{I} c_i Y_i^n\right] \\ &\quad < E\left[r \sum_{i=1}^{I} X_i^{n-1} - \sum_{i=1}^{I} c_i Y_i^{n-1}\right]. \end{aligned} \tag{16}$$

Rearranging to have rewards on the left-hand side and costs on the right-hand side:

$$r E\left[\sum_{i=1}^{I} \left(X_i^n - X_i^{n-1}\right)\right] < E\left[\sum_{i=1}^{I} \left(c_i Y_i^n - c_i Y_i^{n-1}\right)\right]. \tag{17}$$

The expectation on the left hand side simply denotes the probability of the nth patient showing up and is p_j. Hence, we can write:

$$\begin{aligned} r p_j &< E\left[\sum_{i=1}^{I} c_i \left(Y_i^n - Y_i^{n-1}\right)\right] \\ &< p_j \sum_{i=1}^{I} c_i P_i^n \quad \text{(from Proposition 1).} \end{aligned} \tag{18}$$

Hence,

$$r < \sum_{i=1}^{I} c_i P_i^n. \tag{19}$$

Now denote the slot to which the nth patient was assigned as i_n. The only way that the assignment of patient n to slot i_n can result in one more patient overflowing from slot i, is when in each slot, from i_n to i, the number of patients serviced is less than the number in the waiting queue. Hence,

$$P_i^n = \begin{cases} \prod_{\tilde{i}=i_n}^{i} \Pr\left\{L_{\tilde{i}} < X_{\tilde{i}}^n + Y_{\tilde{i}-1}^n \middle| \mathcal{A}_n\right\} & \text{if } i_n \leq i \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

Now since the assignment of slot i_n to the nth patient is made by Equation (4), we have due to the fact that the chosen slot yielded the best objective:

$$f(\mathbf{Q}^n, \mathbf{R}^n) \geq f\left(Q^n_{i_{\tilde{n}}}, R^n_{i_{\tilde{n}}}\right) = f(\mathbf{Q}^{n-1} + \mathbf{\Delta}^{i_{\tilde{n}}j}, \mathbf{R}^{n-1} + \mathbf{\Delta}^{i_{\tilde{n}}j}) \quad \forall i_{\tilde{n}} \in [1, \ldots, I]. \tag{21}$$

Subtracting $f(\mathbf{Q}^{n-1}, \mathbf{R}^{n-1})$ from both sides:

$$f(\mathbf{Q}^n, \mathbf{R}^n) - f(\mathbf{Q}^{n-1}, \mathbf{R}^{n-1}) \geq f\left(Q^n_{i_{\tilde{n}}}, R^n_{i_{\tilde{n}}}\right) - f\left(\mathbf{Q}^{n-1}, \mathbf{R}^{n-1}\right) \quad \forall i_{\tilde{n}} \in [1, \ldots, I].$$

Substituting for $f(\cdot, \cdot)$, and simplifying using Proposition 1 and Equation (20):

$$p_j r - p_j \sum_{i=1}^{I} c_i \prod_{\tilde{i}=i_n}^{i} \Pr\left\{L_{\tilde{i}} < X^n_{\tilde{i}} + Y^n_{\tilde{i}-1} | \mathcal{A}_n\right\} \geq p_j r - p_j \sum_{i=1}^{I} c_i \times \prod_{\tilde{i}=i_{\tilde{n}}}^{i} \Pr\left\{L_{\tilde{i}} < X^n_{\tilde{i}} + Y^n_{\tilde{i}-1} | \mathcal{A}_n\right\} \qquad \forall i_{\tilde{n}} \in [1, \ldots, I].$$

That is,

$$\sum_{i=1}^{I} c_i \prod_{\tilde{i}=i_n}^{i} \Pr\left\{L_{\tilde{i}} < X^n_{\tilde{i}} + Y^n_{\tilde{i}-1} | \mathcal{A}_n\right\} \leq \sum_{i=1}^{I} c_i \prod_{\tilde{i}=i_{\tilde{n}}}^{i} \Pr\left\{L_{\tilde{i}} < X^n_{\tilde{i}} + Y^n_{\tilde{i}-1} | \mathcal{A}_n\right\} \qquad \forall i_{\tilde{n}} \in [1, \ldots, I]. \tag{22}$$

Now consider P^m_i for $m > n$ and say the mth patient is assigned to slot i_m. Since $m > n$, we have for any realization, $X^m_i \geq X^n_i$ and $Y^m_i \geq Y^n_i$, therefore:

$$\begin{aligned}
\sum_{i=1}^{I} c_i P^m_i &= \sum_{i=1}^{I} c_i \prod_{\tilde{i}=i_m}^{i} \Pr\left\{L_i < X^m_{\tilde{i}} + Y^m_{\tilde{i}-1} | \mathcal{A}_n\right\} \\
&\geq \sum_{i=1}^{I} c_i \prod_{\tilde{i}=i_m}^{i} \Pr\left\{L_i < X^n_{\tilde{i}} + Y^n_{\tilde{i}-1} | \mathcal{A}_n\right\} \\
&\geq \sum_{i=1}^{I} c_i \prod_{\tilde{i}=i_n}^{i} \Pr\left\{L_i < X^n_{\tilde{i}} + Y^n_{\tilde{i}-1} | \mathcal{A}_n\right\} \\
&\qquad \text{(from Equation (22))} \\
&= \sum_{i=1}^{I} c_i P^n_i.
\end{aligned} \tag{23}$$

From Equation (19):

$$r < \sum_{i=1}^{I} c_i P^n_i \leq \sum_{i=1}^{I} c_i P^m_i. \tag{24}$$

Say patient m is of type $\tilde{j}$, then multiplying by both sides by $p_{\tilde{j}}$:

$$r p_{\tilde{j}} < p_{\tilde{j}} \sum_{i=1}^{I} c_i P^m_i. \tag{25}$$

Which implies as earlier:

$$r E\left[\sum_{i=1}^{I} \left(X^m_i - X^{m-1}_i\right)\right] < E\left[\sum_{i=1}^{I} c_i \left(Y^m_i - Y^{m-1}_i\right)\right], \tag{26}$$

or equivalently

$$f(\mathbf{Q}^m, \mathbf{R}^m) < f\left(\mathbf{Q}^{m-1}, \mathbf{R}^{m-1}\right). \tag{27}$$

■

Corollary 1. *If n is such that $f(\mathbf{Q}^n, \mathbf{R}^n) \geq f(\mathbf{Q}^{n-1}, \mathbf{R}^{n-1})$ then for all $m \leq n$, $f(\mathbf{Q}^m, R^m) \geq f(\mathbf{Q}^{m-1}, \mathbf{R}^{m-1})$.*

Proof. Follows directly from Theorem 1. ■

Proposition 2. *There exists an n such that $f(\mathbf{Q}^n, \mathbf{R}^n) < f(\mathbf{Q}^{n-1}, \mathbf{R}^{n-1})$ if and only if $r < c_I$.*

Proof.

$$f(\mathbf{Q}^n, \mathbf{R}^n) - f(\mathbf{Q}^{n-1}, \mathbf{R}^{n-1}) = p_j \left(r - \sum_{i=1}^{I} c_i P^n_i\right). \tag{28}$$

hence, we need to show that there exists an n such that $r < \sum_{i=1}^{I} c_i P^n_i$ if and only if $r < c_I$.

First say that there exists an n such that $r < \sum_{i=1}^{I} c_i P^n_i$, then for all $m > n$ from Theorem 1, $r < \sum_{i=1}^{I} c_i P^m_i$. Let the assignment of the mth patient be to slot i_m. Then $m \to \infty$, $P^m_i \to 1$ for $i \geq i_m$ and $P^m_i = 0$ for $i < i_m$. Hence, from the minimization in Equation (4), for very large m, $i_m = I$. Which implies that:

$$\begin{aligned}
r &< \sum_{i=1}^{I} c_i P^m_i \\
&= c_I.
\end{aligned} \tag{29}$$

Now if $r < c_I$, say there does not exist an n such that $r < \sum_{i=1}^{I} c_i P^n_i$. This implies that:

$$\begin{aligned}
r &> \lim_{n \to \infty} \sum_{i=1}^{I} c_i P^n_i \\
&= c_I,
\end{aligned} \tag{30}$$

a contradiction. Hence, we have that a necessary and sufficient condition for the existence of an n such that $f(\mathbf{Q}^n, \mathbf{R}^n) - f(\mathbf{Q}^{n-1}, \mathbf{R}^{n-1}) < 0$ is $r < c_I$. ■

Proposition 3. *A sufficient condition for $f(\mathbf{Q}^1, \mathbf{R}^1) > f(\mathbf{Q}^0, \mathbf{R}^0)$ is given by: $r > \sum_{i=i_n}^{I} c_i e^{-\lambda \Delta t_i (i - i_n + 1)}$ for all i_n.*

Proof. Obviously $f(\mathbf{Q}^0, \mathbf{R}^0) = 0$, since we have no scheduled patients. Hence, we are interested in establishing the necessary and sufficient conditions for $f(\mathbf{Q}^1, \mathbf{R}^1) > 0$. First we show that if $r > \sum_{i=i_n} c_i e^{-\lambda \Delta t_i (i - i_n + 1)}$ for all i_n then $f(\mathbf{Q}^1, \mathbf{R}^1) > 0$.

If the first arriving patient is of type j:

$$\begin{aligned} f(\mathbf{Q}^1, \mathbf{R}^1) &= p_j \left(r - \sum_i^I c_i P_{1,i} \right) \\ &= p_j \left(r - \min_{i_n} \left\{ \sum_{i=1}^{I} c_i \prod_{\tilde{i}=i_n}^{i} \Pr\left\{ L_{\tilde{i}} < X_{\tilde{i}}^1 + Y_{\tilde{i}-1}^1 | \mathcal{A}_1 \right\} \right\} \right) \\ &\geq p_j \left(r - \sum_{i=1}^{I} c_i \prod_{\tilde{i}=i_n}^{i} \Pr\left\{ L_{\tilde{i}} < X_{\tilde{i}}^1 + Y_{\tilde{i}-1}^1 | \mathcal{A}_1 \right\} \right) \\ &\quad \forall i_n = 1, \ldots, I. \end{aligned} \tag{31}$$

Since only one patient exists in the system entering in slot i_n, the patient overflows out of slot i only when zero patients are served in each slot from i_n to i. Hence, the above probability corresponds to serving zero patients in each slot from i_n to i:

$$\begin{aligned} f(\mathbf{Q}^1, \mathbf{R}^1) &\geq p_j \left(r - \sum_{i=1}^{I} c_i e^{-\lambda \Delta t_i (i - i_n + 1)} \right) \\ &> 0. \end{aligned} \tag{32}$$

■

Proposition 4. *The necessary condition for* $f(\mathbf{Q}^1, \mathbf{R}^1) > f(\mathbf{Q}^0, \mathbf{R}^0)$ *is given by:* $r > \min_{i_n} \sum_{i=i_n} c_i e^{-\lambda \Delta t_i (i - i_n + 1)}$.

Proof. Next we show that if $f(\mathbf{Q}^1, \mathbf{R}^1) > 0$ then $r > \min_{i_n} \sum_{i=1}^{I} c_i e^{-\lambda \Delta t_i (i - i_n + 1)}$. Proceeding similarly as in above we have:

$$\begin{aligned} f(\mathbf{Q}^1, \mathbf{R}^1) &= p_j \left(r - \sum_i^I c_i P_{1,i} \right) \\ &= p_j \left(r - \min_{i_n} \left\{ \sum_{i=1}^{I} c_i \prod_{\tilde{i}=i_n}^{i} \Pr\left\{ L_{\tilde{i}} < X_{\tilde{i}}^1 + Y_{\tilde{i}-1}^1 | \mathcal{A}_1 \right\} \right\} \right) \\ &> 0. \end{aligned} \tag{33}$$

Hence,

$$\begin{aligned} r &> \min_{i_n} \left\{ \sum_{i=1}^{I} c_i \prod_{\tilde{i}=i_n}^{i} \Pr\left\{ L_{\tilde{i}} < X_{\tilde{i}}^1 + Y_{\tilde{i}-1}^1 | \mathcal{A}_1 \right\} \right\} \\ &= \min_{i_n} \sum_{i=1}^{I} c_i e^{-\lambda \Delta t_i (i_n - i + 1)}. \end{aligned} \tag{34}$$

■

5. Results and insights

This section will discuss some insights into various aspects of our scheduling policy. Using examples, we will illustrate the objective evolution as the call-in period progresses, observe the resulting slot assignments, and compare these with a policy that does not consider no-show probabilities or overflows. We will also examine the effect of overflow cost coefficients on slot assignments and expected profits, and we will investigate the "sequence" effect by generating schedules for the same set of patient call-ins, sequenced in many different ways.

Unless specified otherwise, for examples considered in this section, we set the number of slots to eight, that is $I = 8$, with $\Delta t_i = 30$ minutes and $\lambda = 3$. There will be three classes of patients, that is, $J = 3$. While the overflow cost for the last slot (c_I) is higher than the overflow costs during the day, the overflow costs during the day will be identical. Hence, in the following, we will always consider cases with $c_I > c_i$ when $i < I$ and take c_i to be a constant for all $i = 1, \ldots, I - 1$. For notational convenience, c_i will denote c_i for all $i = 1, \ldots, I - 1$. The reward per patient processed will be $r = 100$. The sequence of patient types for the examples are generated by sampling the J types uniformly.

5.1. *Illustrating the scheduling mechanism*

The objective of this subsection is two-fold. First, we illustrate the evolution of the objective using a specific example. Second, we will shed some light into the dependence of our approach's impact on no-show probabilities. As the measure of success, one would like to compare the proposed approach to the one most commonly used. Unfortunately there is no specific approach adopted by most clinics. In most cases, including our clinical partners, the scheduling approach is somewhere between arbitrary/random scheduling based on preferences and availability and a Round-Robin-based scheduling with accommodations for patient preferences. A Round Robin policy assigns the ith customer to slot $((i - 1) \bmod I) + 1$. The advantage with the Round Robin approach is its simplicity and ease of implementation. It also aligns with the notion that spreading out the schedule throughout the day in an even manner is likely to diffuse chaotic situations. This as one can expect is a reasonable policy provided all patients show up, but when patients sometimes do not show up and have varied probabilities of showing up, the Round Robin approach is easily and significantly bettered by the proposed scheduling policy. Hence, in this section we will compare our policy with a Round-Robin-based policy.

Figure 2 illustrates the evolution of our profit objective for an example with $c_i = 40$ and $c_I = 200$. We use a $\boldsymbol{p} = (0.1, 0.5, 0.9)$ and the call-in sequence was simulated using uniform sampling among the three classes. Note that p_j refers to the probability that a patient of type j will show up. The sequence of patient call-ins is given along the abscissa. The left ordinate represents the expected profit of a current schedule and the right ordinate represents the slot. For each patient (on the abscissa), we can read the slot assignment from the right ordinate and the expected profit associated with the current schedule from the left. For example, the first patient is assigned slot 1 with a corresponding profit value of 48.95, the second to slot 4 with profit 97.90, and so

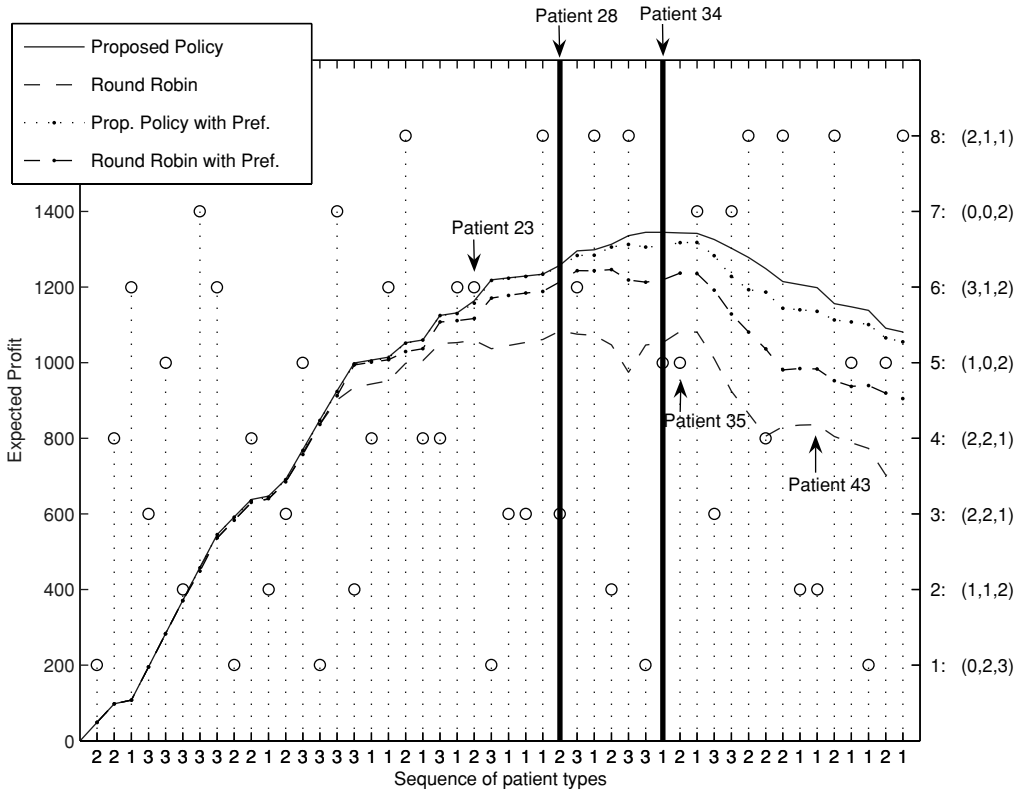

Fig. 2. The schedule and expected profit evolution, $\boldsymbol{p} = (0.1, 0.5, 0.9)$.

forth. Four profit curves are displayed. The curves marked Proposed Policy and Round Robin plot the evolution of the objectives associated with the schedule constructed by our booking policy and the Round Robin policy, respectively.

The right ordinate also gives the final assignment of patients to slots using our policy. For example, slot 4 has (2, 2, 1) indicating that there are two patients of types 1 and 2 and one patient of type 3 in the slot. Figures 3 and 4 provide additional information on the evolution of expected overflow. The expectation of Y_I provides expected number of patients that need to be served at overtime costs and $\sum_i E[Y_i]/n$ indicates the waiting time per patient in terms of the expected number of slots each patient is expected to overflow. Again, the solid lines represent the overflow associated with the schedule constructed by our booking policy, while the dashed presents the overflow of the Round Robin approach. Note that these curves terminate at patient 34.

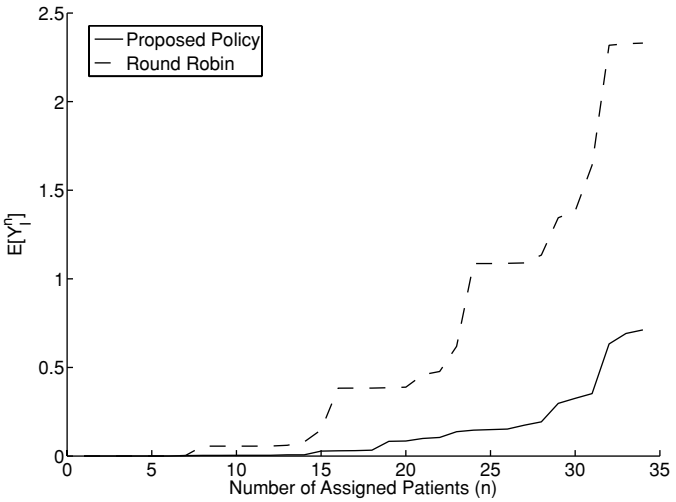

Fig. 3. Expected overflow from slot I.

Next we consider how patient preferences might affect performance of the Proposed Policy. To this extent we will let each incoming patient have either $U_n = (1, \ldots, 4)$ or $U_n = (5, \ldots, 8)$ with equal probability, implying patient preferences to mornings and afternoons with equal probability. We will compare this to the performance of a Round Robin approach where the nth morning preferred patient is assigned to slot $((n-1) \bmod 4) + 1$ and the nth evening preferred patient is assigned to slot $((n-1) \bmod 4) + 5$. The curves marked Prop. Policy with Pref. and Round Robin with Pref. in Fig. 2 plot the respective objective evolutions.

There are several points that we want to address. First, note that the profit curve of our approach exhibits a unique local maximum, as we established in the last section. In general, this property does not hold across all scheduling policies (as demonstrated by the Round Robin objective evolution), which complicates the selection of a stopping criterion. For example, the Round Robin approach exhibits local maxima at patients 23, 28, 35 and 43, and thus it

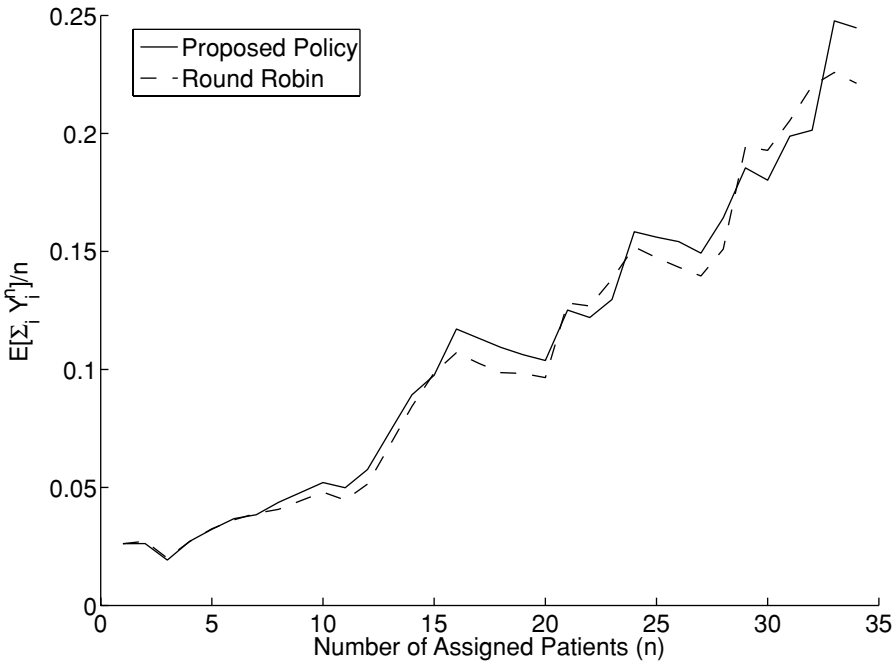

Fig. 4. Expected number of slot overflows per patient.

would not be clear when patient scheduling should be terminated. Furthermore, our global maximum occurs at patient 34 with a profit of 1344.83, while the Round Robin approach yields a global maximum profit of 1084.10 at 28 patients and a profit of 1053.10 at 34 patients. From Fig. 3, we see that the overflow from period 8 is significantly higher for the Round Robin approach, while the average overflow from the other slots (Fig. 4) is approximately the same for the two approaches. This reflects the fact that our booking policy optimally responds to the more severe overtime cost, while the Round Robin approach does not.

Accommodating patient preferences decreases the choices available to the scheduler. Hence, under the Proposed Policy we expect the objective evolution to decrease. For the example in Fig. 2, the decrease in objective due to patient preference is around 2%. On the other hand the modified Round Robin approach (which accommodates patient preferences) could do better or worse than the simple Round Robin approach.

Since the above illustrations are based on one sequence of call-ins, we run 2500 simulated call-in sequences and summarize our observations in Fig. 5. This figure illustrates the distribution of the percentage improvement of the Proposed Policy over the Round Robin approach, that is

$$100 \times \left. \frac{f^n_{\text{Proposed}} - f^n_{\text{Round Robin}}}{f^n_{\text{Proposed}}} \right|_{n=\arg\max_m f^m_{\text{Round Robin}}} \tag{35}$$

for the 2500 runs. Note that we evaluate the difference at the stopping point of our approach since there is no specific stopping criteria for the Round Robin approach. This results in a distribution with a mean 5.22 and a standard deviation of 3.92. Furthermore, since the global maxima is not identifiable with the Round Robin approach, termination at the first local maxima yields a 11.65% greater objective for the Proposed Policy for the 2500 runs. Also on average the decrease in objective due to patient preferences under the Proposed Policy is only 0.86%.

We next consider a $\boldsymbol{p} = (0.25, 0.5, 0.75)$, that has the same average no-show probability as the earlier $p = (0.1, 0.5, 0.9)$ but a smaller spread. For the same call-in sequence (as in Fig. 2), Fig. 6 plots the objective evolution. For this specific case, the percentage difference in profit observed between the proposed and the Round Robin policies as measured by the statistic in Equation (35) is now 12.32% as opposed to 27.70% when $\boldsymbol{p} = (0.1, 0.5, 0.9)$. This suggests that the impact of the proposed scheduling policy increases with the variation in no-show probabilities. One might be inclined to believe that a decrease in no-show rates might decrease the impact of the Proposed Policy over the Round Robin approach. However, the variation in the no-shown rates is the primary factor that influences the impact and not the no-show rates. To illustrate this, Fig. 7 plots the objective evolution for the same sequence with $\boldsymbol{p} = (0.25, 0.5, 0.9)$. This decreases the average no-show rate while increasing the variation and this increases the percentage difference to 20.67%, thereby clearly indicating that a decrease in no-show rates can actually amplify the impact of our policy just because the variation increased.

We next consider another set of no-show probabilities $\boldsymbol{p} = (0.81, 0.86, 0.91)$, $(0.83, 0.88, 0.93)$, $(0.85, 0.90, 0.95)$,

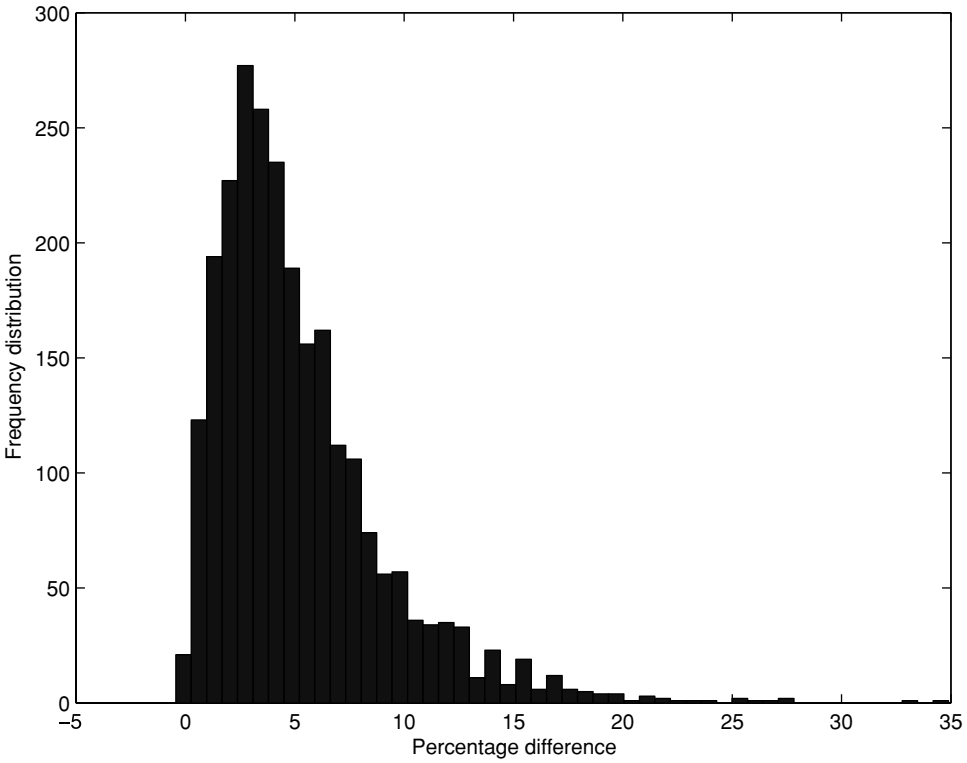

Fig. 5. Percentage improvement over Round Robin.

(0.87, 0.92, 0.97), (0.89, 0.94, 0.99). These have the same constant variance but much lower average no-show rates. The statistic in Equation (35) for the specific sequence considered in Fig. 2 is plotted in Fig. 8. The point of this figure is to illustrate that the average no-show rate has no specific direction of impact. The low percentage difference in profit observed in Fig. 8 is due to the very low variance in the $\boldsymbol{p}$ values used, that is, 0.001 67.

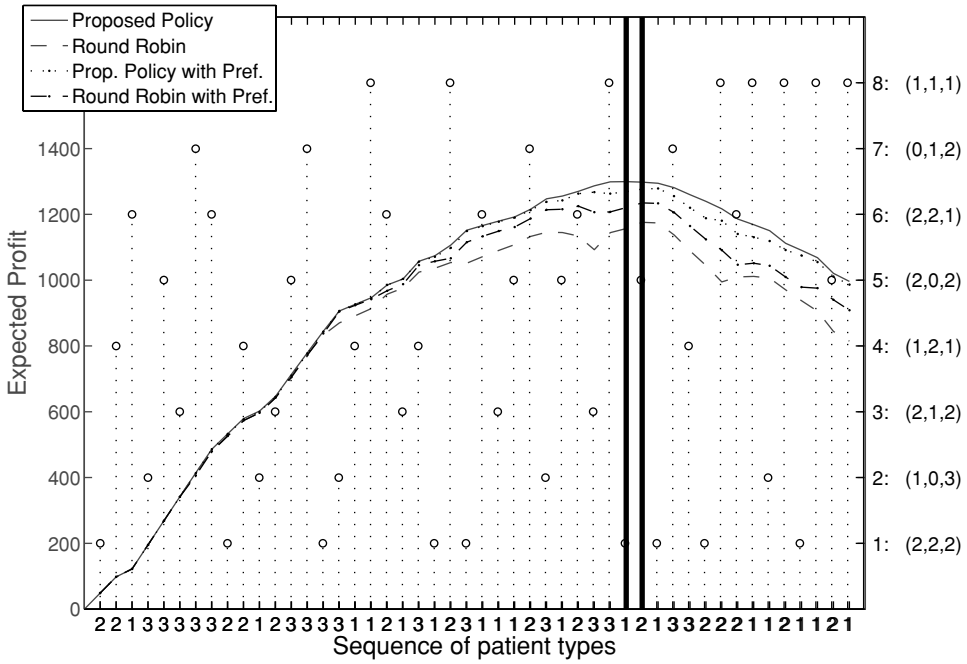

Fig. 6. The schedule and expected profit evolution, $\boldsymbol{p} = (0.25, 0.5, 0.75)$.

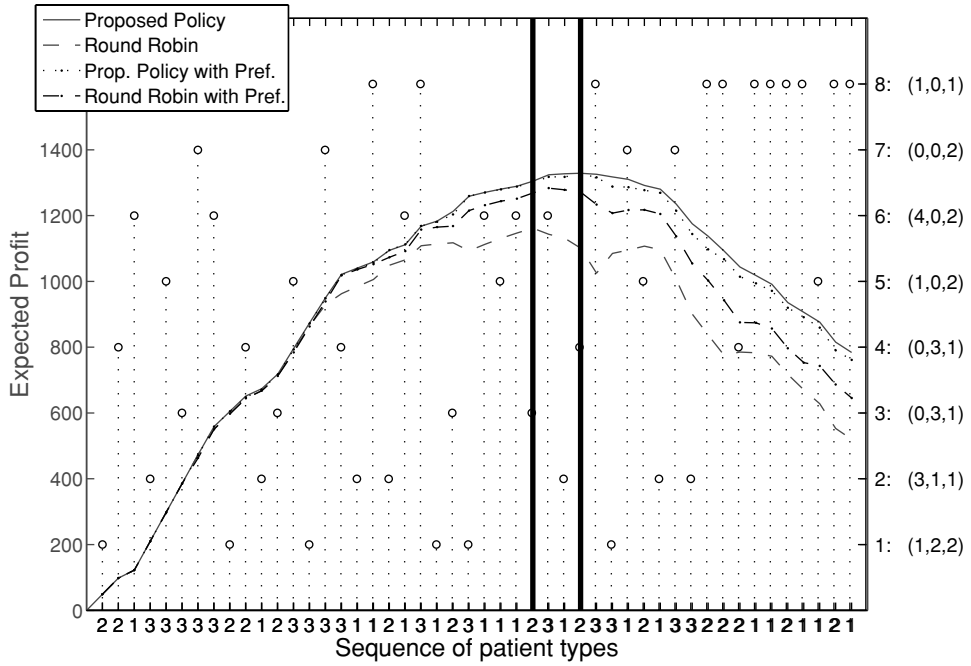

Fig. 7. The schedule and expected profit evolution, $\boldsymbol{p} = (0.25, 0.5, 0.9)$.

Again we confirm these observations on 2500 simulated call-in sequences for each case. The average improvement in the objective function yielded by the Proposed Policy is 5.22, 2.76 and 3.98 for the cases $\boldsymbol{p} = (0.1, 0.5, 0.9)$, $\boldsymbol{p} = (0.25, 0.5, 0.75)$ and $\boldsymbol{p} = (0.25, 0.5, 0.9)$, respectively. Note that the magnitudes of these percentage improvements by themselves do not matter since arbitrarily large values can be obtained by choosing appropriate objective coefficients r and c_i. It is the relative differences that provide insight. The variation in no-show rates have a strong positive impact in the percentage difference over Round Robin while the average no-show rate does not have a specific direction of impact. Restricting attention to a single $\boldsymbol{p}$ will aid comparisons and interpretations. In the interest of not overestimating the potential of the Proposed Policy, we will choose the $\boldsymbol{p}$ what has the lowest average gain over the Round Robin policy. That is, in the following $\boldsymbol{p} = (0.25, 0.5, 0.75)$ unless mentioned otherwise.

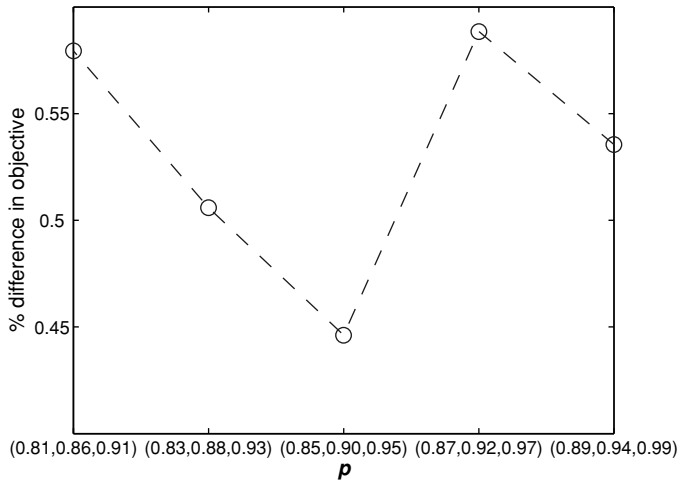

Fig. 8. Percentage difference in objective, for low no-show rates with constant variance.

It is also interesting to examine the sequence of slot assignments. Our approach assigned patient 1 to slot 1, patient 2 to slot 4, patient 3 to slot 6, and so forth. Enumerating further, we have the slot assignment sequence 1,4,6,3,5,2,7,6,1,4,2,3,. . ., and we see that, in this case at least, consecutively assigned slots tend to be spaced well apart. This results from the policy's attempt to reduce overflow between slots and is a function of the overflow costs and service rates. It is also not surprising that the average number of assignments to the later slots would be less than the earlier one. On average, for the cost structure in this example ($c_i = 40$, $c_I = 200$ and $r = 100$), our approach tends to load the first slot more heavily, the intermediate slots uniformly at a lower level, and the last slot at the lowest level, as illustrated in Fig. 9. Note that Fig. 9 gives the average percent of patients assigned to each slot for our example (based on 2500 call-in sequences) with $\boldsymbol{p} = (0.25, 0.5, 0.75)$. On average, about 18% of patients go to the first slot, 12 to 13% go to each intermediate slot, and around 7% go to the last slot. Of course these percentages are highly dependent on the cost structure, an issue we will address in the next subsection.

Finally, in Fig. 2, we continued assigning patients after reaching the local maxima just to see how the cost curve and assignment process behaves. In practice, this represents the case where the scheduler is forced to keep accepting patients

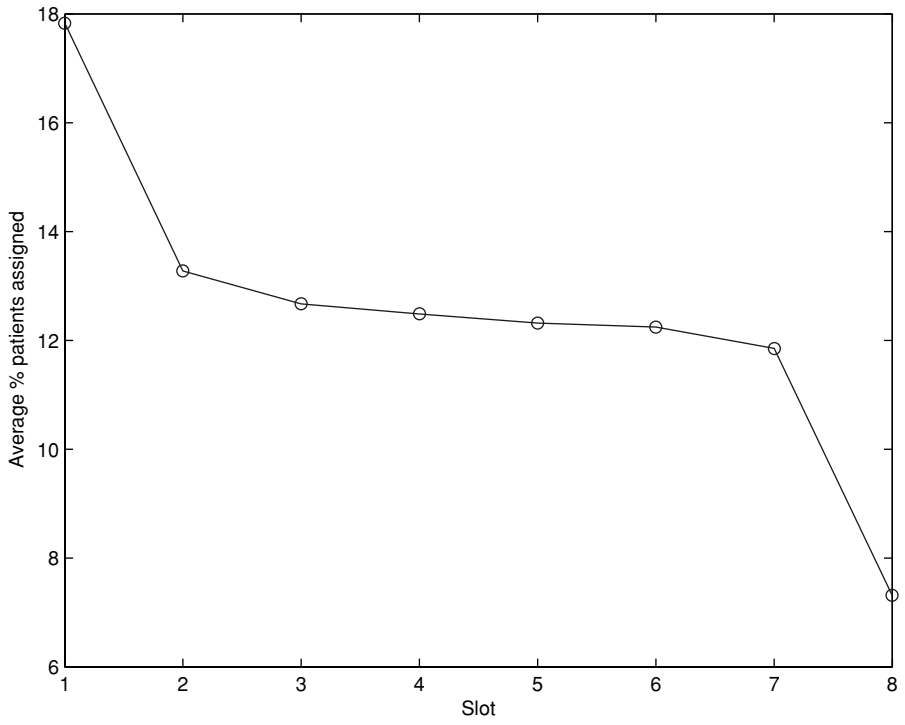

Fig. 9. Percentage assignments per slot.

beyond the global maximum. After the global maximum is attained, the profit curve declines rapidly, indicating that overtime and waiting costs for additional patients increasingly outweigh additional revenues. During this period of decline, close to half of the 14 additional patients go to the last two slots, with four going to the last slot. This indicates that these patients will almost certainly cause additional overflow in all subsequent slots, and thus the least expensive assignment will be to the last slot.

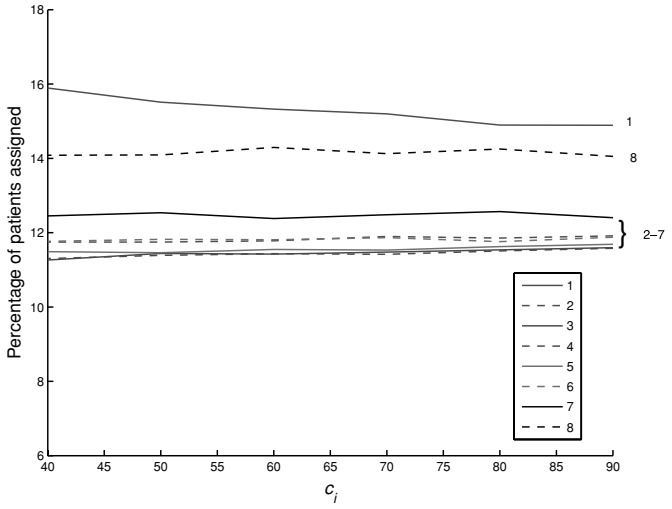

Fig. 10. Percentage assignments per slot for various c_i (averaged over 5000 sequences).

5.2. *Sensitivity to cost coefficients*

The previous subsection illustrates the case for $c_i = 40$, $c_I = 200$ with the reward $r = 100$. Under such a cost structure, fewer assignments are made towards the end of the day and the percentage of assignments to slots 2–7 is roughly uniform. We next want to investigate how these slot assignments are affected by changes in the cost coefficients. To see this, we generated 5000 call-in sequences and used our policy to schedule these for various cost coefficients. Fig. 10 plots the percentage assignment to each slot with $c_I = 100$, $r = 100$ and varying c_i, while Fig. 11 plots the percentage of assignment to each slot with $c_i = 40$, $r = 100$ and varying c_I. We see that assignments to slots 2–7 tend to be much less sensitive to the cost coefficients, while assignments to all slots are more sensitive to changes in c_I than to changes in c_i. With increasing c_I, as one would expect, there is a significant decrease in numbers assigned to slot 8, with most of the decrease in slot 8 going to slot 1, and the rest being evenly assigned to slots 2–7.

Figures 12 and 13 plot the expected profits with increase in the cost coefficients. While the dotted line plots the average expected profit over the 5000 sequences, the solid lines plot the tenth and the 90th percentiles. In this case, unlike in Figs. 10 and 11, the sensitivity to c_i is greater than the sensitivity to c_I. This is understandable, since increasing c_I by say a dollar would make the scheduler less inclined towards slot 8, at which point patients not assigned to slot 8 can be distributed across seven other slots. On the other hand, when c_i is increased by a dollar, the scheduler becomes less

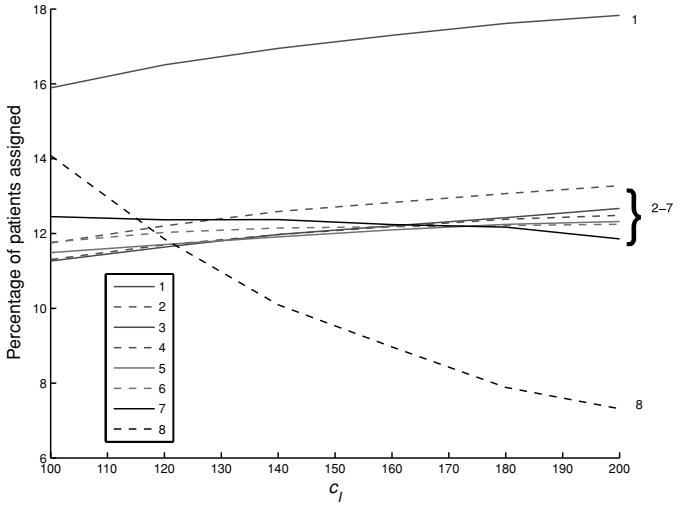

Fig. 11. Percentage assignments per slot for various c_I (averaged over 5000 sequences).

inclined towards assignment to slots 1–7, at which point patients not assigned to slots 1–7 have to go to slot 8. Notice that these expected cost curves are convex and will go to zero as the costs go to infinity. This is because when there is an infinite cost for overflow and a finite reward, the optimal decision is to schedule none. Figures 14 and 15 plot the three components of the expected profit equation. These show that the decrease in expected profit is the result of a large decrease in revenues as well as a smaller decrease in the cost.

5.3. *Sensitivity to average no-show rates and number of patient types*

In this subsection we will consider the effect that changes in the no-show rates and number of slots have on the maximum profit and the number of patients scheduled. To measure the effect of change in average no-show probability, we consider three cases, all three with $\boldsymbol{p} = (0.25, 0.5, 0.75)$, but with different sets of arrival rates for the three cases. We use the weights $(1, 2, 3)$ for the first, $(1, 1, 1)$ for the second and $(3, 2, 1)$ for the third. These imply for example that in the first case an arrival is of type 1 with probability $1/6$, type 2 with probability $2/6$ and type 3 with probability $3/6$. The three cases then correspond to average no-show probabilities of 0.417, 0.5 and 0.583, respectively.

It is important to use a consistent set of sample paths for each of the 1000 simulations. Hence, we first generate 1000 sequences of underlying call-ins with six types of patients. Then to obtain the sample path that feeds the scheduler for case 1, we relabel patient type 1 as type 1, patient types 2 and 3 as type 2 and patient types 4–6 as type 3. Using the same underlying sample paths we similarly obtain the sample paths for cases 2 and 3 with relabeling that is consistent with the weights. We summarize

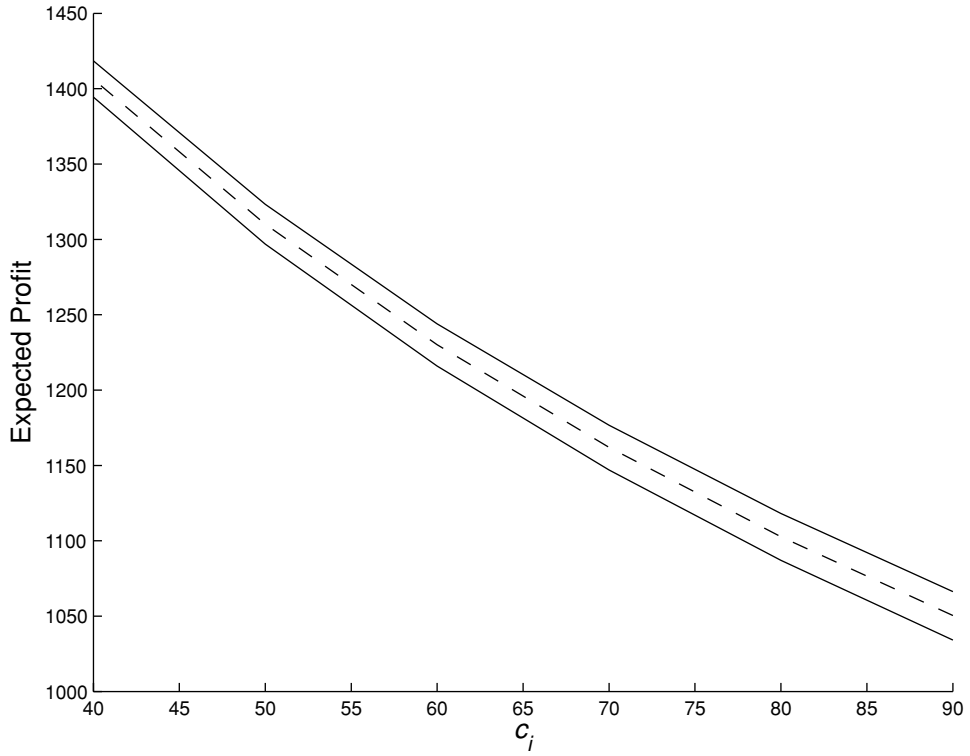

Fig. 12. Expected profit for various c_i (averaged over 5000 sequences).

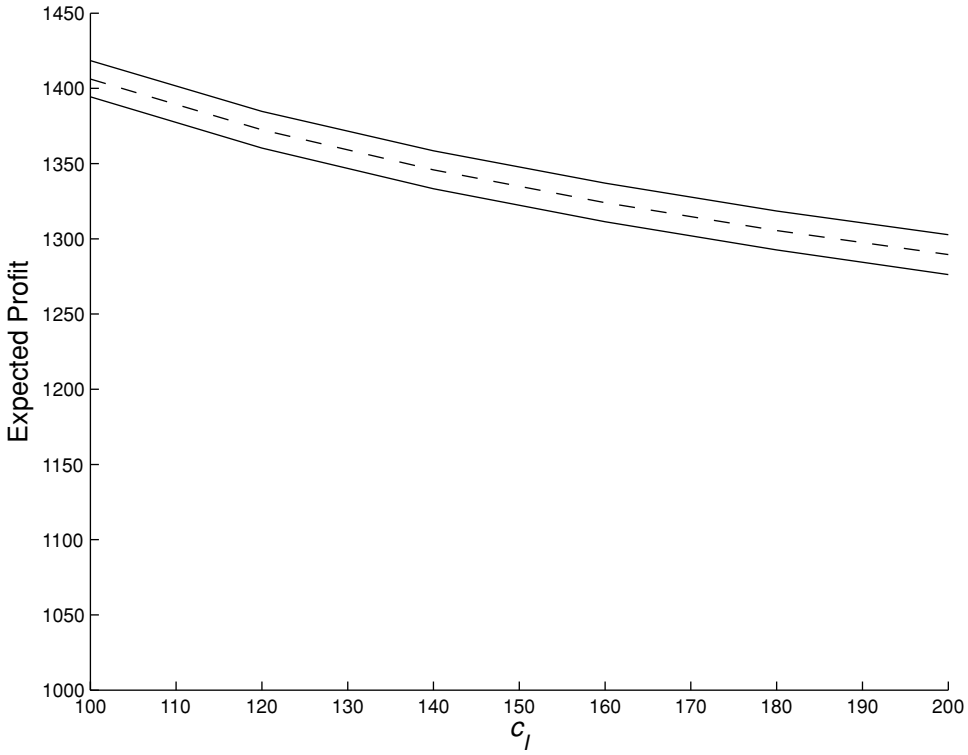

Fig. 13. Expected profit for various c_I (averaged over 5000 sequences).

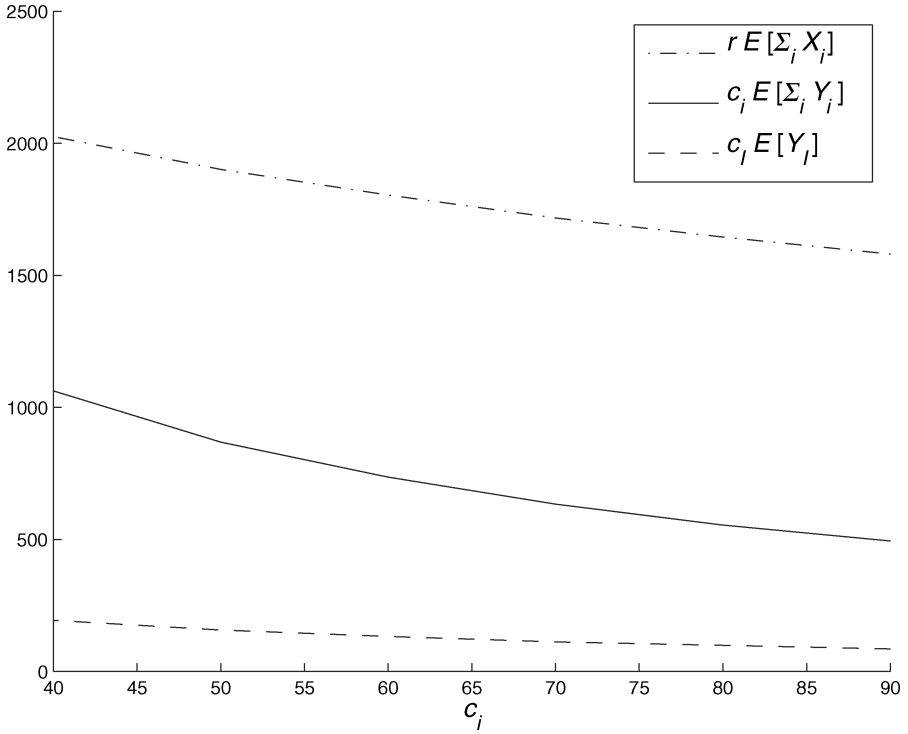

Fig. 14. Components of expected profit for various c_i (averaged over 5000 sequences).

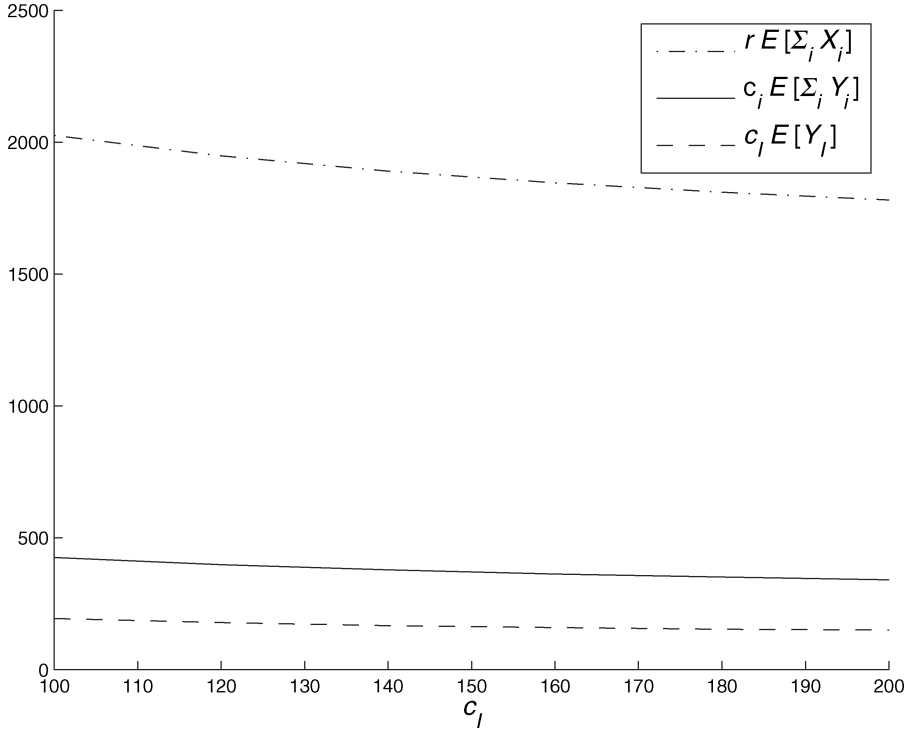

Fig. 15. Components of expected profit for various c_I (averaged over 5000 sequences).

results from these 1000 simulations. The average maximum profit for case 1 is 1310.8, case 2 is 1289.4 and case 3 is 1262.0. The decrease in profit with increasing no-show probabilities does conform with basic intuition. Also, as can be expected, the number of patients scheduled increases with increasing no-show probabilities. The respective average number of patients scheduled are 30.67, 35.58 and 42.13.

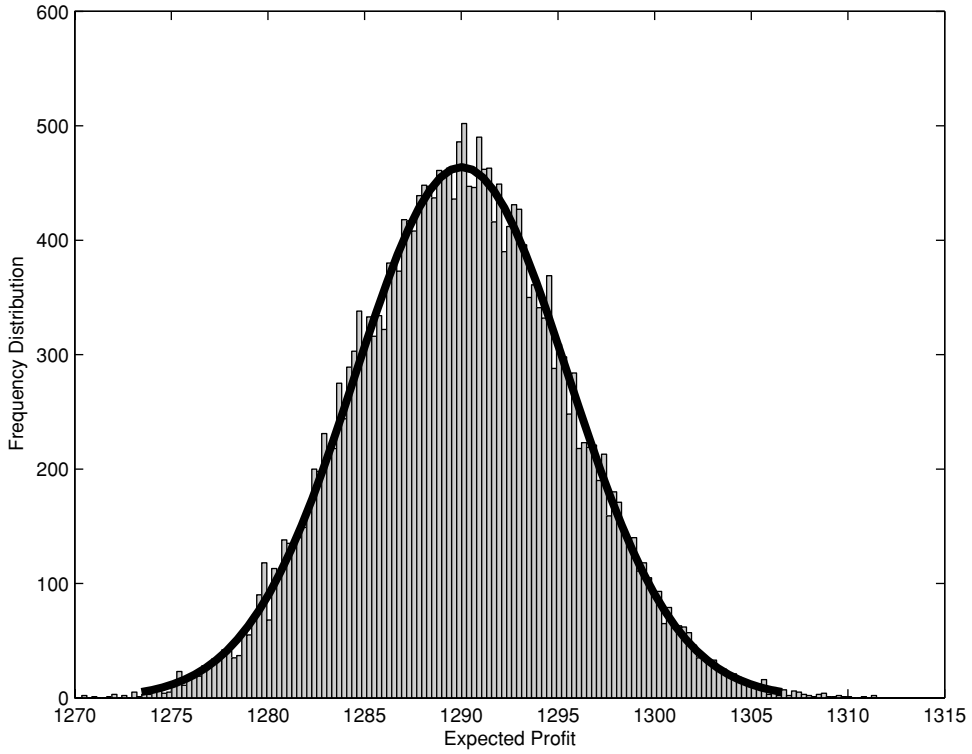

Fig. 16. Frequency histogram of expected profit for 25 000 sequences.

Next, to measure the effect of the number of types of patients, we consider three cases. The first with $J = 2$ and $p = (0.33, 0.67)$, second with $J = 3$ and $p = (0.25, 0.5, 0.75)$ and the third with $J = 4$ and $p = (0.2, 0.4, 0.6, 0.8)$. Note that the average no-show probability is always 0.5 for each of the three cases. As earlier to use a consistent set of sample paths we generate underlying sample paths with 12 types of patients. Then for case 1, types 1–6 are relabeled type 1 and the rest type 2. Similar relabeling of the underlying sample paths generate the sample paths for cases 2 and 3 as well. As the number of patient types increases so does the maximum objectives, although not significantly. The respective values are 1279.2, 1289.4 and 1295.0 for cases 1, 2 and 3. The marginal increase in objective is possibly due to the increase in flexibility made available to the decision maker by larger number of patient types.

5.4. *Effect of call-in sequence on schedule profit*

In this subsection, we experimentally examine the effect of the call-in sequence of a set of patients on the schedule generated by our booking policy. Our procedure is as follows:

Step 1. Randomly generate a set of N patients.
Step 2. Randomly select M sequences of the N patients.
Step 3. For each of the M sequences, use the booking policy to generate a schedule.
Step 4. Develop the frequency distribution of schedule profits for the M schedules.

Figure 16 illustrates this distribution for our previous example with $c_i = 40$, $c_I = 200$, $r = 100$ for 25 000 sequences of 48 patients. The maximum observed profit is \$1310, the average is \$1290 and the minimum is \$1275. Thus, we estimate that, in the worst case, the sequence effect costs us \$35 or around 2.6%, and in the average case, \$20 or around 1.5%. Furthermore, the histogram is very symmetric and has a normal appearance, and thus can be used to make approximate probability statements about daily profit, which provides some predictive capability for the clinic. For example, assuming normality, sufficient demand, and estimating μ at \$1290 and σ at 5.52, we can be approximately 95% confident that the clinic's daily profit will fall between \$1279 and \$1301.

6. Conclusions

In this work, we formulated an overbooking model and presented a myopic scheduling policy for outpatient clinics that explicitly leverages on patient no-show probability estimates. We developed an objective function that captures patient waiting time, staff overtime and patient revenue, and we derived the necessary and sufficient conditions for the expected profit evolution to be unimodal. The local maxima can then conveniently serve as a natural stopping criterion for the scheduling policy. Furthermore, we examined the behavior of the policy with respect to slot loading, changes in cost coefficients and call-in sequence effects. We believe that the work provides a significant contribution to the research literature on appointment scheduling and that it is easily implemented in practice.

The model formulated in this paper is readily extendable in many ways, often easily. First, the number of patient types need not be finite, we could assume that each patient has a different no-show probability. We take a finite set of patient types only for the convenience in presentation. Second, walk-ins can be easily added to the model. Only the estimate of $Q^n_{i,m}$ would change depending on the model describing the walk-ins.

The assumption of exponential service times are arguably restrictive for the use of the model in a variety of situations and related health care problems. Future work is needed in extending the model and the analysis in this paper for general distributions. Our future work will include some of these extensions and focus on characterizing non-myopic optimal policies and implementing the approach with our clinical partners.

Acknowledgements

We thank Purdue's Regenstrief Center for Healthcare Engineering for supporting this work. We also thank the physicians, administrators and staff of the Indiana University Medical Group and Wishard Primary Care Clinic of Indianapolis, Indiana for their interactions, comments and feedback. Finally, we thank the anonymous referees and Prof. Benneyan for their comments and suggestions.

References

Babes, M. and Sarma, G.V. (1991) Out-patient queues at the Ibn-Rochd health centre. *The Journal of the Operational Research Society*, **42**(10), 845–855.

Bean, A.G. and Talaga, J. (1995) Predicting appointment breaking. *Journal of Health Care Marketing*, **15**(1), 29–34.

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

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

Centers for Medicare and Office of the Actuary Medicaid Services. (2007) National health care expenditures projections: 2007–2017.

Chatwin, R. (1998) Multiperiod airline overbooking with a single fare class. *Operations Research*, **46**(6), 805–819.

Chatwin, R. (1999) Continuous-time airline overbooking with time-dependent fares and refunds. *Transportation Science*, **33**, 182–191.

Coughlan, J. (1999) Airline overbooking in the multi-class case. *The Journal of the Operational Research Society*, **50**(11), 1098–1103.

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

Dervin, J.V., Stone, D.L. and Beck, C.H. (1978) The no-show patient in the model family practice unit. *Journal of Family Practice*, **7**(6), 1177–1180.

Deyo, R.A. and Inui, T.S. (1980) Dropouts and broken appointments. *Medical Care*, **18**(11), 1146–1157.

Feng, Y. and Xiao, B. (2001) A dynamic airline seat inventory control model and its optimal policy. *Operations Research*, **49**(6), 938–949.

Forjuoh, S.N., Averitt, W.M., Cauthen, D.B., Couchman, G.R., Symm, B. and Mitchell, M. (2001) Open-access appointment scheduling in family practice: Comparison of a demand prediction grid with actual appointments. *Journal–American Board of Family Practice*, **14**(4), 259–265.

Goldman, L., Freidin, R., Cook, E.F., Eigner, J. and Grich, P. (1982) A multivariate approach to the prediction of no-show behavior in a primary care center. *Archives of Internal Medicine*, **142**(3), 563–567.

Harper, P.R. and Gamlin, H.M. (2003) Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach. *OR Spectrum*, **25**, 207–222.

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

Ho, C., Lau, H. and Li, J. (1995) Introducing variable-interval appointment scheduling rules in service systems. *International Journal of Operations & Production Management*, **15**(6), 59–68.

Jun, J.B., Jacobson, S.H. and Swisher, J.R. (1999) Application of discrete-event simulation in health care clinics: a survey. *The Journal of the Operational Research Society*, **50**(2), 109–123.

Karaesmen, I. and Van Ryzin, G. (2004) Overbooking with substitutable inventory classes. *Operations Research*, **52**(1), 83–104.

Kennedy, J.G. and Hsu, J.T. (2003) Implementation of an open access scheduling system in a residency training program. *Family Medicine*, **35**(9), 666–670.

Klassen, K.J. and Rohleder, T.R. (2004) Outpatient appointment scheduling with urgent clients in a dynamic multi-period environment. *International Journal of Service Industry Management*, **15**(2), 167–186.

Kopach, R., De Laurentis, P.-C., Lawley, M., Muthuraman, K., Ozsen, L., Rardin, R., Wan, H., Intrevado, P., Qu, X. and Willis, D. (2007) Effects of clinical characteristics on successful open access scheduling. *Health Care Management Science*, **10**(2), 111–124.

Lau, H. and Lau, A.H. (2000) A fast procedure for computing the total system cost of an appointment schedule for medical and kindred facilities. *IIE Transactions*, **32**(9), 833–839.

Liu, L. and Liu, X. (1998) Block appointment systems for outpatient clinics with multiple doctors. *Journal of the Operational Research Society*, **29**(12), 1254–1259.

McGill, J. and Van Ryzin, G. (1999) Revenue management: research overview and prospects. *Transportation Science*, **33**(2), 233–256.

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

Murray, M., Bodenheimer, T. and Rittenhouse, D. (2003) Improving timely access to primary care. *The Journal of the American Medical Association*, **289**, 1042–1046.

O'Hare, C. D. and Corlett, J. (2004) The outcomes of open-access scheduling. *Family Practice Management*, Feb, 35–38.

Robinson, L. (1995) Optimal and approximate control policies for airline booking with sequential nonmonotonic fare classes. *Operations Research*, **43**(2), 252–263.

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

Rohleder, T.R. and Klassen, K.J. (2002) Rolling horizon appointment scheduling: a simulation study. *Health Care Management Science*, **5**(3), 201–209.

Rothstein, M. (1985) OR and the overbooking problem. *Operations Research*, **33**(2), 237–248.

Subramanian, J., Stidham, S. and Lautenbacjer, C. (1999) Airline yield management with overbooking, cancellations, and no-shows. *Transportation Science*, **33**(2), 147–167.

Vanden Bosch, P.M. and Dietz, D.C. (2000) Minimizing expected waiting in a medical appointment system. *IIE Transactions*, **32**(9), 841–848.

Vanden Bosch, P.M. and Dietz, D.C. (2001) Scheduling and sequencing arrivals to an appointment system. *Journal of Service Research*, **4**(1), 15–25.

Vanden Bosch, P.M., Dietz, D.C. and Simeoni, J.R. (1999) Scheduling customer arrivals to a stochastic service system. *Naval Research Logistics*, **46**(5), 549–559.

Biographies

Kumar Muthuraman is currently an Assistant Professor at the McCombs School of Business at the University of Texas in Austin. He received his Ph.D in 2003 from Stanford University and has previously been an Assistant Professor at the School of Industrial Engineering at Purdue University. His primary interests are in computational methods for stochastic control problems especially in quantitative finance. His secondary interests include stochastic optimization problems in inventory management and healthcare.

Mark Lawley is an Associate Professor in the Weldon School of Biomedical Engineering at Purdue University. Before joining Biomedical Engineering in 2007, he served as an Assistant and an Associate Professor of Industrial Engineering, also at Purdue, and he 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 80 technical papers and has won three best paper awards. In January 2005, he was appointed Regenstrief Faculty Scholar in support of Purdue's Regenstrief Center for Health Care Engineering. He is particularly interested in developing optimal decision policies for system configuration and resource allocation in large healthcare systems. As a Regenstrief Scholar, he has focused on research initiatives with Wishard Hospital, the Regenstrief Institute of Indianapolis, the Richard L. Roudebush Veterans Administration Medical Center, Ascension Health, and St. Vincent Hospitals. His research has been supported by the National Science Foundation, Union Pacific Railroads, Consilium Software, General Motors, Ascension Health, the Indiana State Department of Health, the Regenstrief Foundation, the St. Vincent Ministry and many others. He received a PhD in Mechanical Engineering from the University of Illinois at Urbana Champaign in 1995.