Abstract No. 0070482

A Stochastic Overbooking Model for Outpatient Clinical Scheduling with No-Shows

Kumar Muthuraman, Mark Lawley
School of Industrial Engineering, School of Biomedical Engineering, Purdue University, West Lafayette, IN 47906
kumar@purdue.edu, (765) 494-5416, malawley@purdue.edu, (765) 494-5415

POMS 18th Annual Conference Dallas, Texas, U.S.A., May 4 to May 7, 2007

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. 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 investigate the behavior of the scheduling policy under a variety of conditions.

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 [13]. 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 a myriad of complex reimbursement structures. Faced with this environment, many hospitals are emphasizing shorter lengths of stay and shifting care from inpatient to outpatient facilities. This is forcing outpatient clinics to re-assess their operations, with the dual objectives of stabilizing revenue streams and improving healthcare access.

Access to outpatient facilities is controlled through appointment scheduling, where patients seeking care call the clinic to book a future appointment. In clinics where patient no-show is a significant problem (some clinics experience up to 40% patient no-show [11, 22]), overbooking can improve patient access and stabilize operations. Although overbooking has has been used by the airlines for many years [6, 7, 8, 12, 17, 23, 26], clinical booking has little in common with the airline problem [21]. Thus, we develop a new overbooking approach that accommodates the dynamics of clinical scheduling and leverages on no-show prediction.

Clinical scheduling is a problem 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 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.

The objectives are to minimize patient wait times, maximize resource utilization, and minimize staff overtime (patients waiting at the end of the day have to be served during overtime). Because of no-shows, clinic capacity will usually be under utilized without some overbooking, which incurs the risk of overloading. Thus, an optimal policy must balance the risks of patient waiting, staff overtime, and clinic under-utilization. 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.

The contributions of this research are as follows. First, it formulates a model of the call-in scheduling problem and develops a sequential policy for scheduling call-ins (Section 3). Next, it presents conditions for the objective evolution to be unimodal (Section 4), that is, the objective is non-decreasing up to a particular call-in patient and then monotone decreasing thereafter, which guarantees an optimal stopping criterion. Finally, the paper develops several insights into the practical characteristics of the policy (Section 5).

2. Literature Review

Cayiril and Veral [4] provide 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) system design; (c) performance measures; and (d) methodology. In the following, we briefly discuss (a) and (b) and refer the reader to [4] or to our working paper [21] for more detailed discussions. We note that [16] provides a review of simulation studies in health care clinics up to 1999.

The first classification 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 [1, 2, 3, 10, 15, 19, 24].

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 [5, 14, 15, 18, 20, 25].

Our work can be classified as static, since we do not adjust future scheduled appointment times as patients arrive. However, 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. 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.

We close this section by quoting Cayiril and Veral [4], who say "No rigorous research exists which investigates possible approaches to adjusting the AS (appointment 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. 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 this analysis, we assume service times are exponential (our work with several clinics supports this assumption [9].).

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,

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

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.

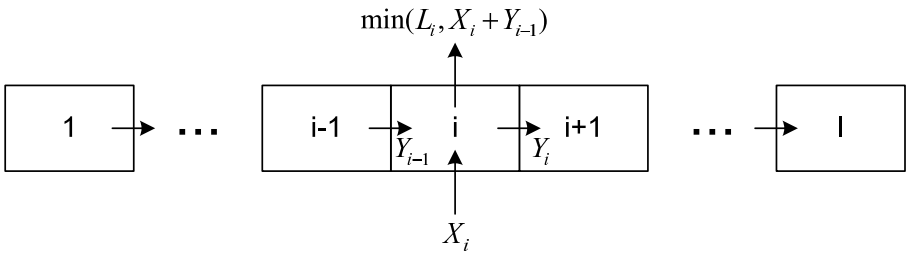

Figure 1 The System

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 $N_i^n = \sum_j S_{ij}^n$. 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 Δ^{ij} is of size $I \times J$ with a 1 at the i, j^{th} position and zeros elsewhere. The function $Q(.)$ 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,

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

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

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

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

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

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

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

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 $S^n_{i,1}, S^n_{i,2}, \ldots, S^n_{i,J}$. Let Π be the set of all non-negative, integer J-vectors $\pi \equiv (\pi_1, \pi_2, \ldots, \pi_J)$ such that $\sum_{j=1}^{J} \pi_j = m$ and $\pi_j \leq S^n_{ij}$ for all j. Then conditioning on the event that π_j number of type j patients show up,

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

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

Further conditioning yields,

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

and similarly for $k > 0$,

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

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

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

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

The memory-less property of the exponential distribution allows us to ignore the amount of time the patient in service at the beginning of a slot has already spent in service. Under a general distribution our model provides an approximation whose quality depends on the service time distribution used.

Equations (7),(9) and (10) enable the calculation of Q^n and R^n for a given S^n. In [21], we give expressions for quickly updating Q^{n+1} and R^{n+1} using Q^n and R^n.

3.2. The Scheduling Policy

The scheduling policy is described below as an algorithm. 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 $i = 1, \ldots, I$ and $j = 1, \ldots, J$
   $Q_{i,0} = R_{i,0} = 1$ for all $i = 1, \ldots, I$, and $n = 1$.
2. Wait for n^{th} call.
3. n^{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 presents results establishing 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 1 establishes that $r < C_I$ is both a necessary and sufficient condition for n being finite. Propositions 2 and 3 establish the sufficient and

necessary conditions for n being greater than 0, respectively. For brevity, we do not present the proofs but refer the reader to [21] for details.

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

Theorem 1 establishes that once the schedule reaches a point where the objective function begins to decrease due to excessive overflow, it will continue to decrease with every additional patient that is added.

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

This corollary establishes that if the objective function increases with the addition of the n^{th} patient, then none of the first n patients has resulted in a decreased objective.

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

Proposition 1 states that the objective will peak at some point and begin to go down if and only if the revenue generated by each additional patient is less than the cost of overtime. This makes intuitive sense, since once the schedule reaches some "point of saturation" every additionally assigned patient can be expected to overflow at the end of the day and generate overtime cost.

Proposition 2 *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.*

Proposition 2 provides a sufficient condition for the objective function to increase with the first assigned patient. It essentially states that the first patient will cause an increase in objective value if the revenue for the patient exceeds the maximum possible expected overflow cost that can result from assigning the patient to a slot.

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

Proposition 3 provides a necessary condition for the objective function to increase with the first assigned patient. It essentially states that for the first patient to cause an increase in objective value, the revenue for the patient must be greater than the minimum expected overflow cost that can result from assigning the patient to a slot.

5. Results and Insights

This section discusses some insights into various aspects of our scheduling policy. Using examples, we 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 also investigate the "sequence" effect by generating schedules for the same set of patient call-ins, sequenced in many different ways. Our working paper [21] contains additional experiments that examine the effects of overflow cost coefficients on slot assignments and expected profits.

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, we will always consider cases with $c_I > c_i$ when $i < I$ and take c_i to be a constant for all $i = 1, \ldots, I-1$. 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

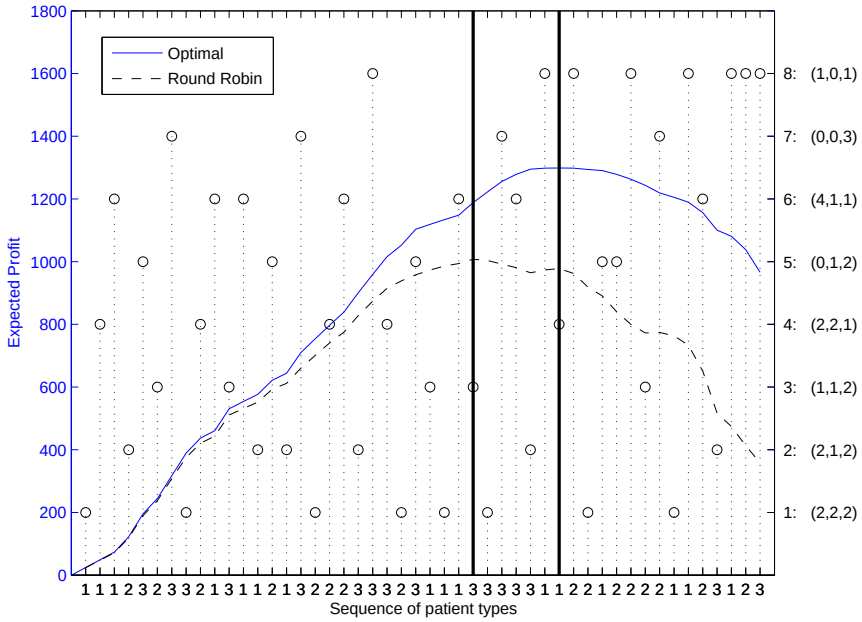

Figure 2 **The schedule and expected profit evolution**

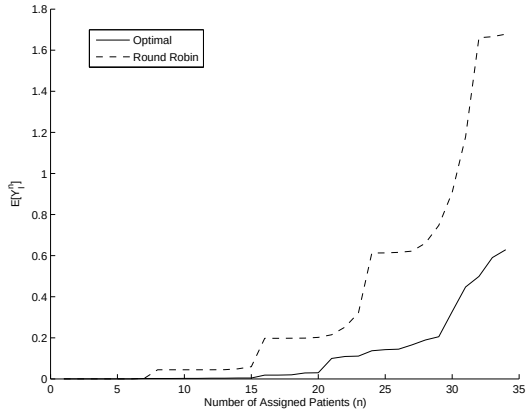

Figure 3 **Expected overflow from slot I**

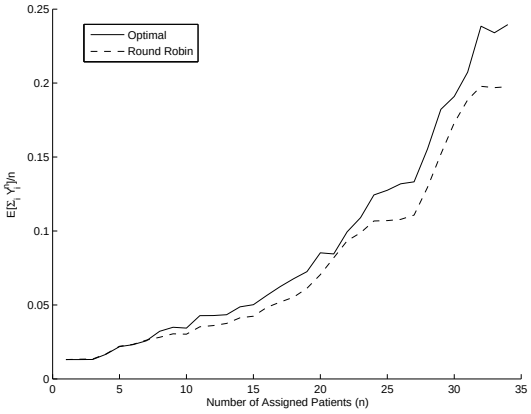

Figure 4 **Expected number of slot overflows per patient**

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.

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. 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 5 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 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 μ at $1,290 and σ at 5.52, we can be approximately 95% confident that the clinic's daily profit will fall between $1,279 and $1,301.

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

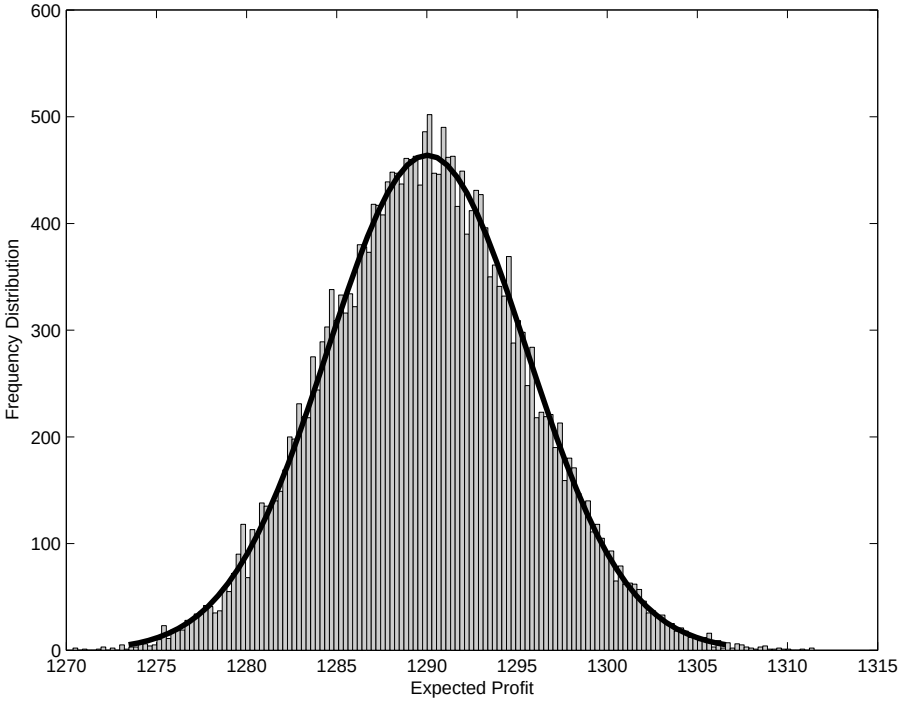

Figure 5 Frequency histogram of expected profit for 25,000 sequences

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 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. 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. 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 the their interactions, comments, and feedback.

References

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

[2] 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.

[3] 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.

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

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

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

[7] R. Chatwin. Continuous-time airline overbooking with time-dependent fares and refunds. *Transportation Science*, 33:182–191, 1999.

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

[9] 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.

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

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

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

[13] Centers for Medicare and Office of the Actuary Medicaid Services. National health care expenditures projections: 2005-2015.

[14] 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.

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

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] K. Muthuraman and M. Lawley. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *Working Paper*, 2006.

[22] 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.

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

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

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

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