Clinic Scheduling Models with Overbooking for Patients with Heterogeneous No-show Probabilities*

Bo Zeng, Ji Lin and Mark Lawley

email: {bzeng, lin35, malawley}@purdue.edu
Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN 47907
Ayten Turkcan,
e-Enterprise Center, Purdue University, West Lafayette, IN 47907
email: aturkcan@purdue.edu

February 26, 2008

Abstract

Clinical overbooking is intended to reduce the negative impact of patient no-shows on clinic operations and performance. In this paper, we study the clinical scheduling problem with overbooking for heterogeneous patients, i.e. patients who have different no-show probabilities. We consider the objective of maximizing expected profit, which includes revenue from patients and costs associated with patient waiting times and physician overtime. We show that the objective function with homogeneous patients, i.e. patients with the same no-show probability, is multimodular. We also show that this property does not hold when patients are heterogeneous. We identify properties of an optimal schedule with heterogeneous patients and propose a local search algorithm to find local optimal schedules. Then, we extend our results to sequential scheduling and propose two sequential scheduling procedures. Finally, we perform a set of numerical experiments and provide managerial insights for health care practitioners.

*Key Words*clinical scheduling, overbooking, patient no-show, multimodularity

1 Introduction

The majority of patient care in the U.S. ($80-90\%$ [2, 4]) is provided by outpatient clinics. Clinic operations are typically driven by appointment schedules, and appointment scheduling is often cited by clinic managers as a major opportunity for improvement. Cayirli and Veral [3] provide a comprehensive review of research in outpatient appointment scheduling. They state that most analytical research does not consider factors such as patient no-shows, walk-ins and emergency. Nevertheless, these factors have significant adverse effect on operational efficiency, total revenue and patient satisfaction.

Among these factors, patient no-show is of particular concern because it wastes the available capacity of valuable resources (physician, staff, equipment) and limits clinic access to the patient population. Cayirli and Veral [3] mention that no-show rates are 5-30%. However, Rust et al. [17] report that for some health care settings, such as public pediatric clinics, no-show rates can reach 80%. To reduce the negative impact of patient no-show, clinic schedulers often overbook. However, naive overbooking can lead to longer patient waiting times, clinic overtime, and deteriorating outcomes for patients who leave without being seen [8, 18]. Thus, modeling and analysis must be used to develop a scheduling methodology that properly balances these competing objectives.

*Supported by NSF Grant CMMI 0729463

An ideal overbooking model depends on four characteristics. The first is a valid patient no-show description that captures the real pattern of patient behavior. The second is the underlying service model that reflects the operational dynamics of the clinic. The third is an objective function that reflects the performance concern of clinic managers. And the last is an efficient algorithm that can generate schedules of desired quality in a timely fashion. We give a brief review of existing clinic overbooking models categorized according to these four characteristics. We also include some studies that do not explicitly consider overbooking but that can be used to obtain overbooked schedules.

No-show probabilities can be correlated to factors such as reservation lead time and patient demographics, see Garuda et al. [5] for details. Even though no-show usually differs by patient, almost all overbooking studies assume that patients are homogeneous, i.e. all patients have the same no-show probability. Liu and Liu [12], Laganga and Lawrence [10, 11], Kim and Giachetti [8] and Kaandorp and Koole [7] consider a single no-show rate for all patients in their models. Muthuraman and Lawley [15] provide an exception by explicitly modeling different no-show probabilities.

With clinic dynamics, most researchers develop single server models (a schedule is created for a single physician). Kim and Giachetti [8] and Laganga and Lawrence [10, 11] assume that the service times of patients are deterministic while Kaandorp and Koole [7] and Muthuraman and Lawley [15] use queuing based models with exponential service times. Liu and Liu [12] consider a model for multiple servers and investigate service times with exponential and general distributions.

The performance criteria considered in appointment scheduling models includes revenue from patients, patient waiting time/cost, physician overtime/cost and physician idle time/cost. Kim and Giachetti [8] consider expected revenue and physician overtime. Since the cost of patient waiting time is not included in their model, all patients are assumed to arrive at the beginning of a clinic session, which implies a single block scheduling model. Laganga and Lawrence [11] consider revenue from patients and costs of patient waiting time and physician overtime, in both linear and quadratic objective functions. Kaandorp and Koole [7] explicitly include the cost of physician idle time, as do Liu and Liu [12]. Muthuraman and Lawley [15] consider revenue from patients and costs from patients waiting time and clinic overtime.

In most cases, computing the optimal schedule is computationally intractable and thus most scheduling algorithms are heuristics or simulation based methods [7, 8, 10, 11, 12]. The research by Kaandorp and Koole [7] is of special interest because they show that their model is multimodular. Multimodularity is a property of functions in discrete space, similar to convexity in continuous space, which guarantees that a locally optimal solution is also globally optimal. In contrast to the work just mentioned, Muthuraman and Lawley [15] consider sequential scheduling in which the schedule is constructed as patients seeking appointments call clinic schedulers. Patients must be given their appointment time before the call ends, and thus once a patient appointment is added to the schedule, it is typically not feasible to alter the time. In this case, the set of patients to be scheduled is not initially known and deciding when a schedule is complete becomes a problem. Although the authors provide a scheduling algorithm and derive optimal stopping criteria, the optimal sequential schedule is not characterized. In this study, we derive some properties of an optimal schedule, which can be used to design better algorithms.

The existing studies do not adequately address the question of how the scheduling problem for heterogeneous patients is different from that of homogeneous patients and whether modeling the heterogeneous nature of patient no-show can lead to superior schedules, particularly in a sequential setting where schedules have to be constructed as patients call-in. Even though different no-show probabilities are taken into account in [15], the patients are treated equally while scheduling. The decision to accept (or reject) a patient is given by only looking at the increase (or decrease) in the objective function. However, scheduling patients with high no-show probabilities leads to higher variabilities in daily workload of clinics. In this study, we also investigate the effect of variability in no-show rates on the quality of the resulting schedules.

The remainder of the paper is structured as follows. Section 2 provides the notation and defines the basic problem. In Section 3, we study the structure of the optimization model and prove that it is not multimodular in general. In Section 4, after deriving some important properties of optimal schedules, we propose a local search algorithm to obtain a good schedule and discuss its extension to sequential scheduling settings. In Section 5, we present a computational study to compare the proposed algorithms with the existing methods. Section 6 concludes with some managerial insights that can improve patients scheduling and clinic performance in practice.

2 Problem Definition

We first state our assumptions and then present required notation. We assume that all patient arrivals to the clinic are scheduled (no walk-ins) and that the patient population can be partitioned into categories based on no-show probability. We further assume that the clinic day is partitioned into a set of time slots and that patient appointment times coincide with the beginning of a slot. We also assume that all arriving patients are punctual and that patients are served according to a first-come-first-serve protocol. Finally, we assume that service times are exponential and that they are independent and identically distributed across patients. Notation is as follows:

I	set of slots
i	slot set index, $i \in \{1, \ldots, \lvert I\rvert\}$
t_i	length of slot i
J	set of patient types based on no-show
j	patient type, $j \in \{1, \ldots, \lvert J\rvert\}$
X_i	number of patients arriving at start of slot i
Y_i	number of patients in the system at the end of slot i, overflow from slot i to slot $i+1$
L_i	number of service completions in slot i
λ	service rate
c_i	unit overflow cost from slot i to slot $i+1$ ($i \neq \lvert I\rvert$), $c_i \geq 0$
c_I	unit overflow cost for $i = \lvert I\rvert$, unit overtime cost, typically $c_{\lvert I\rvert} > c_i$
r	revenue per patient, $r > 0$
p_j	probability that patient of type j arrives as scheduled, ($p_1 > \cdots > p_{\lvert J\rvert}$)
n_j	number of patients of type j
S	a schedule ($\in \mathbb{Z}^{\lvert I\rvert \times \lvert J\rvert}$)
$\Delta_{i,j}$	unit matrix such that cell i, j is 1, all others 0
$F(S)$	objective function value of S
$R(S)$	overflow matrix of schedule S
$Q(S)$	arrival matrix of schedule S
$a \sim b$	random variables a and b are iid

For a given set of heterogeneous patients, we formulate the following overbooking model to obtain an optimal schedule S that maximizes the expected total profit.

$$\begin{aligned} \max \ F(S) =& r \sum_{i \in I} E[X_i] - \sum_{i \in I} c_i E[Y_i] \\ \text{s.t.} & \sum_{i \in I} S_{i,j} \leq n_j \\ & S_{i,j} \in \mathbb{Z} \quad \forall\, i \in I, j \in J \end{aligned} \tag{1}$$

We note that $X_i + Y_{i-1}$ is the number of patients in the system at the beginning of slot i and that the number of patients in the system at the end of slot i is given as:

$$Y_i = \max\{Y_{i-1} + X_i - L_i, 0\}. \tag{2}$$

To compute probabilities for X_i and Y_i, Muthuraman and Lawley [15] introduce two matrices, an arrival matrix $[Q_{i,l}]$ such that $Q_{i,l}$ is the probability that l patients arrive at the beginning of slot i, and an overflow matrix $[R_{i,k}]$ such that $R_{i,k}$ is the probability that k patients overflow from slot i to slot $i+1$. These are computed as follows:

$$Q_{i,l}(S) \quad = \quad Pr(X_i = l) = \sum_{\pi \in \Omega} \prod_{j \in J} \frac{S_{i,j}!}{\pi_j!(S_{i,j} - \pi_j)!} p_j^{\pi_j} (1-p_j)^{S_{i,j}-\pi_j},$$

where $\pi = \{\pi_1, \ldots, \pi_{|J|}\}$ with $\pi_j \in \mathbb{Z}_+$ for $j \in J$, $\sum_{j \in J} \pi_j = l$ and Ω is the set of all such vectors.

$$R_{i,m}(S) = \begin{cases} \sum\limits_l \sum\limits_k (1 - F_{L_i}(l+k)) Q_{i,l} R_{i-1,k} & \text{if } m = 0 \\ \sum\limits_l \sum\limits_k f_{L_i}(l+k-m) Q_{i,l} R_{i-1,k} & \text{if } m \geq 1 \end{cases} \tag{3}$$

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

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

Given these equations, we can compute $E[X_i] = \sum_l iQ_{i,l}$ and $E[Y_i] = \sum_k kR_{i,k}$.

Typically, optimization problems such as (1) arising from appointment service systems are very difficult to solve since the objective functions are nonlinear and decision variables are discrete. However, it has recently been shown that if the objective function is multimodular over $\mathbb{Z}^n$, a property similar to convexity in $\mathbb{R}^n$, and constraints are simple upper or lower bound constraints, a well-defined local search algorithm can be used to obtain (global) optimal solutions, see Hajek [6], Altman et al. [1] and Koole and van der Sluis [9]. Based on these results, Kaandorp and Koole [7] prove that their scheduling model for homogeneous patients is multimodular and implement a local search algorithm to obtain an optimal schedule. As a natural extension, it is important to see whether our overbooking model is multimodular. If so, we can use the results to obtain an optimal scheduling method, and if not we are justified in seeking heuristic approaches. Section 3 addresses this problem.

3 Structure of the Overbooking Scheduling Model

In this section, we investigate the multimodularity of the scheduling model given in (1). As an aid to the reader, we make the following informal note about multimodularity before providing the definition. Let f be function on $\mathbb{Z}^m$. When we join the integer points of f by lines, we obtain a new function g on $\mathbb{R}^m$. g is convex if and if f is multimodular. This implies that a local optimum is also a global optimum.

More formally, let $\vec{e}_i$ be the i^{th} standard unit vector of $\mathbb{R}^m$. Then, we define a set of vectors $\Gamma = \{\vec{v}_0, \ldots, \vec{v}_m\} \in \mathbb{Z}^m$ such that $\vec{v}_0 = -\vec{e}_1$, $\vec{v}_i = \vec{e}_i - \vec{e}_{i+1}$, for $i = 1, \ldots, m-1$ and $\vec{v}_m = \vec{e}_m$.

Definition 1. *A function $f : \mathbb{Z}^m \to \mathbb{R}$ is multimodular if for all $\vec{x} \in \mathbb{Z}^m, \vec{u}, \vec{v} \in \Gamma, \vec{u} \neq \vec{v}$,*

$$f(\vec{x} + \vec{u}) + f(\vec{x} + \vec{v}) \geq f(\vec{x}) + f(\vec{x} + \vec{u} + \vec{v}). \tag{4}$$

Because of the connection between multimodular and convex functions, Koole and van der Sluis [9] propose a local search algorithm that searches all the neighbors of a particular point x in the form $\vec{x} + \sum_{\vec{v} \in U} \vec{v}$ where U is a subset of Γ. They show that this local search will lead to an optimal solution of f. Later, Kaandorp and Koole [7] use this concept to obtain an optimal schedule for their scheduling model. In this section,we use the following equivalent form

$$f(\vec{x} + \vec{u}) - f(\vec{x}) \geq f(\vec{x} + \vec{v} + \vec{u}) - f(\vec{x} + \vec{v}) \tag{5}$$

to verify the multimodularity of a function. We note that (5) can be interpreted as the improvement from perturbing $\vec{x}$ by $\vec{u}$ is greater or equal to that from perturbing $\vec{x} + \vec{v}$ by $\vec{u}$. Because $\vec{u}$ and $\vec{v}$ are closely related to the unit vectors $\vec{e}_i$ for some i, we first derive some result on the improvement of f obtained from perturbing $\vec{x}$ by $\vec{e}_i$. This result will be frequently used in this section to help us simplify the proof of multimodularity.

Proposition 1. *For a given schedule S^0, we have*

$$\frac{F(S^0 + \Delta_{i^*,j_1}) - F(S^0)}{F(S^0 + \Delta_{i^*,j_2}) - F(S^0)} = \frac{p_{j_1}}{p_{j_2}} \tag{6}$$

for all $i^* \in I$ and $j_1, j_2 \in J$.

Proof. Assume that W is a patient with no-show probability p_{j_1} being added to slot i^* in schedule S^0 such that the schedule is updated by $S^1 = S^0 + \Delta_{i^*,j_1}$. We use X_i^0 and Y_i^0 to denote the number of arrivals in slot i and the size of overflow from slot i, respectively, for schedule S^0. Then, we define X_i^1 and Y_i^1 for S^1 similarly. Also, we introduce $P_i(i^*)$ to be the conditional probability that the arrival of patient W increases the overflow from slot i to $i+1$ by 1.

Let $\mathcal{W}$ denote the arrival of patient W. Then, from (1), on the condition of $\mathcal{W}$, we have

$$\begin{aligned} F(S^1) - F(S^0) =& rp_{j_1} + \sum_{i\in I} c_i E[Y_i^1 - Y_i^0] \\ =& rp_{j_1} - (1-p_{j_1})\sum_{i\in I} 0 - p_{j_1}\sum_{i\in I} c_i E[Y_i^1 - Y_i^0|\mathcal{W}] \\ =& p_{j_1}(r - \sum_{i\in I} c_i P_i(i^*)). \end{aligned} \tag{7}$$

It can be easily seen that if $P_i(i^*)$ is independent of p_{j_1} for all i, we have $F(S^0 + \Delta_{i^*,j_2}) - F(S^0) = p_{j_2}(r - \sum_{i\in I} c_i P_i(i^*))$. Then, the conclusion follows.

In fact, we observe that the only situation where the arrival of patient W leads to one more patient overflowing from slot i is the case where for each slot k such that $i^* \le k \le i$, the number of patients served is less than or equal to the number patients in slot k in schedule S^0. So, we have

$$P_i(i^*) = \begin{cases} \prod_{k=i^*}^{i} Pr(L_k \le X_k^0 + Y_{k-1}^0) & \text{if } i^* \le i \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Clearly, both $P_i(i^*)$ and $r - \sum_{i\in I} c_i P_i(i^*)$ are independent of the no-show probability of patient W. The desired results follows. □ □

Because the result of Proposition 1 is about the value difference from unit changes, we call the result of (7) the *local perturbation* of a given schedule. Next, we show that the concept of local perturbation can help us simplify the proof of multimodularity significantly as compared to that of Kaandorp and Koole [7]. Because (1) is a maximization problem, we use $\mathfrak{F}$ to denote $-F$ and verify that $\mathfrak{F}$ is multimodular. When $|J| = 1$, we use S_i instead of $S_{i,j}$ and use p to denote the patient show-up probability in our derivation.

Theorem 2. *When $|J| = 1$, $\mathfrak{F}$ is a multimodular function over $\mathbb{Z}^{|I|}$.*

Proof. Because for the simple cases where $\vec{u}$ or $\vec{v}$ is $-\vec{e}_1$ or $\vec{e}_{|I|}$, (5) can easily be proven using the argument similar to the following, we focus on the general case where neither $\vec{u}$ and $\vec{v}$ are standard unit vectors.

Without loss of generality, we let $\vec{u} = \vec{e}_k - \vec{e}_{k+1}$ and $\vec{v} = \vec{e}_l - \vec{e}_{l+1}$ such that $1 \le k < l \le |I| - 1$. To make use of our local perturbation, for any particular schedule S, we define two base schedules S^L and S^R for LHS and RHS of (5) as

$$S^L = S - \vec{e}_{k+1} \tag{9}$$

and

$$S^R = S - \vec{e}_{k+1} + \vec{e}_l - \vec{e}_{l+1} \tag{10}$$

with $S_{k+1} \geq 1$ and $S_{l+1} \geq 1$.

By using S^L and S^R, both LHS and RHS can be interpreted as the difference of two local perturbations. Correspondingly, we use $X_i^L, X_i^R, Y_i^L, Y_i^R$ to denote the number of arrivals in slot i and the number of overflows from slot i in S^L and S^R, respectively. We also use $P_i^L(i_0)$ and $P_i^R(i_0)$ to denote the overflow effect from adding one more patient to slot i_0 in schedule S^L and S^R on the condition of this patient's arrival.

From (8), we have (11) for LHS of (5)

$$\begin{aligned}
\mathfrak{F}(S+\vec{u}) - \mathfrak{F}(S) &= (\mathfrak{F}(S+\vec{u}) - \mathfrak{F}(S^L)) - (\mathfrak{F}(S) - \mathfrak{F}(S^L)) \\
&= p\sum_{i=k}^{|I|} c_i P_i^L(k) - p\sum_{i=k+1}^{|I|} c_i P_i^L(k+1) \\
&= p\{c_k P_k^L(k) + Pr(L_k \leq X_k^L + Y_{k-1}^L)\sum_{i=k+1}^{|I|} c_i \prod_{h=k+1}^{i} Pr(L_h \leq x_h^L + Y_{h-1}^L) - \\
&\quad \sum_{i=k+1}^{|I|} c_i \prod_{h=k+1}^{i} Pr(L_h \leq x_h^L + Y_{h-1}^L)\} \\
&= p\{c_k P_k^L(k) + (Pr(L_k \leq X_k^L + Y_{k-1}^L) - 1)\sum_{i=k+1}^{|I|} c_i P_i^L(k+1)\}.
\end{aligned} \tag{11}$$

Note that, in the first equality of (11), the first sum evaluates overflow characteristics when the patient is added to slot k and the second sum evaluates overflow characteristics when the patient is added to slot $k+1$. Similarly, we have (12) for RHS of (5).

$$\begin{aligned}
\mathfrak{F}(S+\vec{u}+\vec{v}) - \mathfrak{F}(S+\vec{v}) &= (\mathfrak{F}(S+\vec{u}) - \mathfrak{F}(S^R)) - (\mathfrak{F}(S) - \mathfrak{F}(S^r)) \\
&= p\sum_{i=k}^{|I|} c_i P_i^R(k) - p\sum_{i=k+1}^{|I|} c_i P_i^R(k+1) \\
&= p\{c_k P_k^R(k) + (Pr(L_k \leq X_k^R + Y_{k-1}^R) - 1)\sum_{i=k+1}^{|I|} c_i P_i^R(k+1)\}.
\end{aligned} \tag{12}$$

Because S^L and S^R have same number of patients per slot up to and including slot $l-1 \geq k$, we have $P_k^L(k) = P_k^R(k) = Pr(L_k \leq X_k^L + Y_{k-1}^L) = Pr(L_k \leq X_k^R + Y_{k-1}^R)$. Also, because $Pr(L_k \leq X_k^L + Y_{k-1}^L) - 1 \leq 0$, it is sufficient to show that

$$\sum_{i=k+1}^{|I|} c_i P_i^L(k+1) \leq \sum_{i=k+1}^{|I|} c_i P_i^R(k+1).$$

Observe that S^R can be obtained from S^L by reassigning one patient who is in slot $l+1$ to slot l. Let W be a such patient. Then, we can compare (11) and (12) conditioned on the arrival of W, $\mathcal{W}$. Clearly, if $\neg\mathcal{W}$, we have (11) and (12) are same. If $\mathcal{W}$, we have $X_l^R \sim X_l^L + 1$, $X_{l+1}^R \sim X_l^L - 1$, $Y_i^R \sim Y_i^L$ for $i = k, \ldots, l-1$ and $X_i^R \sim X_i^L$ for $i \neq l, l+1$. Furthermore, on the condition of $\mathcal{W}$, we also have

$$Pr(L_l \leq X_l^R + Y_{l-1}^R) = Pr(L_l \leq X_l^L + Y_{l-1}^L + 1) = Pr(L_l \leq X_l^L + Y_{l-1}^L) + Pr(L_l = X_l^L + Y_{l-1}^L + 1).$$

Next, we compare the dynamics of queuing model in S^L and S^R in the case where $L_l \leq X_l^L + Y_{l-1}^L$. Because $X_l^R \sim X_l^L + 1$, $Y_{l-1}^R \sim Y_{l-1}^L$ and $L_l \leq X_l^L + Y_{l-1}^L$, we have $Y_l^R \sim Y_l^L + 1$. Also, because $X_{l+1}^R \sim X_{l+1}^L - 1$, we have $X_{l+1}^R + Y_{l+1}^R \sim X_{l+1}^L + Y_{l+1}^L$ and therefore $Y_{l+1}^R \sim Y_{l+1}^L$. From the fact that $X_j^R \sim X_j^L$ for $j \geq l+1$, we have $X_j^R + Y_{j-1}^R \sim X_j^L + Y_{j-1}^L$ for $j = l+1, \ldots, |I|$. Clearly, from slot $l+1$ to slot $|I|$, the queuing dynamics in schedule S^R and S^L are identical, as shown in Figure 1.

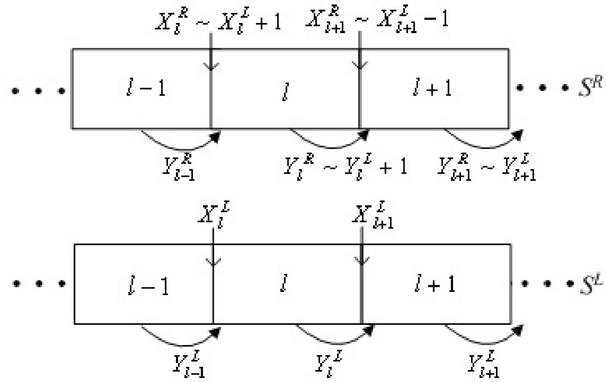

Figure 1: Dynamics of Slot Queuing System in S^R and S^L Conditioned on $\mathcal{W}$ and $L_l \leq X_l^L + Y_{l-1}^L$

Let $H_i^{R^*} = \prod_{j=l+1}^{i} Pr(L_j \leq X_j^R + Y_{j-1}^R)$ given $L_l \leq X_l^L + Y_{l-1}^L$ and $H_i^{R'} = \prod_{j=l+1}^{i} Pr(L_j \leq X_j^R + Y_{j-1}^R)$ given $L_l = X_l^L + Y_{l-1}^L + 1$ for $i \geq l+1$. $H_i^{R^*}$ represents the probability that $\mathcal{W}$ causes additional overflow from slot i and $H_i^{R'}$ represents the probability that $\mathcal{W}$ causes no additional overflow flow from slot i. Further, $H_i^{R^*} Pr(L_l \leq X_l^L + Y_{l-1}^L) = \prod_{j=l}^{i} Pr(L_j \leq X_j^L + Y_{j-1}^L)$ for for $i \geq l+1$. As a consequence, we obtain

$$\begin{aligned}
\sum_{i=k+1}^{|I|} c_i P_i^R(k+1) &= \sum_{i=k+1}^{l-1} c_i P_i^L(k+1) + c_l P_{l-1}^L(k+1)(Pr(L_l \leq X_l^L + Y_{l-1}^L) + Pr(L_l = X_l^L + Y^L + 1)) \\
&\quad + P_{l-1}^L(k+1)(\sum_{i=l+1}^{|I|} c_i H_i^{R^*} Pr(L_l \leq X_l^L + Y_{l-1}^L) + \sum_{i=l+1}^{|I|} c_i H_i^{R'} Pr(L_l = X_l^L + Y^L + 1)) \\
&\geq \sum_{i=k+1}^{l-1} c_i P_i^L(k+1) + c_l P_{l-1}^L(k+1) Pr(L_l \leq X_l^L + Y_{l-1}^L) + \\
&\quad P_{l-1}^L(k+1) Pr(L_l \leq X_l^L + Y_{l-1}^L) \sum_{i=l+1}^{|I|} c_i H_i^{R^*} \\
&= \sum_{i=k+1}^{|I|} c_i P_i^L(k+1).
\end{aligned} \tag{13}$$

The inequality follows from the fact that $c_i \geq 0$ for $i \in I$ and probabilities are non-zero. The last equality follows from the definition of $P_i^L(k)$ in (8). □ □

Since (1) is multimodular for $|J| = 1$, we can apply the local search method by Koole and van der Sluis [9] to obtain an optimal schedule. However, for the general case where $|J| \geq 2$, we have the following result. Note that we express a schedule S in the form of a vector, $[S_{1,1}, \ldots, S_{1,|J|}, \ldots, S_{|I|,1}, \ldots, S_{|I||J|}] \in \mathbb{Z}^{|I||J|}$.

Theorem 3. *The function $\mathfrak{F}$ is not multimodular over $\mathbb{Z}^{|I||J|}$ for $|J| \geq 2$.*

Proof. It is sufficient to show that for some $\vec{u}, \vec{v} \in \Gamma$, (5) does not hold. Let $\vec{u} = \vec{e}_l - \vec{e}_{l+1}$ and $\vec{v} = \vec{e}_{l+k|J|} - \vec{e}_{l+k|J|+1}$ for some l, k such that $1 \leq l, l+k|J|+1 \leq |I||J|$. Denote $j_0 = \lfloor \frac{l}{|J|} \rfloor + 1$. Then, we observe that the operation corresponding to $\vec{u}$ ($\vec{v}$, respectively) is to move one patient of p_{j_0} (p_{j_0+k}, respectively) from slot $i_0 + 1 = l + 1 - (j_0 - 1)|J|$ and to slot i_0.

Similar to our proof for Theorem 2, for a particular schedule S, we define two base schedule S^L and S^R for LHS and RHS of (5) as

$$S^L = S - \vec{e}_{l+1} \tag{14}$$

and

$$S^R = S - \vec{e}_{l+1} + \vec{e}_{l+k|J|} - \vec{e}_{l+k|J|+1}. \tag{15}$$

We also use X_i^L, X_i^R, Y_i^L, and Y_i^R to denote arrival and overflow in S^L and S^R respectively. Comparing S^L and S^R, we observe that $S^R_{i_0,j_0+k} = S^L_{i_0,j_0+k} + 1$, $S^R_{i_0+1,j_0+k} = S^L_{i_0+1,j_0+k} - 1$ and $S^R_{i,j} = S^L_{i,j}$ for all other (i,j). We can easily see that S^R the scheduling resulting when we reassign one patient of type $j_0 + k$ from slot $i_0 + 1$ to slot i_0 in schedule S^L. Let W be a such patient.

Again, we use $P_i^L(i_0)$ and $P_i^R(i_0)$ to denote the overflow effect from adding one more patient to slot i_0 in schedule S^L and S^R conditioned on $\mathcal{W}$. If $\neg\mathcal{W}$, $S^R = S^L$. So, we need only consider the case where W shows up.

Similar to the proof of Theorem 2, for LHS of (5), we have

$$\begin{aligned}\text{LHS} = \mathfrak{F}(S+\vec{u}) - \mathfrak{F}(S) =& c_{i_0} Pr(L_{i_0} \le X^L_{i_0} + Y^L_{i_0-1}) \\ &+ (Pr(L_{i_0} \le X^L_{i_0} + Y^L_{i_0-1}) - 1)(\sum_{i=i_0+1}^{|I|} c_i \prod_{k=i_0+1}^{i} Pr(L_k \le X^L_k + Y^L_{k-1})\end{aligned} \tag{16}$$

For RHS of (5), we have

$$\begin{aligned}\text{RHS} = \mathfrak{F}(S+\vec{u}+\vec{v}) - \mathfrak{F}(S+\vec{v}) =& c_{i_0} Pr(L_{i_0} \le X^R_{i_0} + Y^R_{i_0-1}) \\ &+ (Pr(L_{i_0} \le X^R_{i_0} + Y^R_{i_0-1}) - 1)(\sum_{i=i_0+1}^{|I|} c_i \prod_{k=i_0+1}^{i} Pr(L_k \le X^R_k + Y^R_{k-1})\end{aligned} \tag{17}$$

Next, we compare the value of (16) and (17) conditioned on the physician's performance in slot i_0, i.e. L_{i_0}. Note that under $\mathcal{W}$, we have $X^R_{i_0} + Y^R_{i_0} \sim X^L_{i_0} + Y^L_{i_0} + 1$ and $X^R_{i_0+1} \sim X^L_{i_0+1} - 1$.

Case (i) $L_{i_0} = X^L_{i_0} + Y^L_{i_0-1} + 1$

Because $Pr(L_{i_0} \le X^R_{i_0} + Y^R_{i_0-1}) = 1$, (17) is equal to c_{i_0}. However, (16) is equal to $-\sum_{i=i_0+1}^{|I|} c_i \prod_{k=i_0+1}^{i} Pr(L_k \le X^L_k + Y^L_{k-1})$. So, LHS − RHS < 0 because $c_{|I|} > c_{i_0+1} \ge 0$.

Case (ii) $L_i \le X^L_{i_0} + Y^L_{i_0-1} - 1$

For this case, we have LHS = RHS = c_{i_0}.

Because the probability of both cases is nonzero, we conclude that (5) does not hold when $|J| \ge 2$. □ □

Since the multimodular property does not hold for the general case, we do not expect to obtain an optimal schedule without implementing an exhaustive search. These observations motivate us to develop a local search algorithm that is efficient and can be used to obtain schedules with good quality. We present our study on the solution methodology in Section 4.

4 Local Search Algorithm and Sequential Heuristics for Clinical Scheduling

Because our scheduling model for heterogeneous patients is not multimodular as shown in Section 3, we first propose a local search algorithm to find good schedules in Section 4.1. The main assumption of the proposed local search algorithm is that the set of patients is known in advance. In many situations, clinics do not know the set of patients that should be scheduled and appointment schedules are generated sequentially along with the patient call-in process. So, in Section 4.2, we extend our study to sequential scheduling and propose two sequential scheduling procedures.

4.1 Local Search Algorithm

First, we derive an important property of optimal schedules, that can be used to generate initial schedules. Then, we define the neighborhood of a given schedule and propose dominance rules to reduce the search space in a local search algorithm. Finally, we give the basic steps of the proposed algorithm.

Theorem 4 shows that an optimal schedule prefers patients with lower no-show probabilities.

Theorem 4. *Let S^* be an optimal schedule of* (1)*. Let $j, j_0 \in J$ with $p_j > p_{j_0}$ and $n_j, n_{j_0} > 0$. If $\sum_{i \in I} S^*_{i,j_0} \geq 1$, then $\sum_{i \in I} S^*_{i,j} = n_j$.*

Proof. Let $j, j_0 \in J$ with $p_j > p_{j_0}$ and $n_j, n_{j_0} > 0$. Let S^* be an optimal schedule such that for some $i_0 \in I$, $S^*_{i_0,j_0} > 0$, and suppose $\sum_{i \in I} S^*_{i,j} < n_j$.

Let $\tilde{S} = S^* - \Delta_{i_0,j_0}$. Then, $F(\tilde{S}) \leq F(S^*)$. From the proof of Proposition 1, we have $p_{j_0} r \geq p_{j_0}(\sum_{i \in I, i \geq i_0} c_i P_i(i_0))$ where $P_i(i_0)$ is the probability of overflow from slot i incurred by adding one patient in slot i_0 to $\tilde{S}$ on the condition of this patient's arrival. Consider the schedule $\hat{S} = \tilde{S} + \Delta_{i_0,j}$. Since $p_{j_0} r \geq p_{j_0}(\sum_{i \in I, i \geq i_0} c_i P_i(i_0))$, we have $r \geq (\sum_{i \in I, i \geq i_0} c_i P_i(i_0))$, and thus $p_j r \geq p_j(\sum_{i \in I, i \geq i_0} c_i P_i(i_0))$. From this, we get $p_j(r - (\sum_{i \in I, i \geq i_0} c_i P_i(i_0)) > p_{j_0}(r - \sum_{i \in I, i \geq i_0} c_i P_i(i_0))$. Thus, $F(\hat{S}) > F(S^*)$, a contradiction. $\square$ $\square$

Theorem 4 shows that patients with lower no-show probabilities contribute more than those with higher no-show probabilities. We propose a local search algorithm that schedules patients according to their no-show probabilities. Before explaining the local search algorithm in detail, we define the neighborhood of a given schedule.

Definition 2. *We say schedule S^1 is a neighbor of schedule S^0 if it satisfies the following:*

(1) $S^1 = S^0 \pm \Delta_{i_0,j_0}$ for some $i_0 \in I$ and $j_0 \in J$, i.e. S^1 is obtained by adding/removing one patient of type j_0;

(2) $S^1 = S^0 - \Delta_{i_0,j_0} + \Delta_{i_1,j_0}$ where $i_0 \neq i_1$, i.e. S^1 is obtained by reassign one patient of type j_0 from slot i_0 to slot i_1;

(3) $S^1 = S^0 - \Delta_{i_0,j_0} + \Delta_{i_1,j_0} - \Delta_{i_1,j_1} + \Delta_{i_0,j_1}$ where $i_0 \neq i_1$ and $j_0 \neq j_1$, i.e. S^1 is obtained by switching the slots for two patients of different no-show probabilities.

Note that the size of the neighborhood is $O(\max\{|I||J|, n\}^2)$ where n is the number of patients in the schedule. However, in the course of local search, the size of the effective neighborhood can further be reduced as follows:

Proposition 5. *Let S^0 be a given schedule that is feasible to* (1)*.*

(i) Assume that $S^k = S^0 - \Delta_{i_0,j_k} + \Delta_{i_1,j_k}$ for $k = 1, 2$ and $p_{j_2} > p_{j_1}$. If $F(S^1) > F(S^0)$, then $F(S^2) > F(S^1) > F(S^0)$;

(ii) Assume that $S^k = S^0 - \Delta_{i_0,j_0} + \Delta_{i_1,j_0} - \Delta_{i_1,j_k} + \Delta_{i_0,j_k}$ for $k = 1, 2$, and $i_0 < i_1$. If $F(S^1) > F(S^0)$ and $p_{j_2} > p_{j_1} > p_{j_0}$, then $F(S^2) > F(S^1) > F(S^0)$; if $F(S^1) > F(S^0)$ and $p_{j_2} < p_{j_1} < p_{j_0}$, then $F(S^2) > F(S^1) > F(S^0)$.

Proof. The is very straightforward to show (i) using the results of Proposition 1 and Theorem 4. Here, we focus on the more difficult proof of (ii).

Let x and y be two patients of types j_0 and j_1, respectively. Let $S|A\overline{B}$ denote schedule S on the condition that all patients in A arrive as scheduled and all patients in B are no-shows. The expected profit of schedules S^1 and S^2 are calculated by conditioning the no-show scenarios of patients.

$$F(S^0) = (1 - p_{j_0})(1 - p_{j_1})F(S^0|\overline{xy}) + p_{j_0}(1 - p_{j_1})F(S^0|x\overline{y}) + (1 - p_{j_0})p_{j_1}F(S^0|\overline{x}y) + p_{j_0}p_{j_1}F(S^0|xy)$$

$$F(S^1) = (1 - p_{j_0})(1 - p_{j_1})F(S^1|\overline{xy}) + (1 - p_{j_1})p_{j_0}F(S^1|x\overline{y}) + (1 - p_{j_0})p_{j_1}F(S^1|\overline{x}y) + p_{j_0}p_{j_1}F(S^1|xy)$$

Note that $(S^0|\overline{xy}) = (S^1|\overline{xy})$ and $(S^1|xy) = (S^0|xy)$. Let $S^* = S^0 - \Delta_{i_0,j_0} - \Delta_{i_1,j_1}$ and $P_i^*(k)$ be the overflow probability from slot k defined in (8). It can be easily seen that $(S^0|\overline{x}y) = (S^* + \Delta_{i_1,j_1}|y)$ and $(S^0|x\overline{y}) = (S^* + \Delta_{i_0,j_0}|x)$. Similar result holds for S^1. Then, we have the following:

$$
\begin{aligned}
F(S^1) - F(S^0) =& p_{j_0}(1-p_{j_1})F(S^* + \Delta_{i_1,j_0}|x) + (1-p_{j_0})p_{j_1}F(S^* + \Delta_{i_0,j_1}|y) \\
& - \{p_{j_0}(1-p_{j_1})F(S^* + \Delta_{i_0,j_0}|x) + (1-p_{i_0})p_{j_1}F(S^* + \Delta_{i_1,j_1}|y)\} \\
=& p_{j_0}(1-p_{j_1})\{F(S^* + \Delta_{i_1,j_0}|x) - F(S^* + \Delta_{i_0,j_0}|x)\} \\
& + (1-p_{j_0})p_{j_1}\{F(S^* + \Delta_{i_0,j_1}|y) - F(S^* + \Delta_{i_1,j_1}|y)\} \\
=& p_{j_0}(1-p_{j_1})\{\sum_{i\in I} c_i P_i^*(i_0) - \sum_{i\in I} c_i P_i^*(i_1)\} \\
& + (1-p_{j_0})p_{j_1}\{\sum_{i\in I} c_i P_i^*(i_1) - \sum_{i\in I} c_i P_i^*(i_0)\} \\
=& \left\{ \sum_{i\in I} c_i\{P_i^*(i_0) - P_i^*(i_1)\}\right\}\left\{p_{j_0}(1-p_{j_1}) - (1-p_{j_0})p_{j_1}\right\}
\end{aligned}
$$

For the case where $p_{j_0} < p_{j_1}$, we observe that the second term in the last equality is always negative because $p_{j_0} < p_{j_1}$ and $1-p_{j_1} < 1-p_{j_0}$. Since $F(S^1)-F(S^0) > 0$, the first term of last equality should be negative. The first term is independent of no-show probabilities of x, y and $p_{j_0}(1-p_{j_2}) - (1-p_{j_0})p_{j_2} < p_{j_0}(1-p_{j_1}) - (1-p_{j_0})p_{j_1}$. Therefore, $F(S^2) > F(S^1) > F(S^0)$.

Similarly, we can prove the desired results for the case where $p_{j_0} > p_{j_1}$. □ □

We observe that the results of Proposition 5 provide guidelines such that local movements can be implemented according to patient no-show probabilities. In particular, using Theorem 4 and Proposition 5, we can obtain better schedules with reduced computational effort. Next, we describe our local search algorithm in detail.

For a given schedule S, we define $\overline{j}_i = \arg\max\{p_j : S_{i,j} \geq 1\}$, $\underline{j}_i = \arg\min\{p_j : S_{i,j} \geq 1\}$ and j^* as the patient type with the lowest no-show probability to be scheduled. From Definition 2, we note that there are four types of neighbors of S obtained by the following local movements: add, remove, reassign and switch, which are numbered by $1, \ldots, 4$ respectively.

Algorithm 1.

(1) Initialization: $S = \emptyset$.

(2) Local Search:
For $l = 1$ *to* 4

if $l = 1$**: (neighbors obtained by adding)** $F_1^* = \max\{F(S + \Delta_{i,j^*}) : i \in I\}$;

if $l = 2$**: (neighbors obtained by removing)** $F_2^* = \max\{F(S - \Delta_{i,\overline{j}_i}) : i \in I\}$;

if $l = 3$**: (neighbors obtained by reassigning)** $F_3^* = \max\{F(S - \Delta_{i,\overline{j}_i} + \Delta_{k,\overline{j}_i}) : i, k \in I\}$;

if $l = 4$**: (neighbors obtained by switching)** $F_4^* = \max\{F_4^1, F_4^2\}$ *where*

$$F_4^1 = \max\{F(S - \Delta_{i,\overline{j}_i} + \Delta_{k,\overline{j}_i} - \Delta_{k,\underline{j}_k} + \Delta_{i,\underline{j}_k}) : 1 \leq i < k \leq |I|, p_{\overline{j}_i} > p_{\underline{j}_i}\};$$

and

$$F_4^2 = \max\{F(S - \Delta_{i,\underline{j}_i} + \Delta_{k,\underline{j}_i} - \Delta_{k,\overline{j}_k} + \Delta_{i,\overline{j}_k}) : 1 \leq i < k \leq |I|, p_{\overline{j}_i} < p_{\underline{j}_i}\}.$$

end for

(3) Find the best schedule, S^**, in the neighborhood of* S*, i.e.* $S^* = \arg\max\{F_l^* : l = 1, \ldots, 4\}$.

(4) If $F(S^*) \geq F(S)$, update current schedule $S = S^*$ and go back to Step (2). Otherwise, go to Step (5).

(5) Return the local optimal schedule S.

In Algorithm 1, the number of neighbors search for a schedule need to be searched is $O(\max\{|I|^2, n\})$, which is smaller than the actual neighborhood size.

4.2 Sequential Scheduling Methods

As mentioned earlier, many clinics generate schedules in a sequential fashion. Typically, a patient calls requesting an appointment. The scheduler will either add the patient to an existing schedule and give an appointment time or reject the patient. Muthuraman and Lawley [15] propose a myopic scheduling method, which sequentially assigns calling patients to the slot that most increases the expected profit of the resulting schedule. It is called myopic since it does not take the possibility of future call-ins into account when making the current assignment. Patients are added to a schedule until the expected total profit starts decreasing. Although the authors consider heterogeneous patients, their method does not differentiate patients according to their no-show probabilities while generating schedules. However, better schedules can be generated by considering the no-show probabilities of patients. We propose two sequential scheduling algorithms that do this by using the properties explained in Section 4.1.

Let S^0 be a fixed schedule for $n-1$ patients and assume that we need to schedule the n^{th} patient of type j. The patient will be inserted into the schedule if adding this patient increases the objective function value. Corollary 6 shows that the decision of accepting (or rejecting) a patient is independent of the patient's no-show probability.

Corollary 6. *For the myopic scheduling method in [15], the decision to accept (or reject) the n^{th} patient of type j is independent of p_j.*

Proof. The myopic scheduling method in [15] is as follows.

Step 1. Set $S^0_{i,j} = 0$ for all $i \in I$ and $j \in J$.

Step 2. Wait for k^{th} patient call.

Step 3. The k^{th} patient calls in and the patient is of type j_0.

Step 4. Compute $F(S^0 + \Delta_{i,j_0})$ for $i \in I$ and set $i^* = \arg\max\{F(S^0 + \Delta_{i,j_0}) : i \in I\}$.

Step 5. If $F(S^0 + \Delta_{i^*,j_0}) > F(S^0)$, assign the k^{th} patient to slot i^* and update $S^0 = S^0 + \Delta_{i^*,j_0}$ and go to Step 2. Otherwise, stop.

Assume that the n^{th} patient is assigned to slot i_n. Let $S = S^0 + \Delta_{i_n,j}$. From Proposition 1, we have $F(S) - F(S^0) = p_j(r - \sum_{i \in I} c_i P_i(i_n))$ where $P_i(i_n)$ is independent of p_j and can be computed without the n^{th} patient. Then, let i_n^* denote the slot index that yields the minimal $\sum_{i \in I} c_i P_i(i_n)$. It is clear that if $\sum_{i \in I} c_i P_i(i_n^*) \leq r$, we can assign patient n to slot i_n^* to increase the expected total profit. Otherwise, patient n will be rejected. Therefore, the decision on the n^{th} patient is independent of p_j. □ □

Based on the results in Proposition 1 and Corollary 6, it is anticipated that overbooking many patients with high no-show probabilities cannot provide the most desirable results. One major drawback of the myopic scheduling method Muthuraman and Lawley [15] is that the objective function depends on the call-in sequence. Different call-in sequences generate schedules which have high variability in the objective function. In order to show the effect of call-in sequences, we consider two sequences. In the first sequence, there are sufficiently many patients of type j_1 before any patient of type j_2 ($p_{j_1} > p_{j_2}$). In the second sequence, there are sufficiently many patients of type j_2 before any patient of type j_1. We apply the myopic scheduling algorithm [15] to generate schedules S_1 and S_2, respectively. Clearly, S_1 has patients of type j_1 and S_2 has patients of type j_2. Figure 2 shows $\frac{F(S_1)}{F(S_2)}$ as a function of $p_{j_1} - p_{j_2}$ for

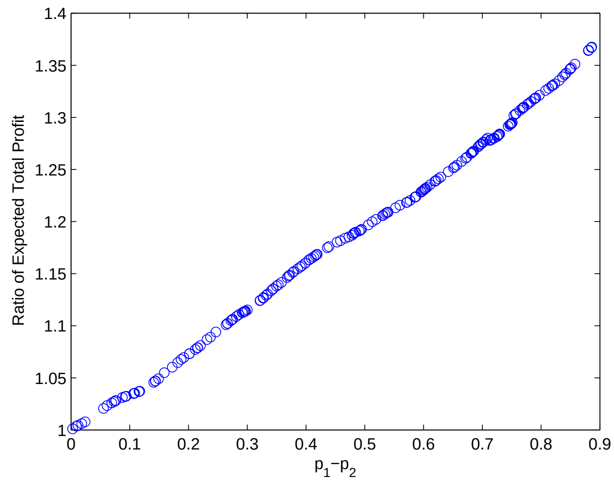

Figure 2: Ratio of Expected Total Profit $F(S_1)/F(S_2)$ vs. $p_{j_1} - p_{j_2}$

200 pairs of randomly generated call-in sequences. The difference between $F(S_1)$ and $F(S_2)$ increases as $p_2 - p_1$ increases.

One may think that it is rare to have all patients at the beginning of the sequence with high no-show probabilities. However, it is commonly observed that patients who make reservations at earlier times tend to have higher no-show probabilities and they often use up the physician's capacity before patients with low no-show probabilities can be scheduled [13, 14, 16]. Therefore, a sequential scheduling method, which accepts all patients regardless of their no-show probabilities, may not generate good schedules in terms of expected profit.

In sequential scheduling, there is an existing schedule and the patients who will call-in should be scheduled or rejected. Assume that S^0 is the existing schedule and there are $\bar{n}_j$ patients for all $j \in J$ that should be scheduled or rejected. The following is the revised model of (1) in the sequential scheduling setting, which adds new patients to an existing schedule:

$$\begin{aligned} \max \ G(S) =& r \sum_{i \in I} \sum_{j \in J} S_{i,j} p_j - \sum_{i \in I} c_i \sum_k k R_{i,k} \\ \text{s.t.} & \sum_{i \in I} S_{i,j} - \sum_{i \in I} S^0_{i,j} \le \bar{n}_j \\ & S_{i,j} - S^0_{i,j} \ge 0 \\ & S_{i,j} \in \mathbb{Z} \forall i \in I, j \in J \end{aligned} \tag{18}$$

We give the following result as a corollary of Theorem 4 which can be easily proven using similar argument to that of Theorem 4.

Corollary 7. *Assume that S^* is an optimal schedule to* (18)*. For $j, j_0 \in J$ with $p_j > p_{j_0}$, either $\bar{n}_j = \sum_{i \in I}(S^*_{i,j} - S^0_{i,j})$ or $\sum_{i \in I}(S^*_{i,j_0} - S^0_{i,j_0}) = 0$, i.e. $S^*_{i,j_0} = S^0_{i,j_0}$ for all $i \in I$.*

Corollary 7 shows that it is always better to include patients of low no-show probabilities into an existing schedule before capacity limit is reached. Also, by Proposition 1 and Corollary 7, our local search algorithm still works for any given existing schedule and patient set. From Corollary 7 and the observation in Figure 2, it is easy to see that limiting the number of patients with high no-show probabilities in the schedule will be an effective way to improve its performance. Following this line, we propose two sequential scheduling methods: the restricted myopic scheduling and the forecasting-based scheduling methods.

The basic idea of the restricted myopic scheduling method is using upper bounds to restrict the number of patients with high no-show probabilities in the schedule. We set upper bounds, $\overline{B}_j$, on the number of patients of type j for $j \neq 1$ in the schedule such that $\overline{B}_1 \geq \overline{B}_2 \geq \cdots \geq \overline{B}_{|J|}$. Let b_j be the number of patients of type j in the current schedule. The basic steps of the restricted myopic scheduling method are as follows:

Restricted Myopic Sequential Scheduling Method 1.

Step 1. *Set* $b_j = 0$ *and* $S^0_{i,j} = 0$ *for all* $i \in I$ *and* $j \in J$.

Step 2. *Wait for* k^{th} *patient call.*

Step 3. *The* k^{th} *patient call occurs and is of type* j_0.

Step 4. *If* $b_{j_0} + 1 > \overline{B}_{j_0}$, *do not accept patient* k *and go to Step 2.*

Step 5. *Perform the traditional myopic scheduling algorithm to compute the best slot* i_0 *for patient* k.

Step 6. *If* $F(S^0 + \Delta_{i_0,j_0}) < F(S^0)$, *i.e. adding patient* k *decreases the expected total profit, stop. Otherwise, update* $b_{j_0} = b_{j_0} + 1$ *and* $S^0 = S^0 + \Delta_{i_0,j_0}$ *and go to Step 2.*

The restricted myopic scheduling method is very simple to implement. However, this method does not consider potential call-ins in the future. We propose another sequential scheduling algorithm, which considers forecasted patient requests from current time to the appointment day. Specifically, for each call-in patient, we generate a schedule from the existing schedule considering the forecasted future patients. We believe that the information about anticipated patients contributes to the scheduling algorithm in two ways: (i) this information can limit the number of patients with high no-show probabilities in the final schedule and (ii) slot allocation decisions anticipate possible future patient call-ins.

In this study, we simply use average historical data to predict future patient demand. Assume that we are generating forecasted patient demand for a day that is T minutes ahead from current time. We can use the average of q pieces of historical patient demand data that happened T or less minutes before virtual appointment days. We denote the forecasted patient demand by $\bar{n}_j$ for $j \in J$. To avoid overestimating future patient arrivals, we may discount our forecasting by α with $0 \leq \alpha \leq 1$. When $\alpha\bar{n}_j$ is not an integer, we can round it to the nearest integer.

Forecasting-based Sequential Scheduling Method 1.

Step 1. *Set* $S^0_{i,j} = 0$ *for all* $i \in I$ *and* $j \in J$.

Step 2. *Wait for* k^{th} *patient call.*

Step 3. *The* k^{th} *patient calls in and the patient is of type* j_0.

Step 4. *Predict future patient demand and obtain* $\bar{n}_j$ *for* $j \in J$.

Step 5. *Perform the scheduling algorithm that considers requests* $\alpha\bar{n}_1, \ldots, \alpha\bar{n}_{j_0-1}, \alpha\bar{n}_{j_0}+1, \alpha\bar{n}_{j_0+1}, \ldots, \alpha\bar{n}_{|J|}$ *to generate a schedule* S^* *from* S^0 *for* (18).

Step 6. *If* $\exists i \in I$ *such that* $S^*_{i,j_0} - S^0_{i,j_0} \geq 1$, *assign the* k^{th} *patient to slot* i *and update* $S^0 = S^0 + \Delta_{i,j_0}$.

Step 7. *If* $S^*_{i,j} - S^0_{i,j} = 0$ *for* $i \in I$ *and* $j \in J$, *stop. Otherwise, go to Step 2.*

When $\alpha = 0$, the forecasting-based scheduling method reduces to the myopic scheduling method in [15].

Comparing these two sequential scheduling methods, the restricted myopic method is conservative because it mostly considers available patient information while the forecasting based method is aggressive since it heavily uses predicted information on potential patient calls. Note that the successful application of both of them requires that the physician has enough patient demand which is always the case in practice. In Section 5, we perform a computational study to compare the proposed algorithms with the traditional myopic scheduling algorithm in [15].

5 Computational Study

We perform a computational study to test the performance of proposed algorithms. We consider three experimental settings. In the first setting, we show the effect of considering heterogeneous patients instead of homogeneous patients. In the second setting, we compare the proposed sequential scheduling algorithms with the traditional myopic scheduling algorithm in [15]. In the last setting, we analyze the effect of overflow cost on expected profit.

Throughout our experiments, we assume that a clinic session is 4 hours and partitioned into 8 equal length slots. The service rate λ, which is equal to 2, is constant during the session. Unless explicitly mentioned, $r = 100$, $c_i = 40$ for $i \neq |I|$ and $c_{|I|} = 200$.

5.1 Homogeneous versus Heterogeneous Patients

A major contribution of this study is that the variability of no-show rates is taken into consideration while designing the scheduling algorithms. Algorithm 1 is used to schedule heterogeneous patients. The heuristic algorithm proposed by Kaandorp and Koole [7] is used to schedule homogeneous patients. We consider three types of patients. p_2 is set to 0.5, and p_1 and p_3 are randomly generated such that $(p_1 + p_3)/2 = 0.5$. We assume equal number of patients in each group ($n_1 = n_2 = n_3 = n$). We consider different values for n ($n = 1, \ldots, 12$) to analyze the effect of variance on expected total profit for different population sizes. The variance of no-show rates is derived from the variance of p_1, p_2 and p_3.

Figure 3 shows the results of 400 randomly generated problems. We observe that the expected profit obtained from the heterogeneous scheduling model dominates the one obtained from the homogeneous scheduling model for all population sizes. The impact of variance of no-show rates on expected profit becomes more significant as the number of patients increases. Figure 4 highlights the difference for 6 and 12 patients. The consideration of heterogeneous patients leads to greater improvements when variance is greater. Especially, when $n = 12$, the improvement on total expected profit could reach up to 20%. Algorithm 1 schedules more patients with low no-show probabilities. However, the total number of patients scheduled is less. As a consequence, the variance in expected profit is less than the one obtained by homogeneous model.

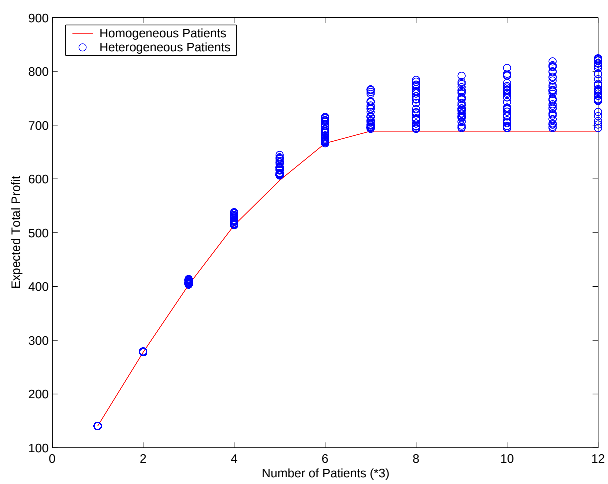

Figure 3: Expected Total Profit vs. Number of Patients

5.2 Sequential Scheduling

We compare the proposed sequential scheduling methods with the traditional myopic scheduling method by Muthuraman and Lawley [15]. We set $|J| = 2$, $p_1 = 0.8$ and $p_2 = 0.2$. We first randomly generate 100

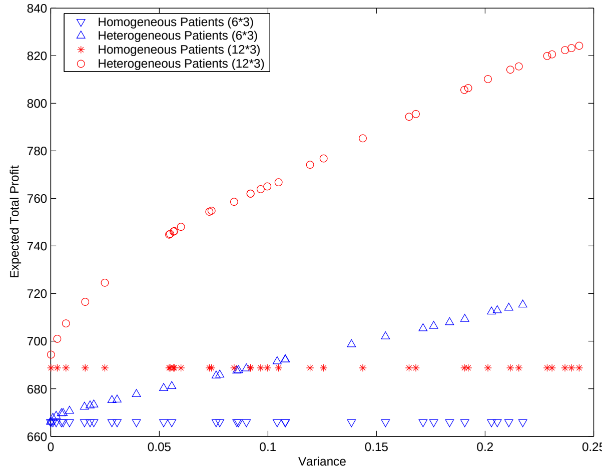

Figure 4: Impact of Variance for 6×3 and 12×3 Patients

call-in sequences that span over 30 days. We assume that the call-in rate increases as the call-in time gets closer to the appointment time. At the beginning, the call-in rate for patients of type j_1 is smaller than the rate for patients of type j_2. As time goes on, it increases and finally becomes larger than that for patients of type j_2. Let $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ be the arrival rates of patients of types j_1 and j_2, respectively. Specifically, once a call-in of type j_1 is generated, we update $\tilde{\lambda}_1 = \gamma\tilde{\lambda}_1$ where γ is a randomly generated positive number that is larger than 1. Similarly, we keep updating $\tilde{\lambda}_2$ by a randomly generated number that is less than 1. We generally set parameters in a way such that the number of expected call-ins of both type 1 and 2 throughout 480×30 mins are more than the number of expected services, $\lambda \times 8 = 16$. In our experiments, we set the initial values $\tilde{\lambda}_1 = \frac{1}{600}$ and for $\tilde{\lambda}_2 = \frac{1}{300}$. Random numbers are generated from $(1, 1 \pm 0.05]$ respectively.

First, we consider the restricted myopic scheduling algorithm. Figure 5 shows the expected total profit for restricted myopic scheduling algorithm (for both $\overline{B}_2 = 4$ represented by $*$, and $\overline{B}_2 = 8$ represented by $\times$) and the traditional myopic method for 100 randomly generated sequences. The results from the restricted myopic method always dominates those from the traditional myopic scheduling method of [15]. The results from the proposed method are very stable (with less variance), while the traditional myopic method gives results with high variance. As expected, the proposed method obtains better results when the upper bound on the number of patients of type j_2 is lower.

To predict potential patient call-ins for a call-in sequence, we randomly choose 4 other sequences to predict numbers of call-ins and set $\alpha = 0.5$ to discount risk. Figure 6 shows the expected total profit for the forecasting-based sequential scheduling method (represented by $*$) and the traditional myopic scheduling method for 100 randomly generated call-in sequences. The results show the advantage of using forecasted patient demand to generate schedules sequentially. In cases where many potential patients with low no-show probabilities are expected, it would be wiser to reserve the capacity for these patients rather than adding patients with high no-show probabilities into the schedule. Obviously, this argument justifies the open-access scheduling model in which the capacity is kept until the day before the appointment day or the appointment day since their no-show probabilities are low in those days.

In fact, both our theoretical analysis in Sections 2-4 and our computational study on the performance of the restricted myopic scheduling method and the forecasting-based scheduling method show that patients with low no-show probabilities should not be restricted by open access model. Actually, our models show that allowing patients with low no-show probabilities to make appointments ahead is effective.

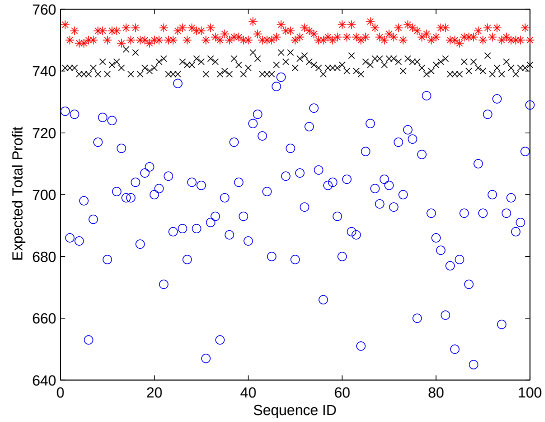

Figure 5: Restricted Myopic Scheduling Method vs Myopic Scheduling Method

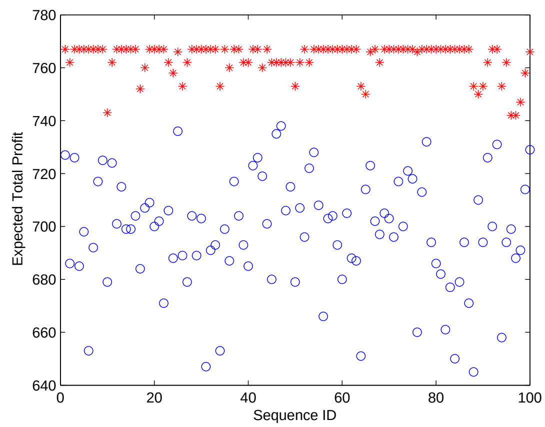

Figure 6: Forecasting-base Method vs Traditional Myopic Method

5.3 Cost of Patient Waiting Time

The expected revenue and overtime cost can typically be estimated by health care practitioners. However, the value of patient waiting time is determined subjectively. Since the cost associated with patient waiting time is not an actual cost paid by the clinics, it is important to investigate the effect of different cost values on scheduling and expected profit. As discussed in Section 2, patient waiting time is directly related to the size of overflow between slots. So, we control the value of patient waiting time through changing the value of c_i for $i \neq |I|$. Similarly, the cost of physician overtime can be controlled through changing the value of c_I. In our experiment, we assume $c_i = c_j$ if $i \neq j$ for $i, j \in I \backslash \{|I|\}$.

We consider two schedules generated by the traditional myopic scheduling method from two calling sequences described in Section 4.2 because of their simple patient structures. For these two sequences, we set $p_1 = 0.8$ and $p_2 = 0.2$. Figure 7 shows the effect of c_i on expected total profit for both sequences. As overflow cost (c_i) increases, expected total profit decreases for both schedules. However, the expected profit of the schedule for the second call-in sequence decreases faster than that of the schedule for the first sequence. When c_i is small (which means that the physician time is more valuable than patient waiting time), the performances of two schedules are close to each other. In such cases, differentiating patients by their no-show probabilities is not very beneficial. When c_i is increasing, the difference between the two schedules becomes larger. If the patients' waiting time is considered as a significant part of the performance measure, the number of patients with high no-show probabilities should be restricted.

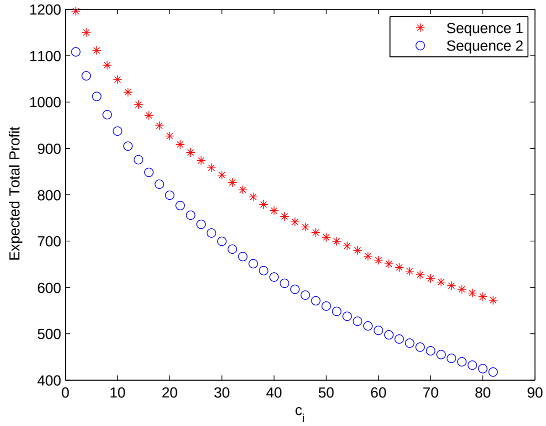

Figure 7: Expected Total Profit vs c_i

Laganga and Lawrence [11] mention that the *net overbooking utility*, which is the expected net return generated by overbooking, is larger in the case where patients have high no-show probabilities than in the case where patients have low no-show probabilities. They further mention that this phenomenon is more significant when the cost of patient waiting time and physician overtime are high. According to our computational results in Figure 7 and Figure 8, the expected total profit of schedules generated using overbooking decrease when cost of patient waiting time and physician overtime increases. Furthermore, the speed of decrease is faster in the case where patients have higher no-show probabilities. These results indicate that overbooking can compensate the loss from patient no-shows to some extent. However, reducing patients' no-show rates should have higher priority than applying overbooking.

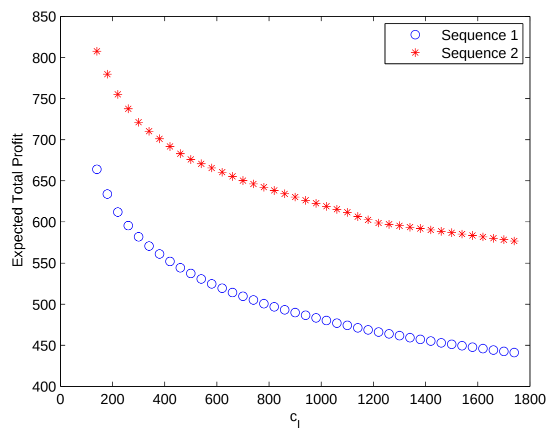

Figure 8: Expected Total Profit vs c_I

6 Concluding Remarks and Managerial Insights

Various scheduling models with overbooking have been proposed to help health care providers alleviate the negative effects of patient no-shows. However, to the best of our knowledge, all existing studies either assume that patients are homogeneous in terms of their no-show probabilities, or do not consider the impact of different no-show rates on general performance measures. In this paper, we systematically study a clinical scheduling model with overbooking for a set of heterogeneous patients, i.e. their no-show probabilities are different. We prove that, unlike the overbooking model for homogeneous patients, the model for heterogeneous patients is not multimodular. It is very difficult to obtain an optimal schedule since the local optimal solution is not guaranteed to be global optimal. We develop a guided local search algorithm based on the properties of an optimal schedule. We observe that homogeneous overbooking models using the mean value of show-up probabilities are not enough to build high quality schedules. The variance of no-show probabilities have a significant impact on the performance of overbooked schedules. Further, we show the disadvantages of the traditional myopic sequential scheduling method and propose two improved sequential scheduling algorithms that give better schedules.

Next, we provide some managerial insights based on our theoretical derivations and computational results. These insights can help health care practitioners better manage clinic scheduling when patients' np-show probabilities are different but can be estimated.

1. Clustering patients according to their no-show probabilities and using our clinical scheduling methods for heterogeneous patients will help to build schedules with better performances.

2. Patients with low no-show probabilities are always preferable in schedule generation. This result justifies the open-access scheduling approach, because no-show probabilities increase as the interval between the call-in time and appointment time increases. However, appointments for patients with low no-show probabilities can be made earlier.

3. Overbooking is beneficial for open-access scheduling systems, because it reduces fluctuations in clinic workload and helps to control demand over time.

4. The traditional myopic scheduling method proposed in [15] performs well when there is enough patient with low no-show probabilities at the beginning of the call-in sequence. Its performance can be improved significantly by restricting the number of patients with high no-show probabilities in the schedule or using the information of potential patient call-ins.

5. If costs of patient waiting time and physician overtime are high, few patients with high no-show probabilities should be scheduled.

6. To reduce overtime cost, patients with low no-show probabilities should be assigned into early slots and patients with high no-show probabilities should be assigned to later slots.

Future research directions include extending our research to multiple physician (server) systems, since several physicians collaborate and share the same set of patients. Another direction is developing scheduling methods considering cancelations and unpunctual arrivals. Finally, as [11, 15] point out, overbooking models can also be used in other appointment-based service systems such as law offices, counseling centers and photo studios.

References

[1] E. Altman, B. Gaujal, and A. Hordijk. Multimodularity, convexity and optimization properties. *Mathematics of Operations Research*, 25:324–347, 2000.

[2] T. Bodenheimer and K. Grumback. *Understanding Health Policy: A Clinical Approach*. Lange Medical Books / McGraw-Hill,Medical Publishing Division, New York, third edition edition, 2002.

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

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

[5] S. Garuda, R. Javalgi, and V. Talluri. Tackling no-show behavior: A market-driven approach. *Health Marketing Quarterly*, 15:25–44, 1998.

[6] B. Hajek. Extremal splitting of point processes. *Mathematics of Operations Research*, 10:543–556, 1985.

[7] G. Kaandorp and G. Koole. Optimal outpatient appointment scheduling. *Health Care Management Science*, 10:217–229, 2007.

[8] S. Kim and R. Giachetti. A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 36:1211–1219, 2006.

[9] G. Koole and E. van der Sluis. Optimal shift scheduling with a global service level constraint. *IIE transactions*, 35:1049–1055, 2003.

[10] L. Laganga and S. Lawrence. Clinic overbooking to improve patient access and provider productivity. *Decision Sciences*, 38:251–276, 2007.

[11] L. Laganga and S. Lawrence. An appointment overbooking model to improve client access and provider productivity. Technical report, University of Colorado at Boulder, 2007.

[12] L. Liu and S. Liu. Dynamic and static job allocation for multi-server systems. *IIE Transactions*, 30:845–854, 1998.

[13] M. Murray and DM. Berwick. Advanced access: reducing waiting and delays in primary care. *JAMA: The Journal of the American Medical Association*, 289:1035–1040, 2003.

[14] M. Murray and C. Tantau. Same-day appointments: exploding the access paradigm. *Family Practice Management*, 7:45–50, 2000.

[15] K. Muthuraman and M. Lawley. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *to appear in IIE Transactions*, 2008.

[16] G. Randolph, M. Murray, J. Swanson, and P. Margolis. Behind schedule: improving access to care for children one practice at a time. *Pediatrics*, 113:230–237, 2004.

[17] C. Rust, N. Clark, W. Clark, D. Jones, and W. Wilcox. Patient appointment failures in pediatric resident continuity clinics. *Archives of Pediatrics and Adolescent Medicine*, 149(6):693–695, 1995.

[18] W. Shonick and B. Klein. An approach to reducing the adverse effects of broken appointments in primary care systems: Development of a decision rule based on estimated conditional probabilities. *Medical Care*, 15:419–429, 1977.