Sequential Clinical Scheduling with Patient No-shows and General Service Time Distributions

Santanu Chakraborty
School Industrial Engineering, Purdue University, West Lafayette, IN
santanu@purdue.edu

Kumar Muthuraman
McCombs School of Business, University of Texas, Austin, TX
kumar.muthuraman@mccombs.utexas.edu

Mark Lawley
Weldon School of BioMedical Engineering, Purdue University, West Lafayette, IN
malawley@purdue.edu

We develop a sequential clinical scheduling method for patients with general service time distributions. Patients are scheduled during a call-in period that precedes the scheduled consultation session. The session is divided into a set of time intervals called “slots”. When a patient calls, the scheduler assigns the patient to a suitable slot if one exists. Each scheduled patient has a probability of not showing up for the appointment, and the scheduler can compensate through overbooking. We show that the proposed scheduling policy always yields a unimodal objective evolution, which provides an optimal stopping criteria for scheduling and is thus critical. We further analytically explore the special case in which service times have a gamma distribution and show how the computational complexity of the scheduling algorithm can be significantly reduced. Finally, we provide exhaustive computational results accompanied by discussion providing insights into the practical aspects of the scheduling approach.

Key words: Overbooking, Appointment Scheduling, Patient No-shows, Outpatient Clinics

1. Introduction

There is an urgent need to provide more efficient and effective patient care. From 80-90% of all US patient care is provided in approximately 200,000 non-psychiatric outpatient clinics (Bodenheimer and Grumback (2002), U.S. Census 2004\(^1\)). Clinical operations are driven by the patient schedule, which determines the arrival time of patients to the clinic. Patient scheduling affects all aspects of clinical operation and is of primary interest in any effort to improve clinic operations. Clinical managers and physicians are quick to identify inadequate patient scheduling as a major source of operational inefficiency and patient dissatisfaction. Although clinical scheduling has a long research history, clinical impact has been limited, a major reason being that published methods inadequately

address routine factors that complicate clinical scheduling, such as sequential patient call-in, patient no-show and walk-in, and on-call physicians.

Perhaps the most vexing problem in patient scheduling is patient no-show, that is, patients not showing up for scheduled appointments. No-show rates can be as high as 40% in some clinics (Lee et al. (2005)). No-show patients introduce significant uncertainty into clinical operations and limit accessibility to other patients by reserving appointment slots that go unused. Although many factors have been cited as indicators of patient no-show including patient demographics, medical conditions, and environment (Deyo and Inui (1980)), there is very little quantitative research that uses patient no-show behavior in the scheduling process. Thus, our research focuses on using no-show behavior to create patient schedules that balance factors such as patient waiting time, staff overtime, physician utilization, and clinic revenue.

In this paper we present a sequential clinical scheduling mechanism for patients with general service time distributions and multiple no-show probabilities. We further focus on the special case of gamma service times and show how the computational requirements can be steeply reduced for this case. Exhaustive computational studies and discussions that illustrate practical behavior and provide insights are also presented.

We refer to our scheduling procedure as myopic since the assignment decision does not consider the future call-in sequence. We consider a multi-objective optimization formulation that balances rewards and costs for patient waiting and staff overtime at the end of the day. This objective can be conveniently interpreted as the expected profit. We show that the proposed policy yields an objective evolution that is unimodal. By unimodal, we mean that the expected profit for a schedule is non-decreasing up to the addition of some patient and then monotone decreasing thereafter. This unimodality is critical since it provides an optimal stopping criterion with the guarantee that the objective will always decrease thereafter.

We note that the optimal sequential scheduling problem can be easily formulated using a dynamic programming tree. However, the state space is very complex and computation is intractable for all but the smallest problem instances. Further, it is easy to show that the optimal policy will not be unimodal in objective evolution, complicating the identification of a stopping criteria.

The paper is structured as follows. Section 2 provides a brief literature review on clinical scheduling, while Section 3 presents our model formulation, the scheduling algorithm, and the derivation of various expressions that are necessary for the scheduling algorithm. The special case of service times having a gamma distribution is considered in Section 4. Section 5 provides the theoretical guarantees that establish the unimodality of the objective evolution. Section 6 presents the computational results and discussions. Concluding remarks are made in Section 7.
2. Literature Review

Cayirli and Veral (2003) provide an extensive review of clinical scheduling research, covering eighty papers from as early as 1952 (Bailey (1952), Lindley (1952)) up to 2003. They categorize the literature as static vs. dynamic; by performance measure; by system design; and by methodology. We will briefly discuss each of these.

In static scheduling, appointment times are not adjusted once the scheduled day begins, while in the dynamic case, schedule adjustments can be made as the day evolves. Dynamic scheduling is appropriate for in-patient situations where patients are available for early service or can wait in their rooms if the schedule is delayed. It can also be applied in outpatient settings with same day scheduling. In most outpatient clinics, however, there is little opportunity to adjust patient appointment times once the schedule is fixed. Most literature focuses on the static case, which typically involves a set of $N$ punctual patients with $iid$ service times to be scheduled for a single session (day) with a single physician (single server). A representative set of recent static papers includes Vanden-Bosch and Dietz (2000, 2001), Vanden-Bosch et al. (1999), Denton and Gupta (2003), Gupta and Wang (2007), Ho and Lau (1992), Lau and Lau (2000) and Robinson and Chen (2003).

Performance measures dictate how a schedule is evaluated and are categorized as time, congestion, and fairness-based. Time measures are a function of patient waiting time, physician idle time, and staff overtime; congestion measures capture features such as waiting room utilization; and fairness measures look at how waiting times are distributed across the day (often, patients arriving later experience greater expected waiting). Gupta et al. (2006) proposes several additional measures. A detailed review is provided in Mondschein and Weintraub (2003).

The design of an appointment system is specified by the “block” (number of patients arriving at the beginning of an appointment period), the “initial block” (number of patients arriving for the initial appointment), and the “interval” (length of the appointment interval which is either fixed or variable). These parameters can be adjusted based on the presence of environmental complexities such as walk-ins, urgent and emergency patients, and patient specific service time distributions. More detailed discussions of system design in complex environments is provided in Cayirli et al. (2006), Harper and Gamlin (2003), Ho and Lau (1992), Klassen and Rohleder (2004), Liu and Liu (1998) and Rohleder and Klassen (2002).

Methodological classes include analytical modeling and simulation. Analytical approaches use stochastic and deterministic operations research and focus on basic appointment scheduling with

Several additional papers on clinical scheduling have appeared since Cayirli and Veral (2003) was published. Three of these address clinical scheduling with patient no-shows and overbooking and thus we will confine our discussion to these.

Kaandorp and Koole (2007) develop a model for a single server system with exponential service and a single no-show probability for all patients. The objective function minimizes expected waiting time, server idle time, and overtime. They prove that the model is multi-modular so that local search can be used to obtain a globally optimal schedule. Their search algorithm is of super-polynomial complexity since the neighborhood of a given schedule is of exponential size.

LaGanga and Lawrence (2007) develop a clinical scheduling model for a single server with deterministic service times and a single no-show probability. Their objective function maximizes the net office utility which is equal to expected profit minus expected waiting and overtime costs. Cost items are modeled in both linear and quadratic forms. Using numerical experiments, they show that overbooking increases with no-show probability and conclude that schedule overbooking is very effective at reducing the negative impact of patient no-show.

Kim and Giachetti (2006) also propose clinical overbooking to cope with patient no-show, again with a single no-show probability for all patients. Their model assumes deterministic service time and maximizes expected revenue minus overtime and penalty costs incurred when patients leave without being seen. They do not explicitly consider patient waiting time as part of the objective. Their numerical results indicate that overbooking can significantly improve the revenue as well as provide increased access to patients.

Although these papers are interesting and provide some insights, they are not applicable and would not perform well in many clinics for at least two reasons. Firstly, they do not consider sequential scheduling, that is, they assume that the complete set of patients is known when the schedule is generated. In our experience with clinical partners, schedules are rarely built in this way. Rather, schedules are constructed as patients call for appointments. Schedulers do not know how many patients will call for appointments and eventually be added to the schedule. Nor do schedulers
know how many should be added, since they have no optimal stopping criteria. Further, there is little opportunity to adjust the schedule once completed. Secondly, these papers consider only a single no-show probability, which does not match well with reality. The research literature shows that patient no-show probabilities can be estimated based on patient demographics, prevailing environmental conditions, and lead time to appointment (Dervin et al. (1978), Deyo and Inui (1980), Goldman et al. (1982), Gruzd et al. (1986), Martinez et al. (1987), Cashman et al. (2004), Lee et al. (2005)). Our own unpublished research shows that significant improvements in scheduling efficiency can be achieved when more accurate no-show modeling is in place.

In Muthuraman and Lawley (2007), we develop a sequential scheduling model with exponential service times and multiple patient no-show probabilities. The objective maximizes the expected revenue for patients seen minus costs for patient waiting and staff overtime. The paper leverages heavily on the memoryless property of the exponential distribution to show that expected profit evolution is unimodal. That is, the expected profit of a schedule is non-decreasing with the number of patients until some critical number is scheduled and then decreases monotonically as more patients are added, which provides an optimal stopping rule, after which patients are rejected for that particular schedule and can be considered for the schedules of other days. The work presented in this paper generalizes this approach by allowing arbitrary service time distributions, which requires that we include variables capturing time spent in service.

3. Problem Formulation

Let the period of interest (often called a “session” and typically 4 to 8 hours) be divided into $I$ intervals referred to as “slots”, and the length of the $i^{th}$ slot be denoted by $\Delta t_i$. Patients calling for an appointment can be scheduled in one of the $I$ slots or rejected. Scheduled patients have a no-show probability, and each patient arrives (we assume on-time) independently of others. Arriving patients join a queue and if they are not serviced in their scheduled slot, they overflow to the next slot.

Suppose $n$ patients have been scheduled for a given session. Let $X^n_i$ denote the number of patients arriving at the beginning of slot $i$ and $Y^n_i$ be the number of patients waiting and in-service at the end of slot $i$ (see Figure 1). Let $\eta_i$ be the time the in-service patient has spent in service at the end of slot $i$. We also define $Y^n_0 = \eta_0 = 0$.

Let $L_i$ be the random number denoting the number of service completions provided the queue does not empty. $L_i$ is drawn from a distribution that implicitly depends on $\eta_{i-1}$. Then $\min(L_i, Y^n_{i-1} + X^n_i)$ represents the number of service completions in slot $i$. The overflow from slot $i$ is $Y^n_i = \max(Y^n_{i-1} + X^n_i - L^n_i, 0)$. The overflow model is illustrated in Figure 1.
As stated earlier, we assume that each patient has an estimated no-show probability. We partition the set of patients into \( J \) groups such that a patient belonging to group \( j \) has a probability \( p_j > 0 \) of arriving and a probability \( 1 - p_j \) of not showing.

![Figure 1 The Slot Model](image)

After \( n \) calls, a session’s schedule is represented by the matrix \( S^n \in \mathbb{R}^{I \times J} \), where \( S^n_{i,j} \) denotes the number of patients of type \( j \) scheduled for slot \( i \). Thus, \( N^n_i = \sum_j S^n_{i,j} \) is the total number of patients scheduled for slot \( i \). When the context is clear, in the sequel, we will often suppress the superscript (as in Figure 1).

Let \( \Delta_{i,j} \) be an assignment matrix 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 \( i \). For notational convenience, we take the matrix \( Q^n \equiv Q(S^n) \).

Consider the \( i \)th row of a given \( S^n \) and let \( \Phi \) 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 \( \pi_j \leq S^n_{i,j} \) for all \( j \). Conditioning on the event that \( \pi_j \) number of type \( j \) patients show,

\[
Q^n_{i,m} = \Pr\{X^n_i = m\} = \sum_{\pi \in \Phi} \Pr\{X^n_i = m|\pi_1, \ldots, \pi_J\}\Pr\{\pi_1, \ldots, \pi_J\} = \sum_{\pi \in \Phi} \Pr\{\pi_1, \ldots, \pi_J\} = \sum_{\pi \in \Phi} \prod_j \frac{S^n_{i,j}!}{\pi_j!(S^n_{i,j} - \pi_j)!} p_j^{\pi_j} (1 - p_j)^{S^n_{i,j} - \pi_j}. \tag{1}
\]

Further simplification yields the convenient recurrence relation

\[
Q^n_{i,m} = \begin{cases} 
Q^n_{i,m-1}(1 - p_j) + (Q^n_{i,m-1}) p_j & \text{when } m \geq 1 \text{ and} \\
Q^n_{i,0}(1 - p_j) & \text{when } m = 0.
\end{cases} \tag{2}
\]
The function $R(\cdot)$ represents 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 $i$. As with $Q^n$, $R^n \equiv R(S^n)$. Obviously, $Q^n, R^n \in \mathbb{R}^{I \times \hat{N}^n}$ where $\hat{N}^n = \max_i N_i^n$. By definition, given $S$,

$$R^n_{ik} = \Pr\{Y^n_i = k\}. \hspace{1cm} (3)$$

The calculation of $R^n_{ik}$ is discussed in subsection 3.1.

Suppose the $n^{th}$ patient calling for an appointment is of type $j$. Letting $U$ be the set of slots (that is, integers from 1 to $I$), we wish to select a slot $i \in U$ for the patient so as to maximize an objective, $f(Q^n, R^n)$. That is, we assign patient $n$ to slot $i^*$ where

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

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

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 to maximize

$$f(Q^n, R^n) = r \sum_i \sum_m mQ^n_{i,m} - \sum_i c_i \sum_k kR^n_{i,k}$$

$$= \mathbb{E}\left[r \sum_{i=1}^I X^n_i - \sum_{i=1}^I c_i Y^n_i\right]. \hspace{1cm} (6)$$

### 3.1. Calculating Overflow Probabilities

Consider $R^n_{i,k}$, that is, the probability of $k$ patients over-flowing into slot $i+1$ from slot $i$,

$$R^n_{i,k} = \Pr\{Y_i = k\}$$

$$= \Pr\{\max(X_i + Y_{i-1} - L_i, 0) = k\}$$

$$= \begin{cases} \Pr\{X_i + Y_{i-1} - L_i = k\} & k > 0 \\ \Pr\{X_i + Y_{i-1} - L_i \leq 0\} & k = 0 \end{cases} \hspace{1cm} (7)$$

Further conditioning yields,

$$R^n_{i,0} = \sum_m \sum_k \Pr\{m + \tilde{k} \leq L_i\}Q^n_{i,m} R^n_{i-1,\tilde{k}}$$

$$= \sum_m \sum_{\tilde{k}} (1 - F_{L_i}(m + \tilde{k}))Q^n_{i,m} R^n_{i-1,\tilde{k}} \hspace{1cm} (8)$$
and similarly for $k > 0$,

$$R_{i,k}^n = \sum_{m} \sum_{k} \Pr\{m + \tilde{k} - k = L_i\} Q^n_{i,m} R_{i-1,k}^n$$

$$= \sum_{m} \sum_{k} f_{L_i}(m + \tilde{k} - k) Q^n_{i,m} R_{i-1,k}^n$$ (9)

where $F_{L_i}(m) = \Pr\{L_i \leq m\}$ and $f_{L_i}(m) = \Pr\{L_i = m\}$ implicitly depend on the distribution of $\eta_{i-1}$ and are obtained from the general service time distribution as detailed below.

Let $f(t)$ and $F(t)$ represent the probability density function (p.d.f.) and the cumulative distribution function (c.d.f.) of the general service times, respectively. By first conditioning on the realization of $\eta_{i-1}$, we evaluate the distribution of $L_i$.

$$\Pr\{L_i = 0|\eta_{i-1}\} = \frac{\Pr\{T_{i,0} \geq \eta_{i-1} + \Delta t_i\}}{\Pr\{T_{i,0} \geq \eta_{i-1}\}} = \frac{1 - F(\Delta t_i + \eta_{i-1})}{1 - F(\eta_{i-1})}$$

where $T_{i,0}$ is a draw from the service time distribution $F(\cdot)$, representing the total service time for the first patient, who is in service at the beginning of slot $i$. For $n \geq 1$ we have,

$$\Pr\{L_i = n|\eta_{i-1}\} = \Pr\{L_i \geq n|\eta_{i-1}\} - \Pr\{L_i \geq n + 1|\eta_{i-1}\}$$ (10)

and letting $U_{i,n} = T_{i,0} + \ldots + T_{i,n}$,

$$\Pr\{L_i \geq n|\eta_{i-1}\} = \Pr\{U_{i,n-1} \leq \Delta t_i + \eta_{i-1}|\eta_{i-1}\} = \frac{1}{1 - F(\eta_{i-1})} \int_{\eta_{i-1}}^{\Delta t_i + \eta_{i-1}} \int_{0}^{\Delta t_i + \eta_{i-1} - u_{i,0}} \ldots \int_{0}^{\Delta t_i + \eta_{i-1} - u_{i,n-2}} \prod_{j=0}^{n-1} f(t_{i,j}) dt_{i,j}$$ (11)

where, $u_{i,n} = t_{i,0} + t_{i,1} + \ldots + t_{i,n}$. When $n = 1$ an integral over $t_{i,0}$ alone resides, that is, only the first integral. In general for any given $n$, the above equation requires the computation of $n$ integrals.

Further simplification using equation (10) yields,

$$\Pr\{L_i = n|\eta_{i-1}\} = \frac{1}{1 - F(\eta_{i-1})} \int_{\eta_{i-1}}^{\Delta t_i + \eta_{i-1}} \ldots \int_{0}^{\Delta t_i + \eta_{i-1} - u_{i,n-2}} \int_{0}^{\Delta t_i + \eta_{i-1} - \sum_{j=0}^{n} f(t_{i,j}) dt_{i,j}$$ (12)

For any given $n$, the above equation requires the computation of $n + 1$ integrals. Finally, integrating with respect to $\eta_{i-1}$ we get,

$$f_{L_i}(n) = \int_{0}^{\infty} \Pr\{L_i = n|\eta_{i-1}\} d\Psi_{i-1}(\eta_{i-1})$$ (13)
where $\Psi_{i-1}()$ represents the c.d.f. of $\eta_{i-1}$. For the case of service times taking an exponential distribution, equation (13) can be further simplified to reveal that the distribution of $L_i$ does not depend on $\eta_{i-1}$ and is simply a Poisson random variable.

### 3.2. Computation of $\Psi_i$

We have from our definition that $\eta_0 = 0$. Note that when no patient services are completed in the first slot, $\eta_1 = \Delta t_1$. Now, for a given $\eta \in [0, \Delta t_i + \eta_{i-1}]$, the c.d.f of $\eta_i$ is given by

$$\Psi_i(\eta) = 1 - \int_0^\infty \Pr\{\eta_i \geq \eta|\eta_{i-1}\}d\Psi_{i-1}(\eta_{i-1})$$

where,

$$\Pr\{\eta_i \geq \eta|\eta_{i-1}\} = 1_{\{\eta \leq \Delta t_i + \eta_{i-1}\}} \Pr\{L_i = 0\} + \sum_{n=1}^\infty \Pr\{U_{i,n-1} \leq \Delta t_i + \eta_{i-1} - \eta|U_{i,n-1} \leq \Delta t_i + \eta_{i-1}, U_{i,n} \geq \Delta t_i + \eta_{i-1}\} \Pr\{L_i = n\}$$

$$= 1_{\{\eta \leq \Delta t_i + \eta_{i-1}\}} \Pr\{L_i = 0\} + \sum_{n=1}^\infty \Pr\{U_{i,n-1} \leq \Delta t_i + \eta_{i-1} - \eta,T_{i,n} \geq \eta\} \Pr\{L_i = n\}$$

$$= 1_{\{\eta \leq \Delta t_i + \eta_{i-1}\}} \Pr\{L_i = 0\} + \sum_{n=1}^\infty \Pr\{U_{i,n-1} \leq \Delta t_i + \eta_{i-1} - \eta\} \Pr\{T_{i,n} \geq \eta\} \Pr\{L_i = n\}$$

$$= 1_{\{\eta \leq \Delta t_i + \eta_{i-1}\}} \Pr\{L_i = 0\} + \sum_{n=1}^\infty \Pr\{U_{i,n-1} \leq \Delta t_i + \eta_{i-1} - \eta\} \Pr\{L_i = n\}. \quad (14)$$

The indicator function for event $\mathcal{A}$, in the above, is denoted by $1_{\{\mathcal{A}\}}$. The distribution of $U_{i,j}$ and $L_i$ in equation (14) can be computed as in subsection 3.1.

### 3.3. The Scheduling Policy

The scheduling policy given enumerates all possible assignments for the current patient and selects the assignment that maximizes the objective. It is sequential in that it assigns patients as they call and myopic in that it does not consider future arrivals when making the assignment. Further, it will reject the patient and terminate when the objective declines.

1. Initialize with $S^n_{i,j} = 0$, $Q^n_{i,0} = R^n_{i,0} = 1$ for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$ and set $n = 1$.
2. Wait for $n^{th}$ call and let the $n^{th}$ call be from a patient of type $j$.
3. For each $i \in U$: Set $S^n_i = S^{n-1} + \Delta t_i$; Compute $Q^n_i$ from $Q^{n-1}$ using equation (1); Compute $\Psi_i(\cdot)$ and $R^n_i$ as described in sections 3.2 and 3.1; Compute $f^n_i = f(Q^n_i, R^n_i)$.
4. If $\max_i f^n_i \geq f^{n-1}$
   
   (a) then $i^* = \arg \max_i f^n_i$, $S^n = S^{n-1} + \Delta t^* j, Q^n = Q^n_{i^*}, R^n = R^n_{i^*}$. Set $n = n + 1$ and iterate by going to 2.
   
   (b) else terminate.
It is likely that calling patients have time preferences. To accommodate this, we can define $U_n \subset U$ as the set of slots that the $n^{th}$ calling patient prefers. Then, step 4 can be modified to maximize over $i \in U_n$. Note that a sequence of singleton $U_n$’s would then eliminate any scheduling flexibility. While such sequences are unlikely in practice, they are theoretically feasible. Any arbitrary objective evolution can be constructed by such sequences and thus a stopping criteria as in step 4 cannot guarantee a maxima. Hence, in section 5, where we seek theoretical guarantees on the behavior of objective evolutions, we will restrict our attention to $U_n = U$ for all $n$.

4. Gamma Distributed Service Times

For general service time distributions, the evaluation of $\Pr\{L_i = n|\eta_{i-1}\}$ using equation (12) involves the evaluation of $n + 1$ integrals. We consider the special case of gamma distributed service times in this section and demonstrate that in this case the computation can be significantly reduced to the evaluation of a single integral.

The gamma distribution is characterized by two parameters, the shape and the scale. These two parameters often provide sufficient flexibility to model different empirical distributions with varying skewness. The sum of iid gamma random variables in turn is also a gamma distributed random variable. Let the service time for a patient be gamma distributed with parameters $\alpha$ and $\lambda_0$. We represent its c.d.f and p.d.f by $G_1(\cdot)$ and $g_1(\cdot)$, respectively. Here

$$g_1(t) = t^{\alpha-1} e^{-\frac{t}{\lambda_0}} \frac{1}{\lambda_0^\alpha \Gamma(\alpha)}, \quad \text{for } t \geq 0.$$  

(15)

The sum of $k$ of these random variables (service times) is then a gamma random variable with parameter $k\alpha$ and $\lambda_0$, the c.d.f. and p.d.f. of which are represented simply by $G_k(\cdot)$ and $g_k(\cdot)$. Figure 2 shows examples of gamma distributions for different values of the parameters $\alpha$ and $\lambda_0$ as well as an exponential distribution.

Now, from equation (11) we have, for $n = 1$

$$\Pr\{L_i \geq 1|\eta_{i-1}\} = \frac{1}{1 - G_1(\eta_{i-1})} \int_{\eta_{i-1}}^{\Delta_t + \eta_{i-1}} g_1(t_0)dt_0$$

and for $n \geq 2$,

$$\Pr\{L_i \geq n|\eta_{i-1}\} = \frac{1}{1 - G_1(\eta_{i-1})} \int_{\eta_{i-1}}^{\Delta_t + \eta_{i-1}} \int_0^{\Delta_t + \eta_{i-1}} g_1(t_0)g_{n-1}(u_{n-1})dt_0du_{n-1}$$

$$= \frac{1}{1 - G_1(\eta_{i-1})} \int_{\eta_{i-1}}^{\Delta_t + \eta_{i-1}} g_1(t_0)G_{n-1}(\Delta_t + \eta_{i-1} - t_0)dt_0$$

Thus, from equation (10) we get,

$$\Pr\{L_i = 1|\eta_{i-1}\} = \int_{\eta_{i-1}}^{\Delta_t + \eta_{i-1}} [1 - G_1(\Delta_t + \eta_{i-1} - t_0)] g_1(t_0)dt_0$$

$$\frac{1}{1 - G_1(\eta_{i-1})}$$
and for \( n > 1 \),

\[
\Pr\{ L_i = n | \eta_{i-1} \} = \frac{\int_{\eta_{i-1}}^{\Delta t_i + \eta_{i-1}} \left[ G_{n-1}(\Delta t_i + \eta_{i-1} - t_0) - G_n(\Delta t_i + \eta_{i-1} - t_0) \right] g_1(t_0) dt_0}{1 - G_1(\eta_{i-1})}
\]

(17)

The distribution of \( \eta_i \) also assumes a simpler form for gamma service times. From equation (14),

\[
\Pr\{ \eta_i \geq \eta | \eta_{i-1} \} = 1_{\{ \eta \leq \Delta t_i + \eta_{i-1} \}} \Pr\{ L_i = 0 \} + \sum_{n=1}^{\infty} \frac{G_n(\Delta t_i + \eta_{i-1} - \eta)}{G_n(\Delta t_i + \eta_{i-1})} \Pr\{ L_i = n \}
\]

5. Unimodal Objective Evaluation

We show that the myopic policy is unimodal under general service times in this section. Theorem 1 shows that if the \( n^{th} \) assignment yields a decrease in expected objective then the \( n + 1^{st} \) assignment would yield a further decrease. By induction, this would imply unimodal evolution of the objective. As mentioned earlier, by unimodal we mean that the expected objective is non-decreasing up to a particular call-in patient and then is monotone decreasing thereafter.

**Theorem 1** If \( n \) is such that \( f(Q^n, R^n) < f(Q^{n-1}, R^{n-1}) \), then \( f(Q^{n+1}, R^{n+1}) < f(Q^n, R^n) \).

Proof: We have,

\[
f(Q^n, R^n) - f(Q^{n-1}, R^{n-1}) = \mathbb{E} \left[ r \sum_{i=1}^{l} X_i^n - \sum_{i=1}^{l} c_i Y_i^n \right] - \mathbb{E} \left[ r \sum_{i=1}^{l} X_i^{n-1} - \sum_{i=1}^{l} c_i Y_i^{n-1} \right]
\]
$$= r \sum_{i=1}^{l} \left[ X^n_i - X^{n-1}_i \right] - \sum_{i=1}^{l} c_i \left[ Y^n_i - Y^{n-1}_i \right] \tag{18}$$

Now consider the first expectation,

$$\mathbb{E} \sum_{i=1}^{l} \left[ X^n_i - X^{n-1}_i \right] = \sum_{i=1}^{l} \sum_x x \Pr \{ X^n_i - X^{n-1}_i = x \}$$

$$= \sum_{i=1}^{l} \Pr \{ X^n_i - X^{n-1}_i = 1 \}$$

$$= \Pr \{ n^{th} \text{ patient arrives for one of the} \ i \text{ slots} \}$$

$$= p_j.$$  

In the above, we let $j_n$ denote the type of the $n^{th}$ patient. Conditioning the second expectation on the event $E_n$, which denotes that the $n^{th}$ patient shows up,

$$\mathbb{E} \left[ Y^n_i - Y^{n-1}_i \right] = p_j \mathbb{E} \left[ Y^n_i - Y^{n-1}_i | E_n \right]. \tag{19}$$

Hence from equation (18),

$$f(Q^n, R^n) - f(Q^{n-1}, R^{n-1}) = p_j \left[ r - \sum_{i=1}^{l} c_i \mathbb{E} \left[ Y^n_i - Y^{n-1}_i | E_n \right] \right]. \tag{20}$$

Using the simplification provided in equation (20), to prove theorem 1, it suffices to show that

$$r < \sum_{i=1}^{l} c_i \mathbb{E} \left[ Y^{n+1}_i - Y^n_i | E_{n+1} \right] \tag{21}$$

if

$$r < \sum_{i=1}^{l} c_i \mathbb{E} \left[ Y^n_i - Y^{n-1}_i | E_n \right]. \tag{22}$$

To show that (22) implies (21), we need more notation. Let the $n^{th}$ and the $n+1^{st}$ patients be of types $j_n$ and $j_{n+1}$. Let the best assignment of the $n^{th}$ and the $n+1^{st}$ patients, using the scheduling policy detail in subsection 3.3, be to slots $i_n$ and $i_{n+1}$, respectively. We have $S^n = S^{n-1} + \Delta^{i_n j_n}$ and $S^{n+1} = S^n + \Delta^{i_{n+1} j_{n+1}}$.

Now also consider the schedule $\bar{S}^n$ created by taking the schedule $S^{n-1}$ and assigning a patient of type $j_n$ to slot $i_{n+1}$ (instead of slot $i_n$ which is the best assignment). That is, $\bar{S}^n = S^{n-1} + \Delta^{i_{n+1} j_n}$. The schedules $S^n, S^{n+1}$, are respectively constructed by taking $S^{n-1}, S^n$ and adding a patient in slot $i + 1$. And note that $S^n - S^{n-1} = \Delta^{i_n j_n}$. Hence,

$$\sum_{i=1}^{l} c_i \mathbb{E} \left[ Y^n_i - Y^{n-1}_i | E_n \right] \leq \sum_{i=1}^{l} c_i \mathbb{E} \left[ Y^{n+1}_i - Y^n_i | E_{n+1} \right] \tag{23}$$
Next compare $S^n$ and $S^n$, since $i_n$ is the best assignment of the $n^{th}$ patient, $f(Q^n, R^n) \geq f(Q^n, R^n)$. Simplification yields,

$$\sum_{i=1}^I c_i E[Y^n_i - Y^{n-1}_i|E_n] \leq \sum_{i=1}^I c_i E[Y^n_i - Y^{n-1}_i|E_n]$$

From (23) and (24), we have

$$\sum_{i=1}^I c_i E[Y^n_i - Y^{n-1}_i|E_n] \leq \sum_{i=1}^I c_i E[Y^{n+1}_i - Y^n_i|E_{n+1}]$$

The above along with (22) yields (21).

6. Computational Results for Empirical Service Time

The primary purpose of this section is to understand the impact of and gain insights into the proposed scheduling methodology. To this extent, first we evaluate the value of being able to use general service time distributions as opposed to approximations like gamma and exponential service times. Next, we evaluate the performance of our scheduling methodology against heuristic scheduling procedures that have been tested and recommended in literature. Finally we look at the effect of both mean and variance of service time distributions on the expected profit and the number of patients scheduled.

6.1. Quality of approximations

One of the primary questions of interest is the value that is added in being able to handle general service times. To answer this, we first need a general service time distribution that is estimated from actual data. Cayirli et al. (2006) use data collected from a primary health care clinic in a New York Metropolitan Hospital which serves approximately 300,000 patients per year. Based on the Kolomogorov-Smirnov test (at $\alpha = 0.05$), they conclude that, a log-normal distribution is the best fit. We will take the same log-normal distribution (with mean $\mu = 15$ minutes and standard deviation $\sigma = 5$ minutes) as our general service time distribution. We assume that there are three types of patients, and the no-show probabilities are $p = (0.25, 0.5, 0.75)$. Slot overflow costs are set to $c_i = $100 for $i = 1, \ldots, I-1$ and $c_I = $300. The reward for each scheduled patient is $r = $200.

For comparisons, we approximate this general service time by the gamma and exponential fits shown in Figure 3. The exponential and gamma are chosen to match the mean and variance of the general service (log-normal) times. For each case, the general service distribution, the gamma fit, and the exponential fit, we schedule patients according to the methodology described in section 3, section 4, and Muthuraman and Lawley (2007), respectively. Figure 4 shows the evolution of the expected profit obtained using these three service time distributions. It is evident from Figure 4
that the expected profit from the gamma fit and the general service time are very close, while the expected profit from the exponential fit is significantly less, where the expected profits are calculated using equation 6. This is to be interpreted as the expected profit if the realized service times are also drawn from the gamma and exponential distributions. Hence, Figure 4 does not provide a measure of how well a gamma or an exponential based schedule performs on in a general service environment, but it does provide a measure of how close the gamma and exponential models are to the general service model.

We now investigate how well a gamma or an exponential based schedule will perform in our general service environment. We seek as a measure, the value lost in approximating the service time distribution for 100 different call-in sequences generated randomly. We use the scheduling methodologies for general, gamma and exponential and prepare 100 schedules each. For each of the 300 schedules we simulate 10 no-show/show and service time realizations (using general service times) and plot the average realized profits for each schedule and methodology in Figure 5. The average realized profit for all these runs are respectively $4574, $4578, and $3881 for general service time, gamma fit and exponential. The differences amongst the general service times and gamma service times are statistically insignificant. Figure 5 clearly illustrates the value gained in being able to handle general service distributions as opposed to approximating the system with an exponential fit. Further it also demonstrates the quality of the gamma approximation at least for this service time distribution observed at the New York Metropolitan Hospital.

![PDF of Lognormal and Gamma with Mean = 0.25 hr and SD = 0.0833 hr.](image.png)
Figure 4 Objective Evolution for Lognormal and Gamma with Mean = 0.25 hr and SD = 0.0833 hr.

Figure 5 Average Profit Obtained from Simulation for Lognormal with Mean = 0.25 hr and SD = 0.0833 hr and Using Schedules from General Service Time, Gamma Fit and Exponential Fit.

Note that the computational requirements for the gamma distributions are much smaller than that for general service times. Calculating $Pr\{L_i = \eta_i\}$ for the Gamma service distributions requires the computation of only a one-dimensional integral, whereas general service distributions require the computation of a $(n+1)$-dimensional integral. For our integral computations here, we use Gaussian Quadratures with 8-points for each integral.
6.2. Comparison with the “Appointment Rules” of Cayirli et al. (2006)

In this section we perform a comparative study of the “Appointment Rules” proposed by Cayirli et al. (2006) with our myopic scheduling policy under general service time. We assume that the service time distribution is lognormal with mean 0.25 hr. and standard deviation 0.0833 hr and schedule different arrival sequences of patients using each of the seven “Appointment Rules” as well as the myopic policy and compare the evolution of the objective function. In the experiments described below we assume that the minimum number of patients to schedule \((N)\) is 20. The “Appointment Rules” are functions of five parameters, viz. \((\beta_1, \beta_2, \beta_3, k_1, k_2)\). In our experiments we use \((\beta_1 = 0.15, \beta_2 = 0.3, \beta_3 = 0.05, k_1 = 10, k_2 = 18)\) (Cayirli et al. (2006)).

6.2.1. IBFI, 2BEG Rules: In both the IBFI and 2BEG the slot length is held constant and is equal to the mean of the service time. According to IBFI rule only one patient is assigned to each slot whereas in 2BEG rule 2 patients are assigned to the first slot and 1 patient is assigned to each of the remaining slots. Figure 6 shows the evolution of the objective functions when patients are scheduled using these two rules as well as the myopic scheduling policy under general service time.

![Figure 6](image_url)  
**Figure 6** Objective Evolution for IBFI & 2BEG and Myopic Policy under General Service Time.

6.2.2. OFFSET Rule: In OFFSET rule first \((k_1 - 1)\) patients are scheduled earlier whereas the rest are scheduled later compared to the IBFI rule. Figure 7 shows the evolution of the objective functions when patients are scheduled using this rule as well as the myopic scheduling policy under general service time.
6.2.3. DOME and 2BGDM Rules: According to the DOME rule, the first \((k_1 - 1)\) patients are scheduled earlier, the following patients up to \((k_2 - 1)\) are assigned later, and the rest are scheduled earlier compared to the IBFI rule. The 2BGDM rule is a combination of the DOME and 2BEG rule. Figure 8 shows the evolution of the objective functions when patients are scheduled using these rule as well as the myopic scheduling policy under general service time.
6.2.4. MBFI Rule: According to the MBFI rule, 2 patients are scheduled at each slot where the length of each slot is twice the mean of the service time distribution. Figure 9 shows the evolution of the objective functions when patients are scheduled using this rule as well as the myopic scheduling policy under general service time.

![Objective Evolution for MBFI and Myopic Policy under General Service Time.](image)

6.2.5. MBDM Rule: The MBDM rule is a combination of the DOME and MBFI rule. Figure 10 shows the evolution of the objective functions when patients are scheduled using this rule as well as the myopic scheduling policy under general service time.

Comparing the objective function from each of the seven “Appointment Rules” and the myopic scheduling policy under general service time we see that maximum expected profit from the latter is always higher, in that it is always able to profitably schedule more patients.

6.3. Effect of service time variance
In this section we study the effect the variance of the service time distribution has on expected profit.

To study the effect of variance of the service time distribution on expected profit, we systematically increase the variance of the lognormal service time keeping the mean constant at 15 minutes. Figure 11 plots the the maximum of the objective evolution as we increase the variance of the service time. It is clear from Figure 11 that the value of the maximum profit decreases as we increase the variance of the service time distribution.
Figure 10  Objective Evolution for MBDM and Myopic Policy under General Service Time.

Figure 11  Maximum Expected Profit with Increasing SD and Fixed Mean of the Service Time. The x-axis units are in hours.

7. Concluding Remarks
In this paper we develop a sequential clinical scheduling policy for a clinic with general service time distributions. We show that the policy developed, though myopic in nature, yields a unimodal objective evolution that provides a convenient stopping criteria. The major challenge with the general service time formulation and scheduling methodology is that it requires the numerical
computation of several large multidimensional integrals. However, we show how the computational needs can be reduced significantly when service times are approximated by a gamma distribution. In this case the multidimensional integrals are shown to reduce to a single integral. We also demonstrate that for an empirical service time distribution observed in the New York Metropolitan Hospital, the gamma approximation to the general service time distribution yields an extremely good approximation as opposed to the simpler memory-less exponential service time approximation. Finally, we also compare the performance of our scheduling method to those studied and recommended in literature.

References


