

# A Stochastic Overbooking Model for Outpatient Clinical Scheduling with No-shows

Kumar Muthuraman, Mark Lawley

School of Industrial Engineering, Purdue University, West Lafayette, IN 47906  
kumar@purdue.edu, malawley@purdue.edu

In this paper we formulate a stochastic overbooking model and develop an appointment scheduling policy 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. We derive conditions under which the objective evolution is unimodal and we investigate the behavior of the scheduling policy under a variety of conditions and make some practical observations on its performance.

*Key words:* Open Access, Appointment Scheduling, Patient No-shows, Outpatient Clinic Operations

---

## 1. Introduction

Healthcare currently consumes 15% of the U.S. Gross Domestic Product and is expected to reach 19% within the coming decade [16]. 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 re-assess 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. Further, 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 [32]. 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 [14].

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 [17, 24, 30, 31] for representative discussions). In essence, there is no 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 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 [2, 13, 18]).

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 [36] provides an engaging review of the evolution of airline overbooking as an acceptable practice, while McGill [28] provides an informative literature review. Other representative research in airline overbooking includes [8, 9, 10, 15, 23, 33, 37].

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. Further, while airlines incur an explicit financial penalty for overshoot situations, system dynamics are not affected. However, overshoots 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 that were unserved 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 under utilized 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. For example, if waiting facilities are limited and patients tend to have companions, then the waiting costs could be set relatively high.

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

In 2003, Cayiril and Veral [6] provided an extensive review of the appointment scheduling literature, covering eighty papers that span fifty years. They categorize the appointment scheduling literature by the following attributes: (a) static vs. dynamic; (b) performance measures; (c) system design; and (d) methodology. They also provide a good discussion of future research directions. In the following, we will briefly discuss (a)-(d) 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 [6].

The first classification attribute is static vs. 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 [3, 4, 5, 12, 20, 26, 34].

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 [29].

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 [7, 19, 20, 25, 27, 35].

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 [3, 4, 5, 12, 34], while simulation studies include [1, 7, 19, 20, 21, 25, 35]. Further, [22] 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. Further, 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 [6], who say “No rigorous research exists which investigates possible approaches to adjusting the AS 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  $i = 1, 2, \dots, I$  is of length  $\Delta t_i$ . 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. We assume that service times are exponentially distributed.

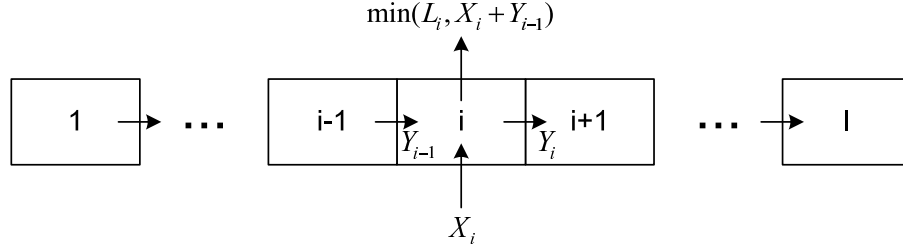
At some point during the call-in period, suppose  $n$  patients have been scheduled. Let the random variable  $X_i^n$  denote the number of patients arriving for slot  $i$  and  $Y_i^n$  be the number of patients waiting for the completion of service at the end of slot  $i$ . That is, the number of patients overflowing from slot  $i$  into slot  $i + 1$  (see Figure 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,

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

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.



**Figure 1** The System

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

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

$$Q_{im}^n = \Pr\{X_i^n = m\} \quad (2)$$

$$R_{ik}^n = \Pr\{Y_i^n = k\}. \quad (3)$$

Suppose the  $n^{th}$  patient calls for an appointment and is of type  $j$ . Letting  $U$  be the set of slots (that is, integers from 1 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  $f(Q^n, R^n)$ , that is, we assign patient  $n$  to slot  $i^*$  where

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

$$S^n = S^{n-1} + \Delta^{i^*j}. \quad (5)$$

While  $S^n$  will denote the state after an optimal assignment  $i^*$ , that is  $S^{n-1} + \Delta^{i^*j}$ , we will use  $S_i^n$  to denote the state where the last assignment is to state  $i$ , which is not necessarily the best

assignment, that is  $S_i^n = S^{n-1} + \Delta^{ij}$ . Similarly  $Q_i^n$  and  $R_i^n$  are the arrival probability matrix and the over-flow probability matrix associated with  $S_i^n$ .

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 over flow from slot  $i$  to slot  $i + 1$ . This provides sufficient flexibility to model the cost of physician and staff overtime by assigning an appropriate over flow 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

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

### 3.1. Calculating $Q^n$ and $R^n$

Consider the  $i^{th}$  row of a given  $S^n$ . We are interested in the probability that  $m$  patients arrive given  $S_{i,1}^n, S_{i,2}^n, \dots, S_{i,J}^n$ . Let  $\Pi$  be the set of all non-negative, integer  $J$ -vectors  $\pi \equiv (\pi_1, \pi_2, \dots, \pi_J)$  such that  $\sum_{j=1}^J \pi_j = m$  and  $\pi_j \leq S_{i,j}^n$  for all  $j$ . Then conditioning on the event that  $\pi_j$  number of type  $j$  patients show up,

$$\begin{aligned} Q_{i,m}^n &= \Pr\{X_i^n = m\} \\ &= \sum_{\pi \in \Pi} \Pr\{X_i^n = m | (\pi_1, \dots, \pi_J)\} \Pr\{(\pi_1, \dots, \pi_J)\} \\ &= \sum_{\pi \in \Pi} \Pr\{(\pi_1, \dots, \pi_J)\} \\ &= \sum_{\pi \in \Pi} \prod_j \frac{S_{i,j}^n!}{\pi_j! (S_{i,j}^n - \pi_j)!} p_j^{\pi_j} (1 - p_j)^{S_{i,j}^n - \pi_j}. \end{aligned} \quad (7)$$

Next consider  $R_{i,k}^n$ , that is, the probability of  $k$  patients over-flowing into slot  $i + 1$  from slot  $i$ ,

$$\begin{aligned} R_{i,k}^n &= \Pr\{Y_i = k\} \\ &= \Pr\{\max(X_i + Y_{i-1} - L_i, 0) = k\} \\ &= \begin{cases} \Pr\{X_i + Y_{i-1} - L_i = k\} & k > 0 \\ \Pr\{X_i + Y_{i-1} - L_i \leq 0\} & k = 0 \end{cases} \end{aligned} \quad (8)$$

Further conditioning yields,

$$\begin{aligned} R_{i,0} &= \sum_m \sum_{\tilde{k}} \Pr\{m + \tilde{k} \leq L_i\} Q_{i,m}^n R_{i-1,\tilde{k}}^n \\ &= \sum_m \sum_{\tilde{k}} (1 - F_{L_i}(m + \tilde{k})) Q_{i,m}^n R_{i-1,\tilde{k}}^n \end{aligned} \quad (9)$$

and similarly for  $k > 0$ ,

$$\begin{aligned} R_{i,n} &= \sum_m \sum_{\tilde{k}} \Pr\{m + \tilde{k} - k = L_i\} Q_{i,m}^n R_{i-1,\tilde{k}}^n \\ &= \sum_m \sum_{\tilde{k}} f_{L_i}(m + \tilde{k} - k) Q_{i,m}^n R_{i-1,\tilde{k}}^n \end{aligned} \quad (10)$$

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  $\frac{1}{\lambda\Delta t_i}$ ,

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

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

It is possible to relax the exponential service time distribution and replace it with general service times. All our calculations above would remain absolutely the same. However relaxing the exponential distribution in our analysis implicitly brings in an approximation. The memory-less 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 suggests exponential service time distributions [11]. 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. This extension might be algebraically tedious depending on the service time distribution assumed.

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

$$Q_{i,m}^n = \begin{cases} Q_{i,m}^{n-1}(1 - p_j) + (Q_{i,m-1}^{n-1})p_j & \text{when } m \geq 1 \text{ and} \\ Q_{i,0}^{n-1}(1 - p_j) & \text{when } m = 0. \end{cases} \quad (13)$$

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  $Q^n$  given  $S^n$ . For the incremental calculation of the overflow probability matrix  $R^n$ , note that  $R_{i,k}^n = R_{i,k}^{n-1}$  for all  $\tilde{i} < i$  and for all  $k$  and the calculation of  $R_{i,k}^n$  for  $\tilde{i} \geq i$  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. Further, the algorithm will reject the patient and terminate when there is no way to schedule the patient without hurting the objective.

1. Set  $S_{i,j} = 0$  for all  $i = 1, \dots, I$  and  $j = 1, \dots, J$   
 $Q_{i,0} = R_{i,0} = 1$  for all  $i = 1, \dots, I$ , and  $n = 1$ .
2. Wait for  $n^{\text{th}}$  call.
3.  $n^{\text{th}}$  call occurs and is of type  $j$ .
4. For each  $i \in U$ 
  - (a) Set  $S_i^n = S^{n-1} + \Delta^{ij}$ .
  - (b) Compute  $Q_i^n$  and  $R_i^n$  from  $Q^{n-1}$  and  $R^{n-1}$  using equations (7),(9) and (10).
  - (c) Compute  $f_i^n = f(Q_i^n, R_i^n)$ .
5. If  $\max f_i^n \geq f^{n-1}$ 
  - (a) Then  $i^* = \arg \max f_i^n$ ,  $S^n = S^{n-1} + \Delta^{i^*j}$ ,  $Q^n = Q_{i^*}^n$ ,  $R^n = R_{i^*}^n$ . Set  $n = n + 1$ . Goto Step 2.
  - (b) Else Stop.

## 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. Further 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 0, respectively.

First we define some notation. The event  $\mathcal{A}_n$  denotes that the  $n^{\text{th}}$  call-in patient actually shows up for the assigned slot.  $P_i^n$  will denote the conditional probability that the assignment of the  $n^{\text{th}}$  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**  $\mathbf{E}[Y_i^n]$  is non-decreasing in  $n$ . Moreover if the  $n^{\text{th}}$  patient is of type  $j$ , then  $\mathbf{E}[Y_i^n] - \mathbf{E}[Y_i^{n-1}] = p_j P_i^n$ .

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

$$\mathbf{E}[Y_i^n - Y_i^{n-1}] = p_j \mathbf{E}[Y_i^n - Y_i^{n-1} | \mathcal{A}_n] > 0. \quad (14)$$

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  $n^{\text{th}}$  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

$$\begin{aligned} \mathbf{E}[Y_i^n - Y_i^{n-1}] &= p_j \mathbf{E}[Y_i^n - Y_i^{n-1} | \mathcal{A}_n] \\ &= p_j \sum_{y=0,1} y \Pr\{Y_i^n - Y_i^{n-1} = y | \mathcal{A}_n\} \\ &= p_j \Pr\{Y_i^n - Y_i^{n-1} = 1 | \mathcal{A}_n\} \\ &= p_j P_i^n \end{aligned} \quad (15)$$

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

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

$$\mathbf{E} \left[ r \sum_{i=1}^I X_i^n - \sum_{i=1}^I c_i Y_i^n \right] < \mathbf{E} \left[ r \sum_{i=1}^I X_i^{n-1} - \sum_{i=1}^I c_i Y_i^{n-1} \right]. \quad (16)$$

Rearranging to have rewards on the LHS and costs on the RHS,

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

The expectation on the LHS simply denotes the probability of the  $n^{\text{th}}$  patient showing up and is hence  $p_j$ . Hence, we can write,

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

Hence,

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

Now denote the slot to which the  $n$ -patient was assigned to be  $i_n$ . 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,

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

Now since the assignment of slot  $i_n$  to the  $n^{\text{th}}$  patient is made by equation (4), we have due to the fact that the chosen slot yielded the best objective,

$$\begin{aligned} f(Q^n, R^n) &\geq f(Q_{i_n}^n, R_{i_n}^n) \\ &= f(Q^{n-1} + \Delta^{i_n j}, R^{n-1} + \Delta^{i_n j}) \quad \forall i_n \in [1, \dots, I]. \end{aligned} \quad (21)$$

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

$$f(Q^n, R^n) - f(Q^{n-1}, R^{n-1}) \geq f(Q_{i_n}^n, R_{i_n}^n) - f(Q^{n-1}, R^{n-1}) \quad \forall i_n \in [1, \dots, I].$$

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

$$p_j r - p_j \sum_{i=1}^I c_i \prod_{\tilde{i}=i_n}^i \Pr\{L_{\tilde{i}} < X_{\tilde{i}}^n + Y_{\tilde{i}-1}^n | \mathcal{A}_n\} \geq p_j r - p_j \sum_{i=1}^I c_i \prod_{\tilde{i}=i_n}^i \Pr\{L_{\tilde{i}} < X_{\tilde{i}}^n + Y_{\tilde{i}-1}^n | \mathcal{A}_n\} \quad \forall i_n \in [1, \dots, I].$$

That is,

$$\sum_{i=1}^I c_i \prod_{\tilde{i}=i_n}^i \Pr\{L_{\tilde{i}} < X_{\tilde{i}}^n + Y_{\tilde{i}-1}^n | \mathcal{A}_n\} \leq \sum_{i=1}^I c_i \prod_{\tilde{i}=i_n}^i \Pr\{L_{\tilde{i}} < X_{\tilde{i}}^n + Y_{\tilde{i}-1}^n | \mathcal{A}_n\} \quad \forall i_n \in [1, \dots, I]. \quad (22)$$

Now consider  $P_i^m$  for  $m > n$  and say the  $m^{\text{th}}$  patient is assigned to slot  $i_m$ . Since  $m > n$ , we have for any realization,  $X_i^m \geq X_i^n$  and  $Y_i^m \geq Y_i^n$ , therefore,

$$\begin{aligned} \sum_{i=1}^I c_i P_i^m &= \sum_{i=1}^I c_i \prod_{\tilde{i}=i_m}^i \Pr\{L_{\tilde{i}} < X_{\tilde{i}}^m + Y_{\tilde{i}-1}^m | \mathcal{A}_n\} \\ &\geq \sum_{i=1}^I c_i \prod_{\tilde{i}=i_m}^i \Pr\{L_{\tilde{i}} < X_{\tilde{i}}^n + Y_{\tilde{i}-1}^n | \mathcal{A}_n\} \\ &\geq \sum_{i=1}^I c_i \prod_{\tilde{i}=i_n}^i \Pr\{L_{\tilde{i}} < X_{\tilde{i}}^n + Y_{\tilde{i}-1}^n | \mathcal{A}_n\} \quad (\text{from equation (22)}) \\ &= \sum_{i=1}^I c_i P_i^n. \end{aligned} \quad (23)$$

From equation (19),

$$r < \sum_{i=1}^I c_i P_i^n \leq \sum_{i=1}^I c_i P_i^m. \quad (24)$$

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

$$r p_{\tilde{j}} < p_{\tilde{j}} \sum_{i=1}^I c_i P_i^m. \quad (25)$$

Which implies as earlier,

$$r \mathbf{E} \left[ \sum_{i=1}^I (X_i^m - X_i^{m-1}) \right] < \mathbf{E} \left[ \sum_{i=1}^I c_i (Y_i^m - Y_i^{m-1}) \right] \quad (26)$$

or equivalently,

$$f(Q^m, R^m) < f(Q^{m-1}, R^{m-1}). \quad (27)$$

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

*Proof:* Follows directly from Theorem 1.

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

*Proof:*

$$f(Q^n, R^n) - f(Q^{n-1}, R^{n-1}) = p_j \left( r - \sum_{i=1}^I c_i P_i^n \right) \quad (28)$$

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

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

Which implies,

$$\begin{aligned} r &< \sum_{i=1}^I c_i P_i^m \\ &= c_I. \end{aligned} \quad (29)$$

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

$$\begin{aligned} r &> \lim_{n \rightarrow \infty} \sum_{i=1}^I c_i P_i^n \\ &= c_I, \end{aligned} \quad (30)$$

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

**Proposition 3** A sufficient condition for  $f(Q^1, R^1) > f(Q^0, R^0)$  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  $f(Q^0, R^0) = 0$ , since we have no scheduled patients. Hence we are interested in establishing the necessary and sufficient conditions for  $f(Q^1, R^1) > 0$ . First we show that if  $r > \sum_{i=i_n}^I c_i e^{-\lambda \Delta t_i (i-i_n+1)}$  for all  $i_n$  then  $f(Q^1, R^1) > 0$

If the first arriving patient is of type  $j$ ,

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

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  $i$ ,

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

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

*Proof:* Next we show that if  $f(Q^1, R^1) > 0$  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,

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

Hence

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

## 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.

In all the 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,  $J = 3$  with no show probabilities for each type given by  $p = (0.25, 0.5, 0.75)$ . 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 sequel, we will always consider cases with  $c_I > c_i$  when  $i < I$  and take  $c_i$  to be a constant for all  $i = 1, \dots, I - 1$ . For notational convenience,  $c_i$  will denote  $c_i$  for all  $i = 1, \dots, I - 1$ . 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

Figure 2 illustrates the evolution of our profit objective for an example with  $c_i = 40$  and  $c_I = 200$ . Note that 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 24.48, the second to slot 4 with profit 48.95, and so forth. Two profit curves are displayed. The solid gives the profit associated with the schedule constructed by our booking policy while the dashed gives the profit of a schedule constructed by a round robin approach that assigns the  $i^{th}$  customer to slot  $((i - 1) \bmod 8) + 1$ . This round robin approach, being simple and easy to implement, is often roughly followed by practitioners. The right ordinate also gives the final assignment of patients to slots at optimal assignment. For example, slot 1 has (2, 2, 2) indicating that there are two patients of each type assigned to 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  $\sum_i \mathbf{E}[Y_i]/n$  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

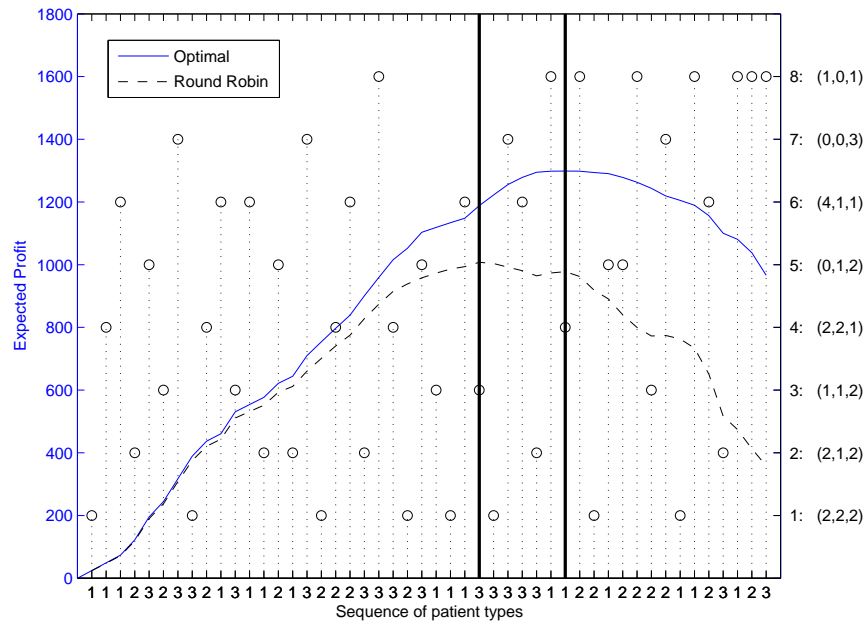


Figure 2 The schedule and expected profit evolution

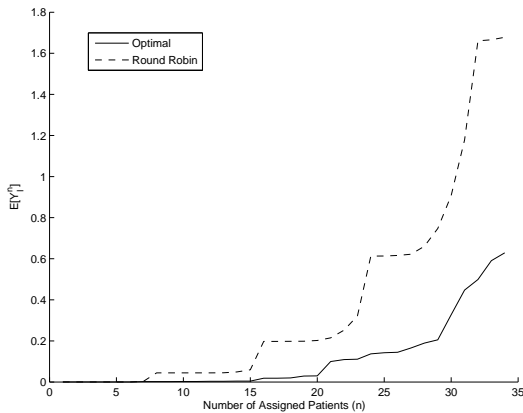


Figure 3 Expected overflow from slot I

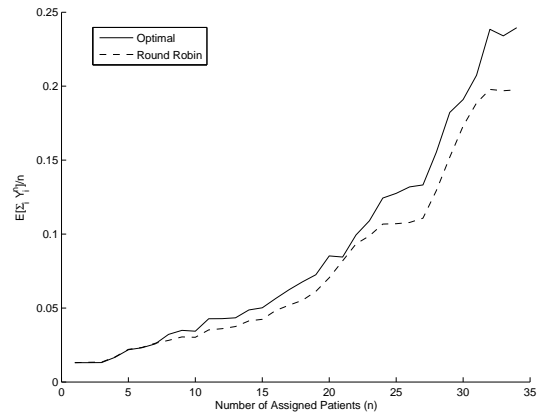
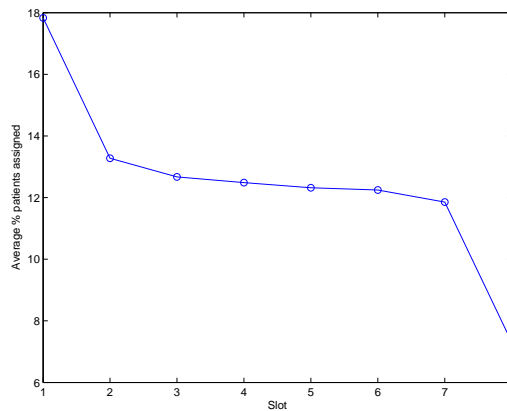


Figure 4 Expected number of slot overflows per patient

constructed by our booking policy, while the dashed presents the overflow of the round robin approach. Note that these curves terminate at the optimal assignment, 34.

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 is not true, which complicates the selection of a stopping criterion. For example, the round robin approach exhibits local maxima at patients 28 and 34, and thus it would be not clear how many patients

need to be scheduled to get the maximum profit. Further, our global maximum occurs at patient 34 with a profit of \$1298.50, while the round robin approach yields a maximum profit of \$1007.70 with 28 patients scheduled (approx. 30% difference). From Figure 3, we see that the overflow from period eight is significantly higher for the round robin approach, while the average overflow from the other slots (Figure 4) is approximately the same for the two approaches. This reflects the fact that our booking policy responds to the more severe overtime cost, while the round robin approach does not.



**Figure 5** Percentage assignments per slot

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 14624371463625..., and we see that, in this case, consecutively assigned slots tend to be at least two slots apart. This results from the policy's attempt to reduce overflow between slots and is a function of the overflow costs and service rates. Further, we see that the number of patients assigned in the optimal schedule to slots 1-8 is 65453632, respectively. Note that the last two slots have significantly fewer assignments than the others, and that the first has the maximum assignment (along with slot 6). On average, for the cost structure of 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 Figure 5. Note that Figure 5 gives the average percent of patients assigned to each slot for our example (based on 25,000 call-in sequences). 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 Figure 2, we continued assigning patients after reaching optimal 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 beyond the global optimal. After the global optimal 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, over half of the fifteen additional patients go to the last three slots, with six 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.

## 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. Figure 6 plots the percentage assignment to each slot with  $c_I = 100$ ,  $r = 100$ , and varying  $c_i$ , while Figure 7 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.

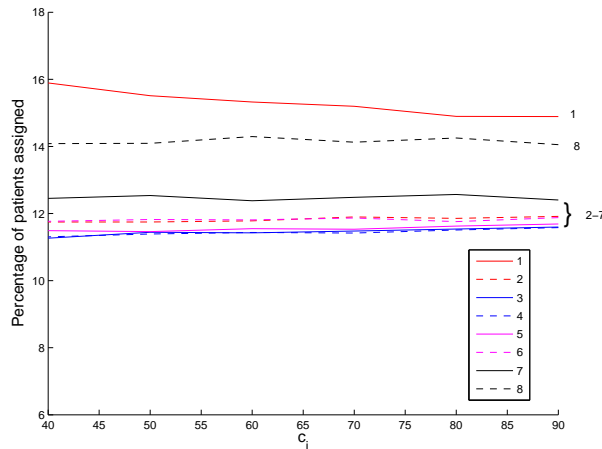


Figure 6 Percentage assignments per slot for various  $c_i$  (averaged over 5000 sequences)

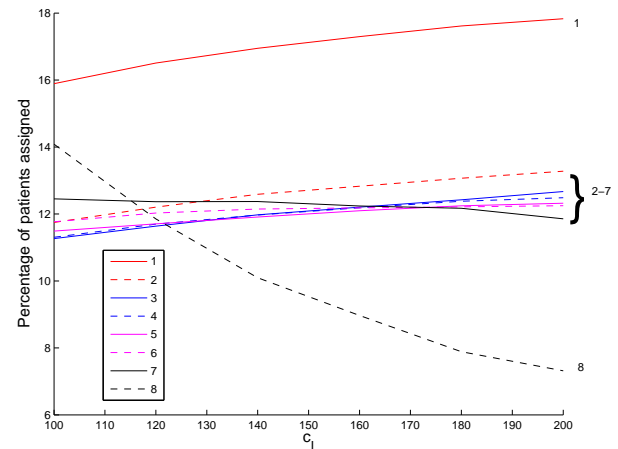
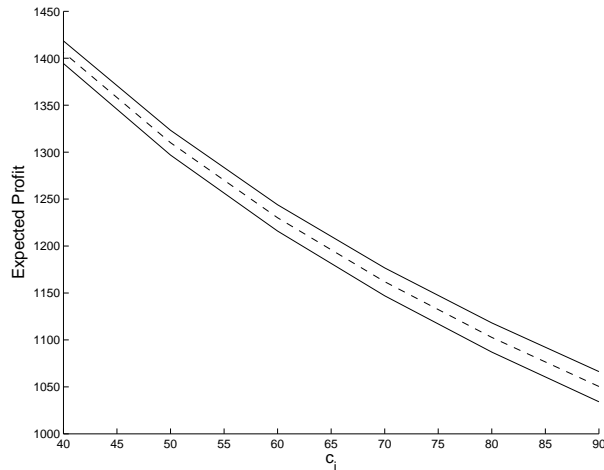
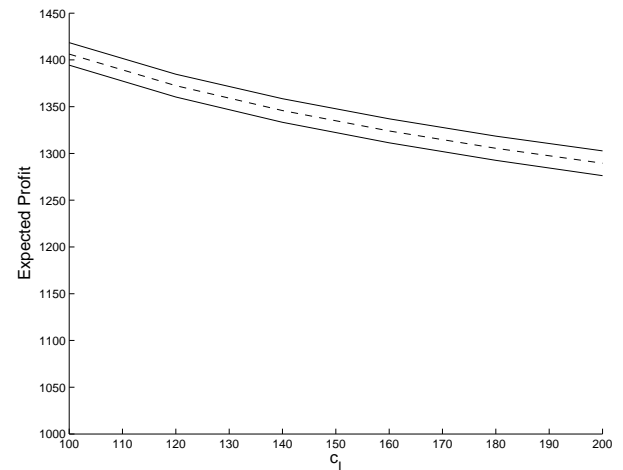


Figure 7 Percentage assignments per slot for various  $c_I$  (averaged over 5000 sequences)



**Figure 8** Expected profit for various  $c_i$  (averaged over 5000 sequences)



**Figure 9** Expected profit for various  $c_I$  (averaged over 5000 sequences)

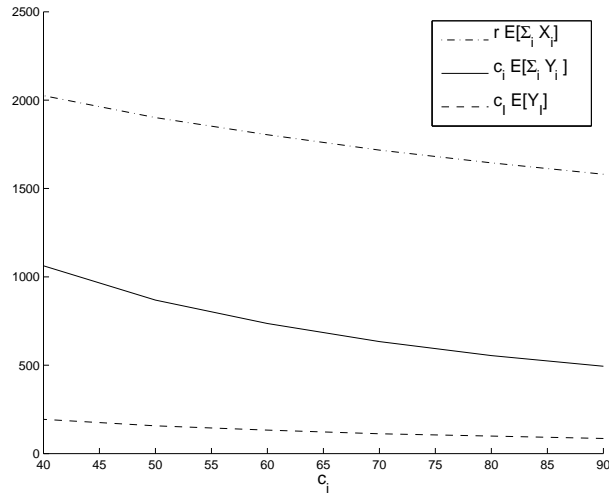
Figures 8 and 9 plot the expected profits with increase in the cost coefficients. While the dotted line plots the average expected profit over the 5000 sequences, the solid lines plot the 10<sup>th</sup> and the 90<sup>th</sup> percentiles. In this case, unlike in Figures 6 and 7, the sensitivity to  $c_i$  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 inclined towards assignment to slots 1 – 7, at which point patients not assigned to slots 1 – 7 have to go to 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 10 and 11 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. Effect of Call-in Sequence on Schedule Profit

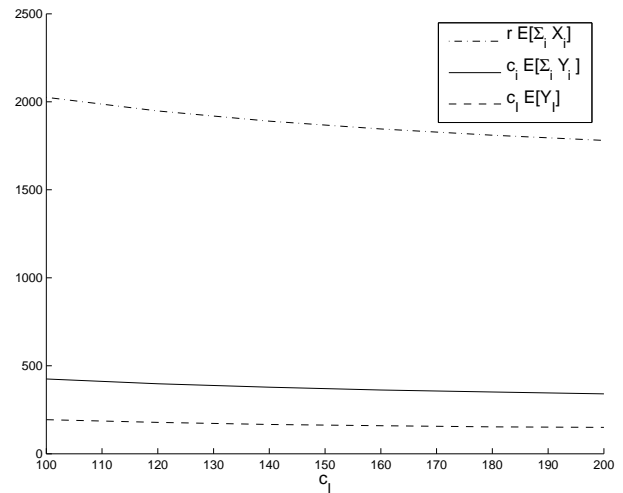
In this section, we experimentally examine the effect of the call-in sequence of a set of patients on the optimal schedule generated by our booking policy. Our procedure is as follows:

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

Figure 12 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 \$1,310, the average is \$1,290, and



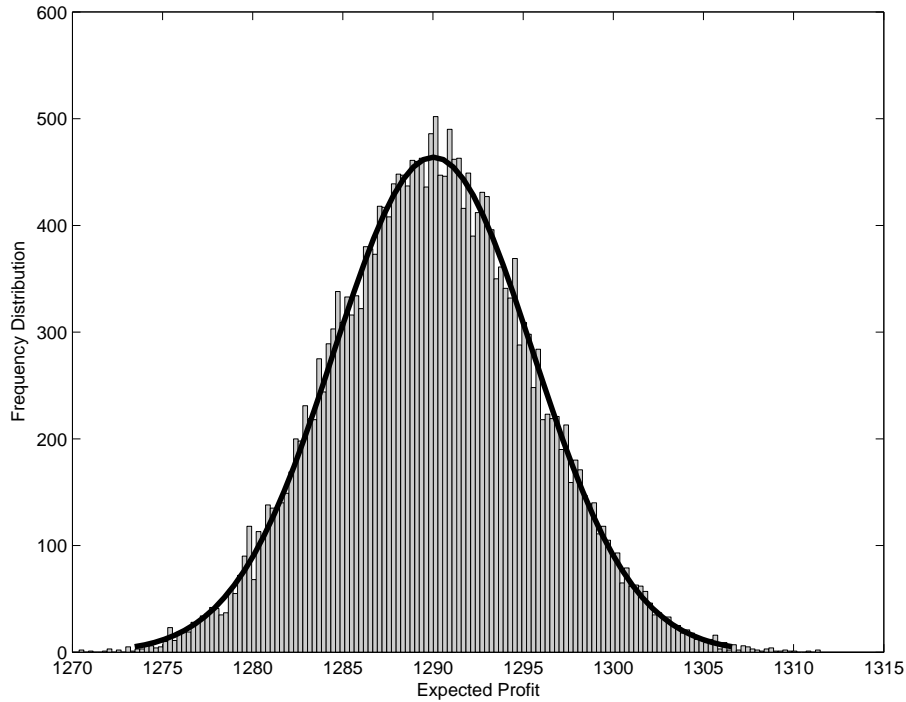
**Figure 10** Components of expected profit for various  $c_i$  (averaged over 5000 sequences)



**Figure 11** Components of expected profit for various  $c_I$  (averaged over 5000 sequences)

the minimum is \$1,275. 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%. Further, 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  $\mu$  at \$1,290 and  $\sigma$  at 5.52, we can be approximately 95% confident that the clinic's daily profit will fall between \$1,279 and \$1,301.

Since our policy is myopic, a relevant question is how well it compares to an optimal policy. Since an optimal call-in policy has not been characterized, we cannot provide rigorous comparisons but provide some discussion. First, consider the case where the call-in sequence is perfectly known. Then the problem is very close to the traditional static problem of optimally scheduling  $N$  patients (see section 2), but extended to include no-show probabilities. In such a case, the optimal mix of patient no-show probabilities is achieved for each slot in order to attain the maximum profit, say  $P$ , for that set of patients. Clearly,  $P$  is an upper bound for our myopic policy, which cannot achieve a better mix of no-show probabilities since it is constrained by the call-in sequence. Now the essential question is whether there exists a call-in sequence for the  $N$  patients for which myopic policy achieves  $P$ . If so, then the myopic policy can only achieve  $P$  when presented with the right sequence. In this case, the low sensitivity of expected profit to the sequence of arrivals (Figure 12) indicates that our myopic policy provides solutions within a few percent of  $P$ . It is also true that, for the given set of  $N$  patients, an optimal sequential policy that uses distributional knowledge of the call-in sequence cannot achieve a profit exceeding  $P$ . Since  $P$  is an upper-bound for both the



**Figure 12** Frequency histogram of expected profit for 25,000 sequences

myopic and the optimal sequential policy, the myopic policy also provides solutions within a few percent of the optimal sequential.

## 6. Conclusion

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 over- time, 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. Further, 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  $Q_{i,m}^n$  would change depending on the model describing the walk-ins. And finally, the restriction of exponential service time can be eliminated by including another state variable that records the amount of time the patient

being serviced has spent in servicing and conditioning all our expectations on this variable. This extension might be algebraically tedious depending on the service time distribution assumed. 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.

## 7. Acknowledgments

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.

## References

- [1] M. Babes and G. V. Sarma. Out-patient queues at the Ibn-Rochd health centre. *The Journal of the Operational Research Society*, 42(10):845–855, 1991.
- [2] A.G. Bean and J. Talaga. Predicting appointment breaking. *Journal of Health Care Marketing*, 15(1):29–34, 1995.
- [3] P. M. Vanden Bosch and D. C. Dietz. Minimizing expected waiting in a medical appointment system. *IIE Transactions*, 32(9):841–848, 2000.
- [4] P. M. Vanden Bosch and D. C. Dietz. Scheduling and sequencing arrivals to an appointment system. *Journal of Service Research*, 4(1):15–25, 2001.
- [5] P. M. Vanden Bosch, D. C. Dietz, and J. R. Simeoni. Scheduling customer arrivals to a stochastic service system. *Naval Research Logistics*, 46(5):549–559, 1999.
- [6] T. Cayirli and E. Veral. Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, 12(4):519–549, 2003.
- [7] T. Cayirli, E. Veral, and H. Rosen. Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science*, 9:47–58, 2006.
- [8] R. Chatwin. Multiperiod airline overbooking with a single fare class. *Operations Research*, 46(6):805–819, 1998.
- [9] R. Chatwin. Continuous-time airline overbooking with time-dependent fares and refunds. *Transportation Science*, 33:182–191, 1999.
- [10] J. Coughlan. Airline overbooking in the multi-class case. *The Journal of the Operational Research Society*, 50(11):1098–1103, 1999.
- [11] P-C. DeLaurentis, R. Kopach, M. Lawley, K. Muthuraman, L. Ozsen, X. Qu, R. Rardin, and H. Wan. A configurable framework for open access scheduling with continuous improvement in outpatient clinics. *Working paper*, 2006.

- [12] B. Denton and D. Gupta. A sequential bounding approach for optimal appointment scheduling. *IIE Transactions*, 35(11):1003–1016, 2003.
- [13] J. V. Dervin, D.L. Stone, and C. H. Beck. The no-show patient in the model family practice unit. *Journal of Family Practice*, 7(6):1177–1180, 1978.
- [14] R. A. Deyo and T. S. Inui. Dropouts and broken appointments. *Medical Care*, 18(11):1146–1157, 1980.
- [15] Y. Feng and B. Xiao. A dynamic airline seat inventory control model and its optimal policy. *Operations Research*, 49(6):938–949, 2001.
- [16] Centers for Medicare and Office of the Actuary Medicaid Services. National health care expenditures projections: 2005-2015.
- [17] Samuel N. Forjuoh, William M. Averitt, Don B. Cauthen, Glen R. Couchman, Barbalee Symm, and Mike Mitchell. 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, 2001.
- [18] L. Goldman, R. Freidin, E. F. Cook, J. Eigner, and P. Grich. A multivariate approach to the prediction of no-show behavior in a primary care center. *Archives of Internal Medicine*, 142(3):563–567, 1982.
- [19] P. R. Harper and H. M. Gamlin. Reduced outpatient waiting times with improved appointment scheduling: A simulation modelling approach. *OR Spectrum*, 25:207–222, 2003.
- [20] C. Ho and H. Lau. Minimizing total cost in scheduling outpatient appointments. *Management Science*, 38(12):1750–1764, 1992.
- [21] C. Ho, H. Lau, and J. Li. Introducing variable-interval appointment scheduling rules in service systems. *International Journal of Operations & Production Management*, 15(6):59–68, 1995.
- [22] J. B. Jun, S. H. Jacobson, and J. R. Swisher. Application of discrete-event simulation in health care clinics: a survey. *The Journal of the Operational Research Society*, 50(2):109–123, 1999.
- [23] I. Karaesmen and G. Van Ryzin. Overbooking with substitutable inventory classes. *Operations Research*, 52(1):83–104, 2004.
- [24] J. G. Kennedy and J. T. Hsu. Implementation of an open access scheduling system in a residency training program. *Family Medicine*, 35(9):666–670, 2003.
- [25] K. J. Klassen and T. R. Rohleder. Outpatient appointment scheduling with urgent clients in a dynamic multi-period environment. *International Journal of Service Industry Management*, 15(2):167–186, 2004.

- 
- [26] H. Lau and A. H. Lau. A fast procedure for computing the total system cost of an appointment schedule for medical and kindred facilities. *IIE Transactions*, 32(9):833–839, 2000.
- [27] L. Liu and X. Liu. Block appointment systems for outpatient clinics with multiple doctors. *Journal of the Operational Research Society*, 29(12):1254–1259, 1998.
- [28] J. McGill and G. Van Ryzin. Revenue management: research overview and prospects. *Transportation Science*, 33(2):233–256, 1999.
- [29] S. V. Mondschein and G. Y. Weintraub. Appointment policies in service operations: A critical analysis of the economic framework. *Production and Operations Management*, 12(2):266–286, 2003.
- [30] M. Murray, T. Bodenheimer, and D. Rittenhouse. Improving timely access to primary care. *The Journal of the American Medical Association*, 289:1042–1046, 2003.
- [31] C. D. O’Hare and J. Corlett. The outcomes of open-access scheduling. *Family Practice Management*, Feb:35–38, 2004.
- [32] Proctor P. Reid, W. Dale Compton, Jerome H. Grossman, and Gary Fanjiang, editors. *Building a Better Delivery System: A New Engineering/Health Care Partnership*. National Academies Press, 2005.
- [33] L. Robinson. Optimal and approximate control policies for airline booking with sequential nonmonotonic fare classes. *Operations Research*, 43(2):252–263, 1995.
- [34] L. W. Robinson and R. R. Chen. Scheduling doctors’ appointments: optimal and empirically-based heuristic policies. *IIE Transactions*, 35(3):295–307, 2003.
- [35] T. R. Rohleder and K. J. Klassen. Rolling horizon appointment scheduling: A simulation study. *Health Care Management Science*, 5(3):201–209, 2002.
- [36] M. Rothstein. OR and the overbooking problem. *Operations Research*, 33(2):237–248, 1985.
- [37] J. Subramanian, S. Stidham, and C. Lautenbacjer. Airline yield management with overbooking, cancellations, and no-shows. *Transportation Science*, 33(2):147–167, 1999.