Sequential Clinical Scheduling with Service Criteria

Ayten Turkcan[a], Bo Zeng[b], Kumar Muthuraman[c], Mark Lawley[d]

[a] *Purdue University, Regenstrief Center for Healthcare Engineering, West Lafayette, IN 47907*
[b] *University of South Florida, Department of Industrial and Management Systems Engineering, Tampa, FL 33620*
[c] *University of Texas, McCombs School of Business, Austin, TX 78712*
[d] *Purdue University, Weldon School of Biomedical Engineering, West Lafayette, IN 47907*

Abstract

This study investigates sequential appointment scheduling with service criteria. It uses a constraint-based approach with service criteria bounded in a constraint set in contrast to the more typical weighted linear objective function. Properties are derived and a sequential scheduling algorithm is developed. Fairness properties of generated schedules are considered in detail, where fairness is the uniformity of performance across patients. New unfairness measures are proposed and used to capture the inequity among patients assigned to different slots. Other criteria such as expectation and variance of patient waiting time, queue length, and overtime are also considered. The fairness/revenue tradeoff is investigated as is the flexibility of the constraint-based approach in handling unavailable time periods.

Keywords: Multiple criteria clinical scheduling, moment-based constraints, overbooking, Pareto optimal set

1. Introduction

The majority of U.S. patient care is delivered through outpatient clinics (U.S. Census Bureau). Access to these clinics is provided through appointment scheduling. Many studies indicate that well-designed appointment scheduling can increase access to care, smooth clinic operations, increase patient satisfaction, improve quality of care, and reduce overall costs (AHRQ (2007), Savin (2006)). Gupta and Denton (2008) discuss the characteristics of appointment scheduling in primary care and specialty clinics. Appointment scheduling is a sequential process where the demand for appointments is realized online. When a patient calls for an appointment, the patient must be worked into a pre-existing (perhaps partially constructed) schedule without changing the appointment times of other patients. This must be done and the selected appointment time given to the patient before the call terminates. There are many

Email addresses: `aturkcan@purdue.edu` (Ayten Turkcan), `bzeng@eng.usf.edu` (Bo Zeng), `kumar.muthuraman@mccombs.utexas.edu` (Kumar Muthuraman), `malawley@purdue.edu` (Mark Lawley)

complicating factors that affect the performance of appointment scheduling methods such as patient no-show, variability in arrival and service times, and physician and patient preferences.

Ho and Lau (1992) show that, among three factors (no-show, service time variability, and number of patients per clinic session), no-show is the major factor that affects clinic performance. Cayirli and Veral (2003) mention that no-show rates range from 5% to 30%. Many factors can affect patient no-show, such as patient condition, distance to the clinic, demographics, insurance, and the lead time between the call and appointment time. Clinics try to reduce no-show with appointment reminders or by using open access scheduling where patients can be seen in 24-48 hours. In contrast, overbooking attempts to reduce the effect of no-shows on clinic operations rather than reducing no-show. Researchers typically develop a linear objective function that includes elements of revenue for seeing patients combined with costs associated with patient waiting, clinic overtime, and/or physician idle time. Service times are typically assumed to be either deterministic or exponential, and most studies assume a single no-show probability for all patients.

More specifically, Kim and Giachetti (2006) determine the number of patients to book in a given session to maximize clinic profit, modeled as the weighted sum of revenue, cost of overtime, and a penalty for patients who leave without being seen. They assume deterministic service times and a single no-show probability. Laganga and Lawrence (2007a, 2007b) balance the utility of serving patients with the costs of waiting time and overtime. They assume deterministic service times and a single no-show probability and use simulation to conclude that their approach works best with high no-show probability and low service time variances. Kaandorp and Koole (2007) assume exponential service times and attempt to minimize a linear combination of patient waiting time, physician idle time, and overtime. To compute the schedule, they develop a local search algorithm, which seeks incremental improvement by shifting patients between adjacent slots (note that this assumes the set of patients to be scheduled is known beforehand and thus cannot be classified as sequential). They establish that this approach yields a global optimum when the objective is multi-modular (which is true when there is only one no-show probability). Hassin and Mendel (2008) consider the problem of scheduling a given number of patients with homogenous no-show. The objective is to minimize the cost of customer waiting and server availability. Their model indicates that appointments should be close together at the beginning and end of a schedule. Green and Savin (2008) balance physician capacity and patient panel size. They assume that the patient no-show rate is a non-decreasing function of the appointment backlog (queue length). They show that queue length and the probability to get same day appointment depend on the patient panel size for deterministic service times.

Most of the studies considering no-show assume that all patients have the same no-show rate. However, patient no-show behavior can be estimated at the individual patient level. Many studies investigate

the relationship between no-show behavior and patient and provider characteristics, environmental factors, health status, lead time, etc. (Daggy et al. (2009)). Muthuraman and Lawley (2008) is the first study that explicitly considers heterogeneous patients, i.e. patients with different no-show rates, in a sequential scheduling environment. They develop a recursive computation method and a sequential scheduling algorithm based on a myopic policy (one that does not consider any information about the unknown future call-in sequence) to maximize the weighted average of expected revenue, patient slot overflow (a surrogate for waiting time), and overtime. The work is extended by Chakraborty et al. (2009a) for general service time distributions, thus relaxing the exponential service time assumption. Zeng et al. (2010) shows that the objective is not multi-modular for heterogeneous no-show (implying that finding an optimal schedule for a given set of patients is computationally hard), and develops non-myopic algorithms for the sequential scheduling problem, illustrating the value of forecasting the residual call-in sequence for a partially constructed schedule.

In contrast to the work discussed above, this research considers sequential scheduling with heterogeneous no-shows in the presence of service constraints. The objective is to maximize revenue from seeing patients subject to constraints on patient waiting time, number of patients in system, and so forth. In our experience, it is much easier for clinic managers to specify bounds such as "the average patient waiting time must be no more than 10 minutes" than it is to specify an actual cost. We also treat overtime as a constraint rather than including it in the objective. Again, in our experience, it is easier for clinic managers to specify a constraint on overtime than to come up with an actual overtime cost. Further, when overtime is included in the objective, the ratio between revenue per patient and overtime cost per patient becomes very important. Indeed, if overtime cost is less than revenue, there is potential for scheduling algorithms to become unstable since the marginal revenue obtained by adding one additional patient is always greater than the marginal overtime cost (see Muthuraman and Lawley, 2008).

We are particularly interested in schedule fairness measures. Cayirli and Veral (2003) define fairness as the uniformity of performance across patients. In their review paper, they suggest that future studies should consider multiple performance measures including fairness to evaluate appointment systems. In the literature, there are few studies that consider fairness while evaluating the quality of the appointment scheduling rules. Bailey (1952) considers the mean waiting time according to the sequence of patients in the schedule. Yang et al. (1998) consider the variance of waiting times, Cox et al. (1985) consider the variance of queue lengths, and Cayirli et al. (2008) consider the standard deviation of waiting times as the fairness measures. These papers use simulation to calculate fairness measures. In this study, we propose new unfairness measures to evaluate schedule uniformity for patients assigned to different slots. We derive analytical expressions for these measures, which are functions of expected waiting time and

the number of patients in the system.

We feel that the main contributions of this paper are as follows.

1. To the best of our knowledge, this is the first study that considers fairness as the primary objective.
2. The first and second moments of waiting time, queue length, and number of patients in the system at each slot are analytically derived for the first time in the literature.
3. The first moments, which are derived based on the assumption of exponential service times, are shown to be upper bounds for general service time distribution.
4. The sequential scheduling problem with service constraints is shown to have different properties than the problem with a weighted linear combination of objectives. With a weighted linear objective, the assignment of a calling patient to the schedule is independent of the patient no-show probability, and thus it is always better for expected profit to schedule low no-show patients. In contrast, when service constraints are in place, adding a patient with low no-show probability may violate constraints, whereas a patient with higher no-show probability might still be added to the schedule.
5. The constraint-based approach can easily handle other constraints such as provider unavailable times, as illustrated in the paper.

The paper is structured as follows. Section 2 defines the sequential scheduling problem with service constraints along with its underlying assumptions. A stochastic programming model is proposed with the objective of minimizing unfairness and maximizing revenue. Service criteria such as the expectation and variance of slot-conditioned waiting time, queue length, and number of patients in the system are bounded in the constraint set. The properties of the service criteria are derived. Section 3 presents a sequential scheduling algorithm to solve the problem. Section 4 presents a variety of computational studies, and Section 5 shows how the proposed approach can handle unavailable provider times. Finally, Section 6 discusses future work and provides some concluding remarks.

2. Problem Definition

In this study, we assume a single server, single stage system. This means that patients are scheduled only for the examination, other stages such as registration and laboratory tests are not considered. The queue for each provider is different, which is a common practice for continuity of care with the primary physician. The service times of patients are assumed to be exponentially distributed. We note that

although this assumption (which is commonly made by scheduling researchers) may not hold in practice, it allows us to develop and test results that give important insights into practical scheduling operations. Also, we show that our first moment derivations provide bounds in general service time settings. In other work, we have developed sequential scheduling results for general service time distributions (Chakraborty et al., 2009a), but we have not yet concisely expressed the performance criteria.

We consider heterogeneous patients, that is, patients with different no-show rates. We assume that patients are punctual, that is, if they arrive for their appointment, they arrive on time. The number of patients that can be scheduled per clinic session is not fixed, because the number may change due to patient mix with different no-show rates. The slot lengths and usage are fixed in advance according to provider preferences. For example, the provider might specify 30 minute slots during some sessions, 20 minute slots for others, or block out certain slots. We note that dividing the service session into slots of fixed length is the prevailing practice in healthcare. In our experience, this is rarely done analytically or optimally, and thus this is an excellent problem for research, which we address in Chakraborty et al. (2009b). It is not, however, the focus of this paper, and thus we take the slot length as given.

Table 1 provides the notation that will be used throughout the paper. In a fixed slot appointment system, the clinic session is divided into I slots of length t_i. There are J types of patients with different no-show rates. Appointment requests are not known in advance, but occur online at any time before the actual appointment session start time. When a patient (say the $(n+1)^{st}$) calls for an appointment in a given session, he/she should be scheduled without changing the appointment times of the previous n patients in session schedule, S^n. A schedule S^n is a matrix of size $I \times J$, where S^n_{ij} represents the number of scheduled patients of type j in slot i. Δ^{ij} is also a matrix of size $I \times J$ where $\delta_{ij} = 1$ and $\delta_{i'j'} = 0$ for any $i' \neq i$ or $j' \neq j$. If patient $n+1$ of type j is assigned to slot i, the new schedule is represented as $S^n + \Delta^{ij}$.

The expected number of arrivals might be less than the number of scheduled patients because of patient no-shows. Let X^n_i be a random variable which denotes the number of arrivals to slot i and Y^n_i be the number of patients waiting for the completion of service at the end of slot i. If L^n_i is a random variable that shows the number of possible service completions assuming infinite demand, then the actual number of service completions is $\min(L^n_i, Y^n_{i-1} + X^n_i)$. The overflow from slot i to slot $i+1$ is calculated as $Y^n_i = Y^n_{i-1} + X^n_i - \min(L^n_i, Y^n_{i-1} + X^n_i) = \max(Y^n_{i-1} + X^n_i - L^n_i, 0)$. Figure 1 shows the relationship among the variables in a fixed slot appointment system.

The arrival (Q) and overflow (R) probability matrices are defined as follows. Let Π be the set of all non-negative, integer, J-vectors $\pi = \langle \pi_1, \pi_2, ..., \pi_J \rangle$ such that $\sum_{j=1}^{J} \pi_j = m$ and $\pi_j \leq S^n_{ij}$ for all j. Q^n_{im} is the probability that m patients arrive in slot i for a given schedule S^n and is calculated by conditioning on the event that π_j patients arrive for their appointments.

I	number of slots
i	slot index, $i \in \{1, \ldots, I\}$
J	number of patient types
j	patient type index, $j \in \{1, \ldots, J\}$
t_i	length of slot i
t^s_{ik}	service time of kth patient overflowing from slot i to $i+1$
t^{rs}_i	remaining service time of the patient in slot i, whose service has started in slot $i-1$
X_i	number of patients arriving for slot i
Y_i	number of overflow patients from slot i to slot $i+1$
L_i	number of possible service completions in slot i
QL_i	Number of patients in the queue at the beginning of slot i, after any arrivals
NS_i	Number of patients in the system at the beginning of slot i, after any arrivals
W_i	Waiting time of patients assigned to slot i
r	revenue per patient
$1-p_j$	no-show probability of patient of type j
S^n	a schedule with n scheduled patients ($\in \mathbf{Z}^{I \times J}$)
S^n_{ij}	number of patients of type j scheduled for slot i in schedule S^n
A_i	set of patients who arrive at slot i
$R(S)$	overflow matrix of schedule S
$Q(S)$	arrival matrix of schedule S
$f(x)$	Probability density function of service time
μ	Expected service time for each patient
b^{eo}	Upper bound on expected overtime
b^{ew}	Upper bound on expected waiting time
b^{eq}	Upper bound on expected queue length
b^{en}	Upper bound on expected number of patients in the system
b^{vw}	Upper bound on variance of waiting time
b^{vn}	Upper bound on variance of number of patients in the system

Table 1: Notation

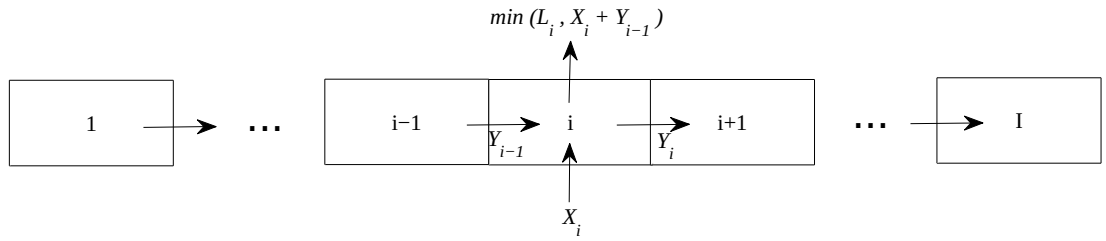


Figure 1: The fixed slot model

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

R^n_{ik} is the probability that k patients overflow from slot i into slot $i+1$ for a given schedule S^n.

$$
\begin{aligned}
R^n_{ik} &= Pr\{Y^n_i = k\} \\
&= Pr\{\max(X^n_i + Y^n_{i-1} - L_i, 0) = k\} \\
&= \begin{cases} Pr\{X^n_i + Y^n_{i-1} - L_i = k\} & k > 0 \\ Pr\{X^n_i + Y^n_{i-1} - L_i \le 0\} & k = 0 \end{cases} \\
&= \begin{cases} \sum_m \sum_{\tilde{k}} Pr\{m + \tilde{k} - k = L_i\} Q^n_{im} R^n_{i-1,\tilde{k}} & k > 0 \\ \sum_m \sum_{\tilde{k}} Pr\{m + \tilde{k} \le L_i\} Q^n_{im} R^n_{i-1,\tilde{k}} & k = 0 \end{cases} \\
&= \begin{cases} \sum_m \sum_{\tilde{k}} f_{L_i}(m + \tilde{k} - k) Q^n_{im} R^n_{i-1,\tilde{k}} & k > 0 \\ \sum_m \sum_{\tilde{k}} (1 - F_{L_i}(m + \tilde{k})) Q^n_{im} R^n_{i-1,\tilde{k}} & k = 0 \end{cases}
\end{aligned}
$$

where $F_{L_i}(l) = Pr\{L_i < l\}$ and $f_{L_i}(l) = Pr\{L_i = l\}$ are calculated using the distribution of service times. For additional detail, see Muthuraman and Lawley (2008).

2.1. Stochastic programming model with service criteria

We now formulate a stochastic programming model that uses revenue and fairness as objectives and service expressions as constraints. We will first define new unfairness measures and then present and discuss the stochastic programming model. (Derivations for the quantities appearing in the constraints are provided in the Appendix A.)

The first proposed measure is the minimization of difference between the number of patients in the system at the beginning of each slot (U_{NS}).

$$U_{NS} = \max_{i=\{1,\ldots,I\}} E[NS_i] - \min_{i'=\{1,\ldots,I\}} E[NS_{i'}] \tag{1}$$

The second unfairness measure is the minimization of difference between the expected waiting time for patients arriving at each slot (U_W).

$$U_W = \max_{i=\{1,\ldots,I\}} E[W_i] - \min_{i'=\{1,\ldots,I\}} E[W_{i'}] \tag{2}$$

Both unfairness measures are non-regular. Thus, they may alternately increase and decrease as more patients are scheduled, and there is no natural criteria to stop scheduling more patients in order to maximize expected profit and minimize unfairness. However, we use other criteria such as overtime ($E[Y_I]\mu$), expected waiting time ($E[W_i]$) and expected number of patients in the queue/system at the beginning of each slot ($E[QL_i]$, $E[NS_i]$) to determine when to stop. We put these criteria into the constraint set and use upper bounds to control each criteria. Variance of waiting times ($Var[W_i]$) and number of patients in the system/queue at the beginning of each slot ($Var[NS_i]$) are the other unfairness criteria that we consider in the constraint set. We proposed the following a stochastic programming model:

$$\begin{aligned}
\min \quad & U_{NS} &&= \max_{i=\{1,\ldots I\}} E[NS_i] - \min_{i'=\{1,\ldots I\}} E[NS_{i'}] \\
\min \quad & U_W &&= \max_{i=\{1,\ldots I\}} E[W_i] - \min_{i'=\{1,\ldots I\}} E[W_{i'}] \\
\max \quad & E[revenue] &&= \textstyle\sum_i rE[X_i] \\
st \quad & E[Y_I]\mu &&= \textstyle\sum_{Ik} kR_{Ik} \leq b^{eo} && (1)\\
& E[W_i] &&= (E[Y_{i-1}](1-Q_{i0}) + \frac{E[X_i]-1+Q_{i0}}{2})\mu \leq b^{ew} && (2)\\
& E[QL_i] &&= E[Y_{i-1}] + E[X_i] - 1 + R_{i-1,0}Q_{i,0} \leq b^{eq} && (3)\\
& E[NS_i] &&= E[Y_{i-1}] + E[X_i] \leq b^{en} && (4)\\
& Var[W_i] &&= \left(E[Y_{i-1}](1-Q_{i0}) + \frac{E[X_i]-1+Q_{i0}}{2}\right)\sigma^2 \\
& &&+ \left(E[Y_{i-1}^2](1-Q_{i0}) + E[Y_{i-1}](E[X_i]-1+Q_{i0}) + \frac{2E[X_i^2]-3E[X_i]+1-Q_{i0}}{6}\right)\mu^2 \\
& &&- \left(E[Y_{i-1}](1-Q_{i0}) + \frac{E[X_i]-1+Q_{i0}}{2}\right)^2 \mu^2 \leq b^{vw} && (5)\\
& Var[NS_i] &&= Var[QL_i] = Var[Y_{i-1}] + Var[X_i] \leq b^{vn} && (6)
\end{aligned}$$

The first constraint guarantees that the expected overtime does not exceed the upper bound b^{eo}. In the second, third, and fourth constraints, the expected waiting time and number of patients in the queue/system are bounded by upper bounds b^{ew}, b^{eq}, and b^{en}, respectively. In the last two constraints, the variance of waiting time and the variance of number of patients in the queue/system are bounded by b^{vw} and b^{vn}, respectively.

Constraints (2), (3) and (4) are functions of $E[X_i]$, $E[Y_{i-1}]$, Q_{i0} and $R_{i-1,0}$. Some of the constraints may become redundant according to the choice of the upper bounds. For example, if $b^{en} = b^{ew}/\mu$, constraint (1) becomes redundant because $E[W_i]/\mu = (E[Y_{i-1}](1 - Q_{i0}) + \frac{E[X_i]-1+Q_{i0}}{2}) < E[Y_{i-1}] + E[X_i] = E[NS_i] \leq b^{en}$. Constraint (3) becomes redundant when $b^{en} = b^{eq}$, since $E[QL_i] = E[Y_{i-1}] + E[X_i] - 1 + R_{i-1,0}Q_{i,0} \leq E[Y_{i-1}] + E[X_i] = E[NS_i] \leq b^{en}$. In that case, using only one constraint is sufficient to achieve the desired effect.

We note that the criteria considered in this study are calculated for each slot, not for the overall schedule. This is not a disadvantage, since the main focus of the paper is providing uniform schedules for patients assigned to different slots. However, to illustrate how composite schedule criteria can be computed from slot-based criteria, we provide a simple example. Table 2 illustrates expected waiting times for a schedule with 4 patients and 3 slots, i.e. $S = (2, 1, 1)$. The no-show probability is 0.2 for all patients, the slot length is one, and the service times are exponentially distributed with a mean one slot length.

(X_1, X_2, X_3)	$Pr\{(X_1, X_2, X_3)\}$	$E[W_1\|(X_1, X_2, X_3)]$	$E[W_2\|(X_1, X_2, X_3)]$	$E[W_3\|(X_1, X_2, X_3)]$	$E[W\|(X_1, X_2, X_3)]$
(2,1,1)	0.4096	0.500	1.104	1.248	0.838
(2,1,0)	0.1024	0.500	1.104	0	0.701
(2,0,1)	0.1024	0.500	0	0.541	0.514
(2,0,0)	0.0256	0.500	0	0	0.500
(1,1,1)	0.2048	0	0.368	0.639	0.336
(1,1,0)	0.0512	0	0.368	0	0.184
(1,0,1)	0.0512	0	0	0.135	0.068
(1,0,0)	0.0128	0	0	0	0
(0,1,1)	0.0256	0	0	0.368	0.184
(0,1,0)	0.0064	0	0	0	0
(0,0,1)	0.0064	0	0	0	0
(0,0,0)	0.0016	0	0	0	0

Table 2: Expected waiting time for all possible realizations (X_1, X_2, X_3) of schedule $S = (2, 1, 1)$

Table 2 shows the expected waiting time for all possible realizations. For a given realization (X_1, X_2, X_3), the expected waiting time of patients arriving to slot i ($E[W_i|(X_1, X_2, X_3)]$) is calculated as $(E[Y_{i-1}] + (X_i - 1)/2)$ (which is equation (2) for $\mu = 1$, $X_i > 0$ and known). The total expected waiting time for the realization (X_1, X_2, X_3) is

$$E[W|(X_1, X_2, X_3)] = \sum_i X_i \cdot E[W_i|(X_1, X_2, X_3)] / \sum_i X_i$$

The expected waiting time at each slot across all realizations is calculated as

$$E[W_i] = \sum_{\forall (X_1, X_2, X_3)} E[W_i|(X_1, X_2, X_3)] \cdot Pr\{(X_1, X_2, X_3)\}$$

where $Pr\{(X_1, X_2, X_3)\}$ is the probability of the realization.

The overall composite measure for expected waiting time for schedule S is calculated as

$$E[W] = \sum_{\forall(X_1,X_2,X_3)} E[W|(X_1, X_2, X_3)] \cdot Pr\{(X_1, X_2, X_3)\}$$

For example, when the realization is (2,0,1), $E[W_1|(2,0,1)] = (0 + (2-1)/2)$, $(E[W_2|(2,0,1)]) = 0$ (since there is no arrival at slot 2), and $(E[W_3|(2,0,1)]) = E[Y_2]$. The overall expected waiting time for realization (2,0,1) is $E[W|(2,0,1)] = (2 \cdot E[W_1|(2,0,1)] + 1 \cdot E[W_3|(2,0,1)])/3 = (2 \cdot (0 + (2-1)/2) + 1 \cdot E[Y_2])/3 = 0.514$. The probability of realizing $(2,0,1)$ is $0.2 \cdot (0.8)^3 = 0.1024$.

Applying these computations to all schedule realizations, we find $E[W_1] = 0.5 \cdot (0.4096) + 0.5 \cdot (0.1024) + 0.5 \cdot (0.1024) + 0.5 \cdot (0.0256) + 0 \cdot (0.0256) + ... + 0 \cdot (0.0016) = 0.320$. Similarly, $E[W_2] = 0.659$ and $E[W_3] = 0.714$. The overall expected waiting time for the schedule (2,1,1) is $E[W] = \sum_{\forall(X_1,X_2,X_3)} E[W|(X_1, X_2, X_3)]Pr\{(X_1, X_2, X_3)\} = 0.567$.

As an aside, note that if we use the slot waiting time formula given by (2), $((E[Y_{i-1}](1 - Q_{i0}) + (E[X_i]-1+Q_{i0})/2)\mu)$, we have to calculate Q_{i0}, $E[Y_{i-1}]$ and $E[X_i]$. According to the arrival and overflow matrices, $Q_{10} = 0.04$, $Q_{20} = 0.2$ and $Q_{30} = 0.2$, $E[Y_1] = 0.824$, $E[Y_2] = 0.892$, $E[X_1] = 1.6$, $E[X_2] = 0.8$ and $E[X_3] = 0.8$. The expected waiting times are calculated as $E[W_1] = (1.6 - 1 + 0.04)/2 = 0.320$, $E[W_2] = 0.824(1-0.2)+(0.8-1+0.2)/2 = 0.659$, and $E[W_3] = 0.892(1-0.2)+(0.8-1+0.2)/2 = 0.714$, which is identical to that given above.

2.2. Criteria properties

We use the memoryless property of exponential distribution to derive the performance criteria incorporating the number of patients overflowing from slot i to slot $i+1$, Y_i. If the service times have a general distribution, the expected overflow calculated based on the exponential service time assumption will be an over-estimation of the overflow value for the general distribution, which is shown by Property 1. Property 1 is an important result which shows that a feasible solution found for the scheduling problem with the upper bounds and exponential service time distribution is also feasible for the problem with general service time distribution.

Property 1. *Given a schedule S^n, the expected overtime ($E[Y_I]\mu$) and the expected service time for patients overflowing from slot $i-1$ to slot i ($E[Y_{i-1}]\mu$), which are calculated for the exponential service time distribution, are upper bounds for other service time distributions assuming identical means.*

The proofs of all properties are provided in Appendix B. Zeng et al. (2010) considered a weighted linear combination of expected revenue, overflow and overtime costs and showed that it was always better to include patients of low no-show probabilities into an existing schedule. This rule is not valid when a constraint-based approach is used, as is shown by Property 2.

Property 2. *Let $S = S^{n-2} + \Delta^{i1} + \Delta^{i4}$ and $S^{'} = S^{n-2} + \Delta^{i2} + \Delta^{i3}$ be two schedules with n patients, where the types of two patients assigned to slot i differ. Schedule S has patients of types 1 and 4 assigned to i and schedule $S^{'}$ has patients of types 2 and 3 assigned to i, where $p_1 + p_4 = p_2 + p_3$ and $p_1 > p_2 \geq p_3 > p_4$.*

a. Both schedules have the same $E[X_i]$.

b. Schedule S has smaller $E[Y_i]$ than schedule $S^{'}$.

c. If $E[Y_{i-1}^{n-2}] < 1/2$, schedule S gives smaller $E[W_i]$ than schedule $S^{'}$.

Property 2 shows that a schedule with a combination of high and low no-show rate patients (schedule S) gives lower $E[Y_i]$ compared to a schedule with medium no-show rates (schedule $S^{'}$). For schedule S', the probability that there will be at least one arrival to slot i is lower than that of schedule S. Therefore, if the expected overflow from previous slot is high ($E[Y_{i-1}^{n-2}] > 1/2$), schedule $S^{'}$ gives a lower expected waiting time than schedule S. The result of Zeng et al. (2010), which shows that it is always better to include patients of low no-show probabilities into an existing schedule, is not valid for the clinical scheduling problem with constraints.

Property 3 shows that the expected overflow ($E[Y_i]$), expected waiting time at each slot ($E[W_i]$), expected queue length ($E[QL_i]$), and the expected number of patients in the system ($E[NS_i]$) at the beginning of each slot are increasing functions of p_j.

Property 3. *For each slot, the expected overflow, expected waiting time, expected queue length and expected number of patients in the system increase as p_j increases.*

As a result of Property 3, if a patient of type j cannot be assigned to slot i due to constraint violation, none of the patients with higher p_j values can be assigned to that slot. Muthuraman and Lawley (2008) proposed a myopic scheduling algorithm with the objective of maximizing the profit, which is a function of expected revenue and expected overflow cost. In their algorithm, the decision to accept (or reject) a patient of type j is independent of p_j (as shown by Zeng et al. (2010)). When a constraint-based approach is used, the stopping condition depends on p_j since all constraints should be satisfied. Property 4 provides the condition to stop scheduling patients to a session.

Property 4. *For a given schedule S, let p_i^c be the critical probability such that no patients with higher p_j values can be assigned to slot i. If a patient of type j has higher p_j than all p_i^c values ($p_j > \max_i p_i^c$), then the patient cannot be assigned to any slot. If $\min_j p_j > \max_i p_i^c$, no patients can be added to the schedule S without violating at least one constraint.*

When more than one slot satisfies all the constraints, the appointment slot should be chosen carefully so that the total number of patients scheduled for a session is maximized. This can be done if the marginal increase in the service criteria is minimized. Property 5 shows that scheduling patients to earlier slots decreases the expected overtime, and hence increases the number of patients that can be scheduled.

Property 5. *The expected increase in overflow at the end of the day* ($E[Y_I]$) *is smaller if a patient is assigned to slot* i_1 *instead of slot* i_2*, where* $i_1 < i_2$*.*

The proposed stochastic programming model should be solved each time a patient calls in for an appointment. However, the schedule for the existing patients cannot be changed. In order to maximize expected revenue, the new patient should be assigned to one of the appointment slots satisfying all constraints. Even though the expected revenue increases, the unfairness measures may increase or decrease. If only the unfairness objective is considered, the proposed model may not schedule the new patient to the current schedule, but may schedule another patient with the same no-show probability at a later time when the schedule changes. In order to avoid this, we proposed a sequential scheduling algorithm, which schedules the patient when a slot satisfying all constraints is found. The algorithm does not accept or reject the patient based on the current decrease or increase in unfairness. However, since the ultimate goal is finding a "fair schedule", the patient is not assigned to the first slot satisfying all constraints. The assignment is made according to the workload at each slot with the objective of minimizing the difference between slots. Properties 3-5 are used to reduce the computation times. In the following section, the proposed sequential scheduling algorithm is explained in detail.

3. Proposed sequential scheduling algorithm

We propose a sequential scheduling algorithm with the objective of minimizing the unfairness measures U_{NS} and U_W. The aim is to find a fair schedule by assigning similar amount of expected workload to each slot. The basic steps of the proposed algorithm are as follows:

Step 1 Initialize $n = 0$, $S^n = \emptyset$, $p_i^c = 1$, and all upper bounds $(b^{eo}, b^{ew}, b^{eq}, b^{en}, b^{vw}, b^{vn})$.

Step 2 Wait for the next patient call before the start time of first slot. Let the patient be of type j. Set i^* to -1.

Step 3 If $p_j > \max_i p_i^c$, patient cannot be assigned, go to Step 5. Otherwise, sort the slots with $p_i^c > p_j$ in non-decreasing order of $\max\{0, E[NS_i^n]\mu - t_i\}$. In case of ties, the slot with a smaller index is chosen first. Let SI be the set of ordered slots.

Step 4 For all $i \in SI$

- *4.1* Temporarily assign the patient to slot i as the $(n+1)^{st}$ patient.
- *4.2* Calculate $E[Y_I^{n+1}]$, $E[W_{i'}^{n+1}]$, $E[QL_{i'}^{n+1}]$, $E[NS_{i'}^{n+1}]$, $Var[W_{i'}^{n+1}]$ and $Var[QL_{i'}^{n+1}] = Var[NS_{i'}^{n+1}]$ for all i'.
- *4.3* If all constraints with an upper bound are satisfied, then $i^* = i$. Go to Step 5. Else if at least one of the constraints among (1)-(4) are not satisfied, update $p_i^c = p_j - \epsilon$ (due to Property 3). Go to Step 4.4.
- *4.4* If all slots in SI are evaluated, go to Step 5. Otherwise, go to Step 4.1 for the next slot in the sorted list SI.

Step 5 If $i^* = -1$, patient cannot be added to schedule S^n, go to Step 6. Otherwise, assign the patient to slot i^*, increase n by 1 and update the schedule as $S^n = S^{n-1} + \Delta^{i^*j}$

Step 6 If $\min_j p_j \leq \max_i p_i^c$, go to Step 2 to wait for the next patient call. Otherwise, terminate since no more patients can be added to schedule S^n (due to Property 4).

The scheduling algorithm starts with an empty schedule of a given session. Since any patient can be added to an empty schedule, the critical probability p_i^c is set to one in Step 1. In Step 2, the scheduler waits until a patient calls for an appointment. When the patient calls, the no-show probability is estimated and the patient is categorized in one of the predetermined patient groups. Let the patient be of type j with no-show probability of $1 - p_j$. In Step 3, p_j is compared with the critical probabilities at each slot (p_i^c). If p_j is greater than all p_i^c values, then the patient cannot be assigned to any of the slots (Property 4). Otherwise, the slots are sorted according to the expected overtime required to serve all the patients in the system at the beginning of the slot ($\max\{0, E[NS_i^n]\mu - t_i\}$). If the expected time to serve all patients at the beginning of the slot is less than the slot length, the slot will have a higher priority while the patient is being scheduled. When there are multiple slots with the same priority, the earlier slot is chosen since it gives smaller increase in expected overtime (Property 5). The main reason for sorting the slots according to the expected workload is finding a "fair" schedule with similar workload at each slot. Once the slots are sorted, the patient is assigned to the first slot in the sorted listed temporarily (Step 4.1). In Step 4.2, the expectation and variance of waiting time, queue length and number of patients in the system are calculated. In Step 4.3, constraint feasibility is checked for all slots. If all constraints are satisfied, the patient is assigned to that slot i, i.e. $i^* = i$. If at least one of

the constraints among (1)-(4) are not satisfied, which means the patient of type j cannot be assigned to slot i, no patients with higher $p_{j'}$ values can be assigned to slot i (Property 3). In order to ensure that, the critical probability of slot i, p_i^c, is updated as $p_j - \epsilon$, where ϵ is a very small number such that $p_j - \epsilon > p_k, \forall p_k < p_j$. Please note that the critical probability is updated only when constraints (1)-(4) are violated. Property 3 cannot be generalized for constraints (5)-(6), because the variance is not an increasing function of p_j. For example, the variance of arrivals to a slot ($Var[X_i]$) is minimum when $p_j = 0$. As p_j increases, $Var[X_i]$ increases until p_j becomes 0.5. After that point, $Var[X_i]$ decreases as p_j gets closer to 1. If the patient cannot be assigned to slot i due to constraint violation, Steps 4.1–4.5 are repeated for the next slot in the sorted list SI. In Step 5, if a slot satisfying all the constraints is found, the patient is assigned to that slot. Otherwise, the patient is not scheduled for that day. Another day is searched for an available slot using the same algorithm. In Step 6, the stopping condition is checked. If $\min_j p_j > \max_i p_i^c$, no more patients can be added to the schedule and the algorithm terminates for the given session. Otherwise, the scheduler waits for next patient call.

4. Computational Study

The computational study is in two parts. In the first, our aim is to show trade-off among multiple criteria. The proposed model is solved with different upper bound values to find the trade-off between expected revenue and unfairness and to show the effect of different bounds on the service criteria, which are in the constraint set. In the second part of the study, we show the effect of different appointment slot lengths on expected revenue and unfairness.

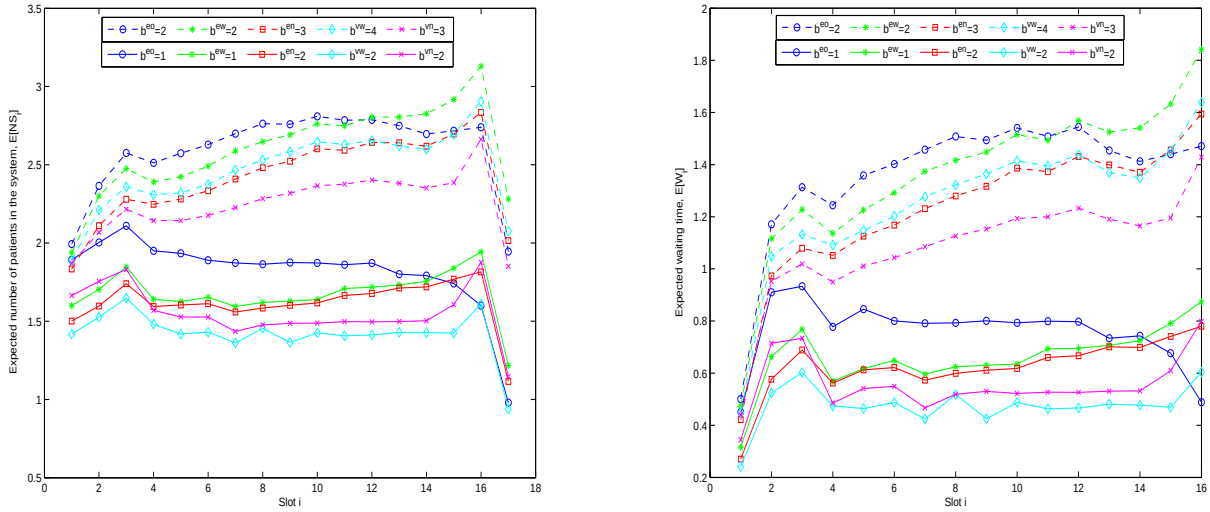
4.1. Trade-off among multiple criteria

We consider a 4-hour session with 16 slots (each slot is 15 minutes). The service times are assumed to be exponentially distributed with mean $\mu = 1$, which corresponds to one slot (expected service time is 15 minutes). Three types of patients with no-show probabilities of 0.1, 0.3, and 0.5 are considered. When the patient calls in, the no-show probability of the patient is estimated and the patient is categorized into one of the three patient groups where $p_1 = 0.9$, $p_2 = 0.7$ and $p_3 = 0.5$. We consider the objective of minimizing U_{NS} and U_W, and maximizing expected revenue. The upper bounds considered in the first part of the computational study are 1 and 2 slots for expected overtime and expected waiting time, 2 and 3 patients for expected number of patients in the system, 2 and 4 for the variance of waiting time and variance of number of patients in the system (b^{eo}= 1 or 2, b^{ew}= 1 or 2, b^{en}= 2 or 3, b^{vw}= 2 or 4, and b^{vn}= 2 or 4). Since all the criteria in the constraint set are functions of R_{ik} and Q_{im}, only one upper bound is fixed for a given experiment and the other bounds are set to infinity, resulting in

8 different problem settings. One hundred different call-in sequences are randomly generated for each problem setting. Figures 2– 5 plot the average of the 100 replications.

Figure 2.(a) shows the expected number of patients in the system at the beginning of each slot ($E[NS_i]$). In the first 14 slots, the problems with an upper bound on expected overtime give higher expected number of patients in the system ($E[NS_i]$). The expected number of patients in the last slot (slot 16) is the minimum. (Please note that there are 16 slots to which a patient can be scheduled.) The expected number of patients in the system at the beginning of slot 17 shows the expected number of patients who are served during overtime hours. The proposed algorithm reduces the number of patients in later slots in order to satisfy the overtime constraint. When there are upper bounds only on the expected waiting time or expected number of patients in the system, the number of patients scheduled at later slots increase, which results in higher overtime. When the upper bounds increase, expected number of patients in the system increases.

Figure 2.(b) shows the expected waiting time at each slot ($E[W_i]$). The expected waiting time at each slot shows the same pattern as $E[NS_i]$. Figure 3 shows the expected number of arrivals at each slot. The expected number of arrivals in the first three slots is higher than the ones in later slots. When the upper bounds increase, the expected number of arrivals at each slot increases, resulting in improved expected profit.



(a) (b)

Figure 2: (a) Expected number of patients in the system at the beginning of each slot ($E[NS_i]$), (b) Expected waiting time at each slot ($E[W_i]$)

Figure 4.(a) shows the trade-off between the expected revenue and unfairness measure U_{NS}. The

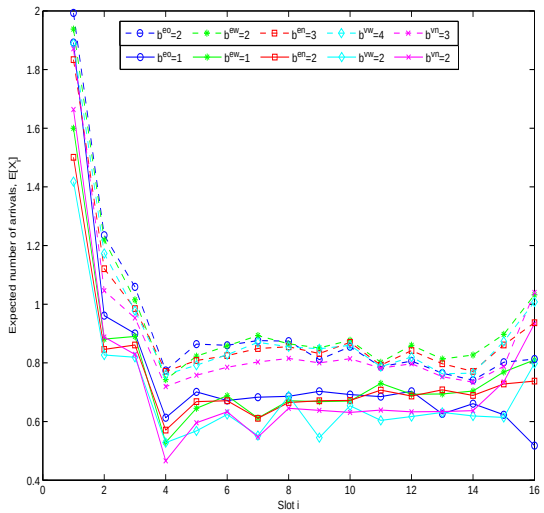
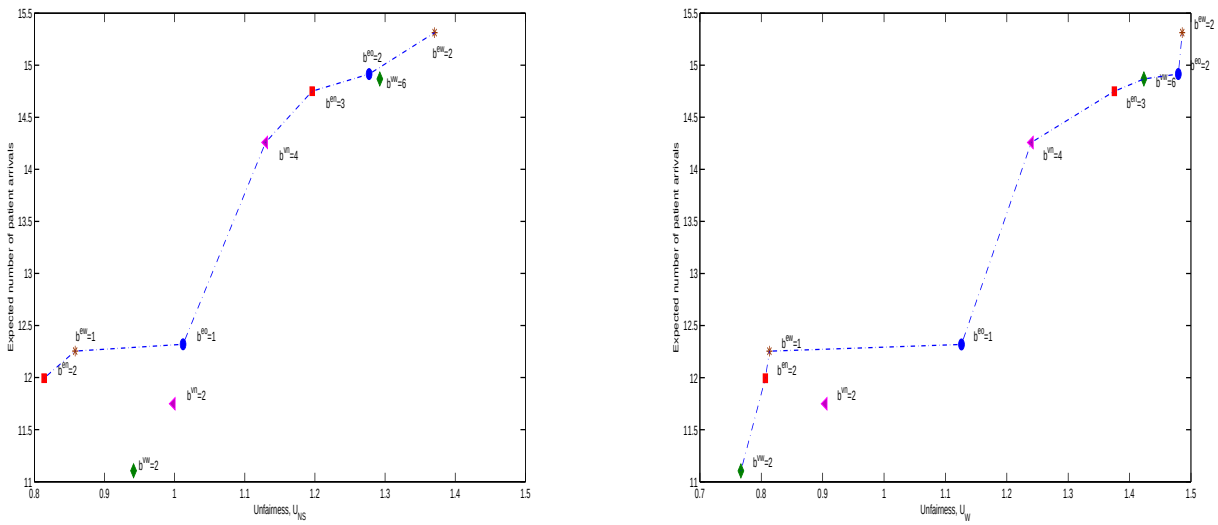


Figure 3: Expected number of arrivals at each slot ($E[X_i]$)

expected revenue increases as the upper bounds increase. However, the unfairness also increases. The connected points are non-dominated solutions. When upper bounds are tight, the solutions found for the problems with the expected overtime, waiting time, and number of patients in the system form the non-dominated solution set. When the bounds increase, all schedules, except the schedule generated for the problem with upper bounds on variance of waiting time, form the non-dominated solution set. Figure 4.(b) shows the trade-off between expected revenue and U_W. We can see that putting a tight upper bound on variance of number of patients in the system does not generate a non-dominated solution.

Figure 5 shows the variance of the number of patients in the system and waiting time. When bounds are tight, the variance in the first three slots increases rapidly and then stabilizes for the remaining slots. As the upper bounds increase, the variance increases as the slot number increases. Thus, patients scheduled to later slots experience more variable waiting times.

When we analyze individual schedules, we can see that, most of the time, only one patient is scheduled to each slot when the bounds are tight. As the bounds increase, the slots are overbooked. The following example shows how the schedules change for different upper bounds. The call-in sequence for the example is (3, 3, 1, 1, 1, 2, 2, 1, 3, 3, 2, 1, 3, 2, 3, 2, 1, 2, 1, 1, 2, 3, 3, 2, 2, 1, 1, 1, 3, 3, 1), where the first call-in patient is of type 3, the second call-in patient is type 3, the third patient is of type 1, and so on. Figure 6 shows a Gantt chart of the schedules for different upper bounds on expected overtime. When the upper bound on expected overtime is 1 (corresponding to 15 minutes), the first 17 patients are scheduled. The schedule is S=[2,0,0; 1,0,0; 0,0,2; 0,0,0; 0,1,0; 0,1,0; 1,0,0; 1,0,0; 0,0,1; 0,0,1; 0,1,0; 1,0,0; 0,0,1; 0,1,0; 0,0,1; 0,1,0], where two patients of type 1 ($p_1 = 0.9$) are scheduled to arrive in the first slot, one patient



(a) (b)

Figure 4: (a) Trade-off between expected revenue and unfairness measure U_{NS}, (b) Trade-off between expected revenue and unfairness measure U_W

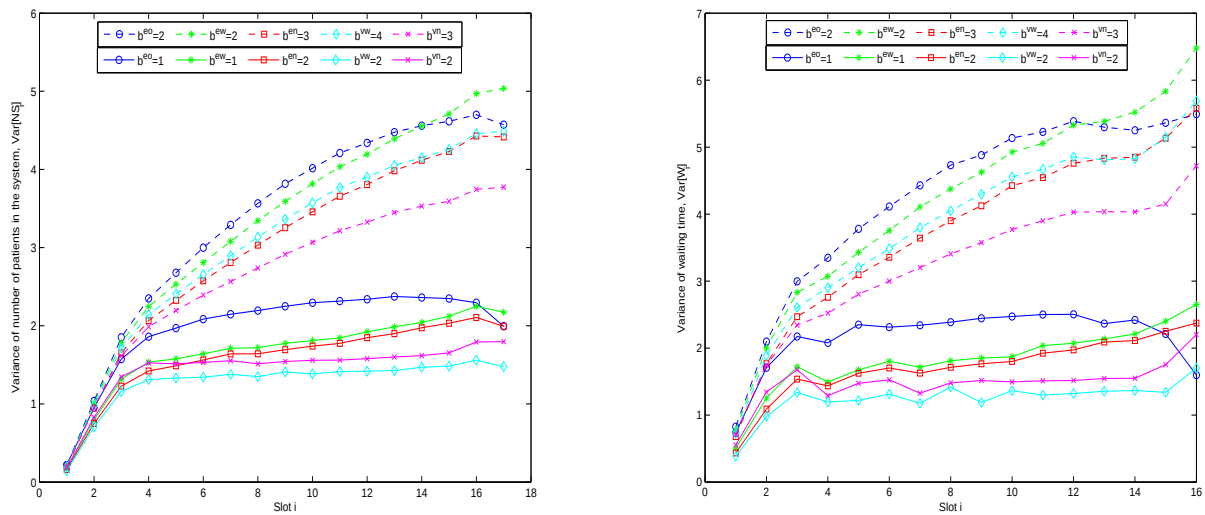
of type 1 is scheduled to arrive in the second slot, two patients of type 3 ($p_3 = 0.5$) are scheduled to arrive at the third slots, and so on. As the upper bound on expected overtime increases, the scheduled number of patients also increases. When the upper bound is 2, the first 20 patients are scheduled. When upper bound is 3, the first 24 patients are scheduled. The expected number of arrivals are 11.9, 14.4 and 17.0, and unfairness measures (U_W) are 0.9267, 1.2543 and 2.4914 for the upper bounds 1, 2 and 3, respectively. When the upper bound is reduced from 3 to 1, the unfairness improves by 1.5647 (2.4914 - 0.9267), which corresponds to 23.5 minutes decrease in the difference between maximum and minimum expected waiting times. However, the expected number of arrivals decreases by 5.1 patients. If the expected revenue per patient is \$150, the expected revenue loss is \$765. The revenue loss per one minute of improvement in unfairness is \$32.6 (\$765/23.5 minutes).

To summarize, we make the following observations:

- When upper bounds are tight, the proposed algorithm schedules less patients leading to less expected profit. As the upper bounds increase, more patients are scheduled. Overbooking increases the expected profit.

- When the bounds are tight, the proposed algorithm gives lower unfairness values (U_{NS}, U_W). Higher upper bounds increase the unfairness between slots.

- When the bounds are tight, the expectation and variance of waiting time and number of patients



(a) (b)

Figure 5: (a) Variance of number of patients in the system at the beginning of each slot ($Var[NS_i]$), (b) Variance of waiting time at each slot ($Var[W_i]$)

in the system become stable after the third slot. When the upper bounds are high, the variance of waiting time and number of patients in the system increases as the slot index increases. This leads to more extreme patient waiting times for the patients scheduled to later slots.

- The service criteria at each slot show similar behavior when the expectation and variance of waiting time and number of patients in the system at each slot are bounded. When expected overtime is bounded, the the proposed algorithm schedules more patients to earlier slots and less patients to later slots. While the expected revenue improves, the unfairness degrades.

- Non-dominated solutions that maximize expected profit and minimize unfairness can be generated by changing the upper bounds on constraints. As the upper bounds increase, more patients are scheduled and higher expected profit is achieved. However, the unfairness also increases. The non-dominated solutions provide alternatives to the decision maker.

4.2. Effect of slot length

In the first part of our study, we assume that the slot lengths are equal to the expected service time. However, the slot length and the number of slots in each session would affect the scheduling decisions. In the second part of the computational study, our aim is to show the effect of slot length on multiple moment-based criteria. We consider 30, 15, and 5-minute slots in a 4-hour session. We assume that the

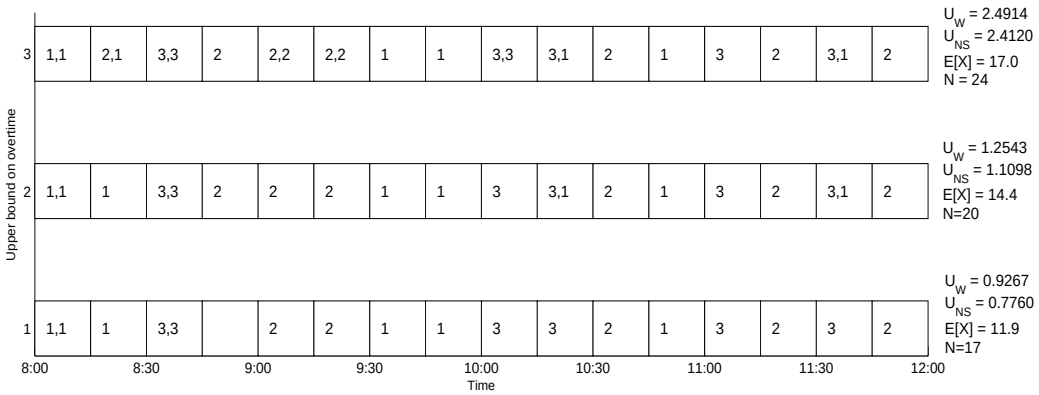
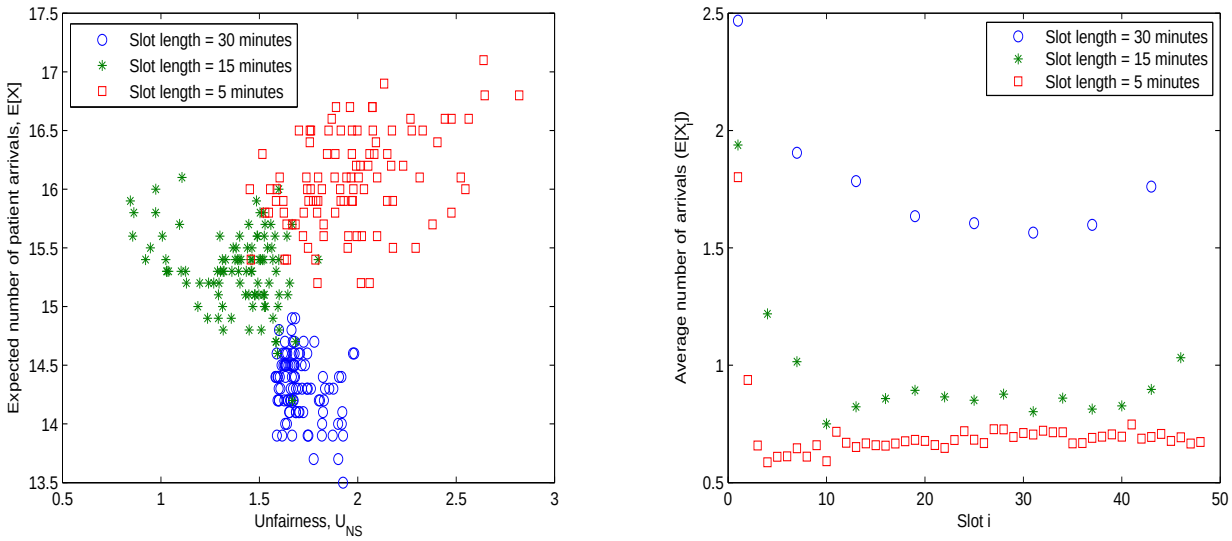


Figure 6: Gantt chart of the schedules for a call-in sequence of (3,3,1,1,1,2,2,1,3,3,2,1,3,2,3,2,1, 2,1,1,2,3,3,2,2,1,1,1,3,3,1), where $p_1 = 0.9$, $p_2 = 0.7$ and $p_3 = 0.5$

service times are exponentially distributed with mean 15 minutes. Three types of patients with no-show probabilities of 0.1, 0.3, and 0.5 are considered. The upper bound for expected waiting time at each slot is set as $ub^{ew} = 30$ minutes. All other upper bounds are set to infinity. The primary objective is minimization of the unfairness measure U_{NS}, that considers the expected number of patients in the system at the beginning of each slot. One hundred different call-in sequences are generated for all slot length alternatives.

In Figure 7 (a), the trade-off between the expected number of arrivals and unfairness (U_{NS}) can be seen. When slot length is 5 minutes (which is shorter than the average service time), the number of patients that can be scheduled increases. However, unfairness also increases due to the slots with no patients. Unfairness is minimized when 15-minute appointment slot lengths are used. The unfairness shows high variability for both 5-minute and 15-minutes slots, which is due to the randomness of call-in sequence. Figure 7 (b) shows the expected number of arrivals ($E[X_i]$) at each slot. The average number of arrivals at each slot gets smaller as the slot length decreases.

In Figure 8 (a), the trade-off between the expected number of arrivals and the second unfairness measure (U_W) can be seen. The unfairness increases as the slot length decreases. However, the proposed unfairness measure (U_W), which is based on the expected waiting times of the patients assigned to different slots, includes the empty slots with zero expected waiting times. That means, when there are empty slots, the unfairness measure shows the maximum expected waiting time rather than the difference of expected waiting times for the patients assigned to different slots. Since the number of empty slots will be higher when slot lengths smaller than service times are used, the unfairness measure will be more than the actual difference between expected waiting times. In Figure 8 (b), the unfairness measure is



(a) (b)

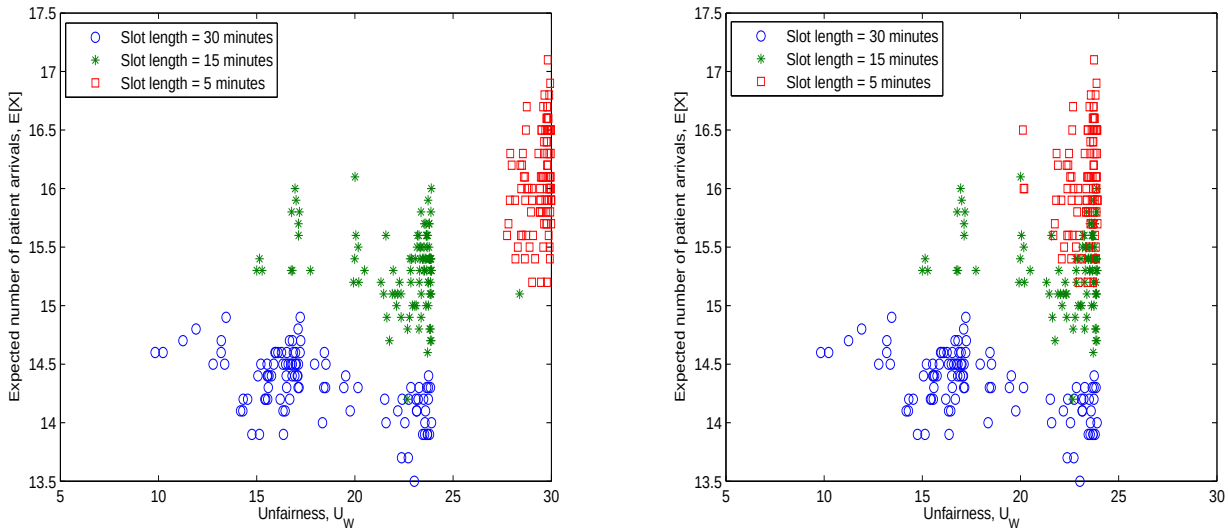
Figure 7: (a) Trade-off curve between unfairness (U_{NS}) and expected profit for slot lengths 30, 15 and 5 minutes, (b) Average number of arrivals at each slot

calculated for the slots with at least one scheduled patient. The actual difference between the expected waiting times of the patients scheduled to different slots has decreased for slot length of five minutes. The maximum unfairness for all slot lengths is 24 minutes. The unfairness has less variability for smaller slot lengths.

Figure 9 shows the average waiting times at each slot for the slots with at least one scheduled patient and the average number of patients in the system at the beginning of each slot. The average waiting time and the number of patients in the system at each slot is lower for 5-minute appointment slots.

Based on our computational results, we make the following observations:

- The expected waiting times at each slot and number of patients in the system at the beginning of each slot are minimized and the profit is maximized when smaller slot sizes are used.
- The average U_{NS} and variability in U_{NS} increase as the slot length decreases. This is due to empty slots and higher number of scheduled patients.
- The average U_W increases as the slot length decreases. However, the variability in U_W decreases.
- When empty slots are excluded from the calculation of U_W, then the maximum unfairness measures for all slots lengths are close to each other.



(a) (b)

Figure 8: Trade-off curve between unfairness (U_W) and expected profit for slot lengths 30, 15 and 5 minutes, considering (a) all slots, and (b) slots with at least one scheduled patient

5. Unavailable time periods

The constraint-based approach can be used to handle several other constraints such as unavailable time periods. Unavailable time periods occur due to providers' other clinical and administrative duties such as making rounds in the hospital, performing surgeries, having clinic sessions in different clinics, being on-call, attending meetings, teaching, etc. In this section, we consider unavailable time periods to show how the proposed approach can handle these other constraints.

We consider two cases of provider unavailability. In the first case, when the provider becomes unavailable, patients in the system have to wait until the provider becomes available. This case might occur when the physician has to attend another scheduled event such as meetings, classes, surgery, etc. It might also occur when the physician is on-call and has to leave the clinic for emergency patients. In the second case, the provider serves all patients who are waiting at the beginning of unavailable period and then becomes unavailable for a fixed period of time. Patients who are scheduled to arrive during the unavailable period will have to wait until the provider becomes available. This case occurs when the provider has to be in different places before and after the unavailable time period. For example, if the provider arrives to the clinic after making rounds in the hospital or performing surgeries, the patients have to wait until the provider comes to the clinic. If the provider has his/her morning and afternoon sessions in different clinics, the provider sees all the patients in one clinic and then goes to the

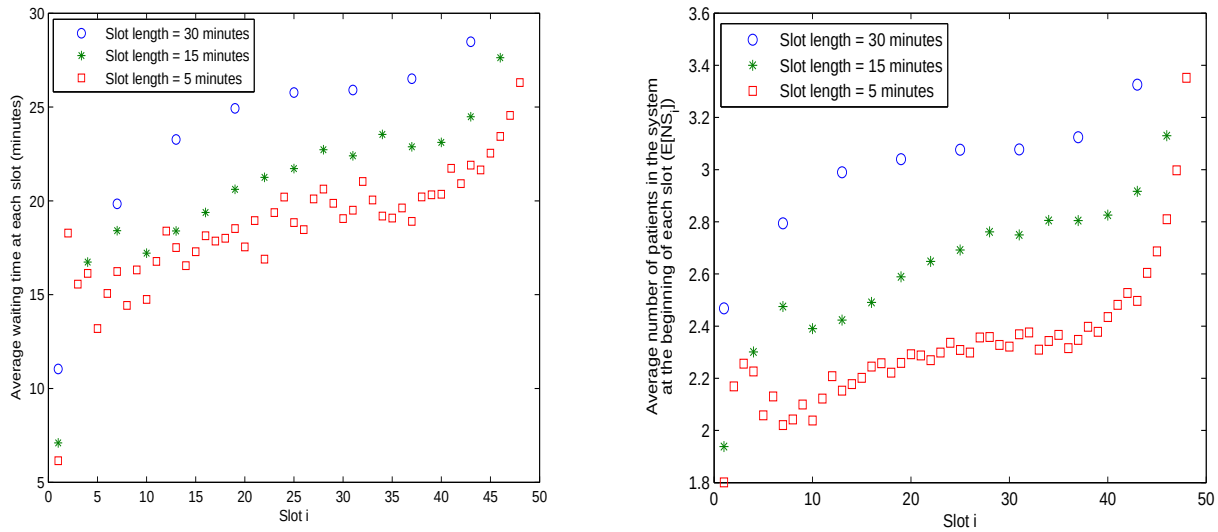


Figure 9: Expected waiting time at each slot and expected number of patients in the system at the beginning of each slot

other clinic. The patients in the second clinic have to wait until the provider arrives. In the following subsections, we will show how the expectation of waiting time, queue length and number of patients in the system at each slot is calculated for the patients affected by the unavailable time periods.

Case 1: The patients in the system at the beginning of unavailable period have to wait until the physician becomes available

Figure 10 is a pictorial representation of the first case. In this study, the moment-based criteria are calculated at each slot for the patients assigned to that slot. If there is an unavailable period during the day (i^u), then the criteria at each slot $i < i^u$ should be re-calculated.

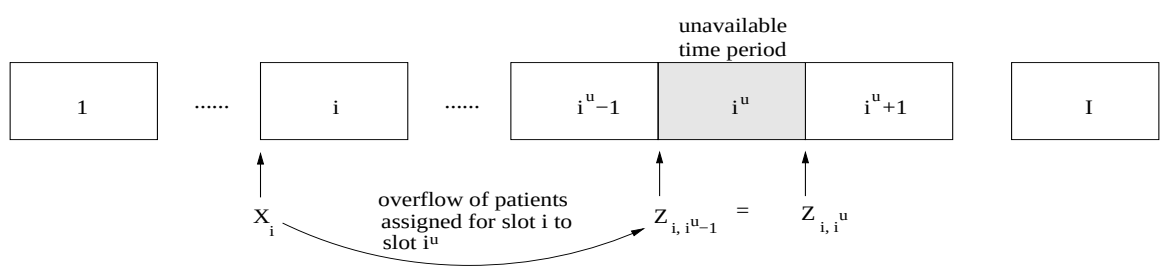


Figure 10: The unavailable period where the patients in the system at the beginning of unavailable period have to wait until the physician becomes available

The arrival matrix for the unavailable period should be $Q_{i^u,0} = 1$ and $Q_{i^u,m} = 0$ for $m > 0$. This will

be guaranteed by putting an additional constraint to the model, i.e. $E[X_i] = 0$. The overflow matrix for the unavailable period should be $R_{i^u,k} = R_{i^u-1,k}$ for all k. This can be guaranteed by putting an additional constraint on service rate in the unavailable slot, i.e., $Pr\{L_{i^u} = 0\} = 1$. The expectation and variance of number of patients in the queue/system are functions of $E[X_i]$, $E[Y_{i-1}]$, $R_{i-1,0}$, Q_{i0}, $Var[X_i]$ and $Var[Y_{i-1}]$ and can be calculated based on the arrival and overflow matrices, which are updated after scheduling each patient.

In order to calculate the expected waiting time at each slot, we need to define a new variable (Z_{i,i^u-1}) to denote the number of patients who are scheduled to arrive at slot i and are still in the system at the end of slot $i^u - 1$. These patients have to wait until the provider becomes available at the beginning of slot $i^u + 1$. The probability that there are z patients, who are scheduled to arrive at slot i and are still in the system at the end of slot $i^u - 1$, is calculated as follows:

$$Pr\{Z_{i,i^u-1} = z\} \quad = \quad \begin{cases} \sum_k \sum_{m \geq z} R_{i-1,k} Q_{im} Pr\{L_{i,i^u-1} = m + k - z\} & z > 0 \\ \sum_k \sum_{m \geq z} R_{i-1,k} Q_{im} Pr\{L_{i,i^u-1} \geq m + k\} & z = 0 \end{cases}$$

The calculation of $Pr\{Z_{i,i^u-1} = z\}$ is very similar to the calculation of $R_{ik} = Pr\{Y_i = k\}$. When R_{ik} is calculated, overflow from slot i to slot $i + 1$ is considered for all patients in the system at the beginning of slot i. When $Pr\{Z_{i,i^u-1} = z\}$ is calculated, overflow from slot i to slot $i^u - 1$ is considered only for the patients scheduled to arrive at slot i. The expected waiting time at each slot i ($i \leq i^u - 1$) is calculated as follows:

$$E[W_i] = (E[Y_{i-1}](1 - Q_{i0}) + \frac{E[X_i] - 1 + Q_{i0}}{2} + E[Z_i, i^u - 1]t_{i^u})\mu$$

where t_{i^u} is the length of the unavailable time period.

There will not be any change in the basic steps of the proposed algorithm, because the arrival and overflow matrices are already calculated at each step, where the patient is temporarily assigned to a slot. However, there will be additional calculations to find $Pr\{Z_{i,i^u-1} = z\}$ and $E[Z_{i,i^u-1}]$ for all slots $i < i^u$.

Case 2: The patients in the system at the beginning of unavailable period are served, but start time for the patients scheduled after the unavailable period might be late

A pictorial representation of the second case can be seen in Figure 11. In this case, the moment-based criteria for the slots before the unavailable time period are not affected. However, there may be additional patient waiting due to provider unavailability at each slot after the unavailable time period ($i > i^u$).

The arrival time of the provider depends on the total time required to see the patients overflowing

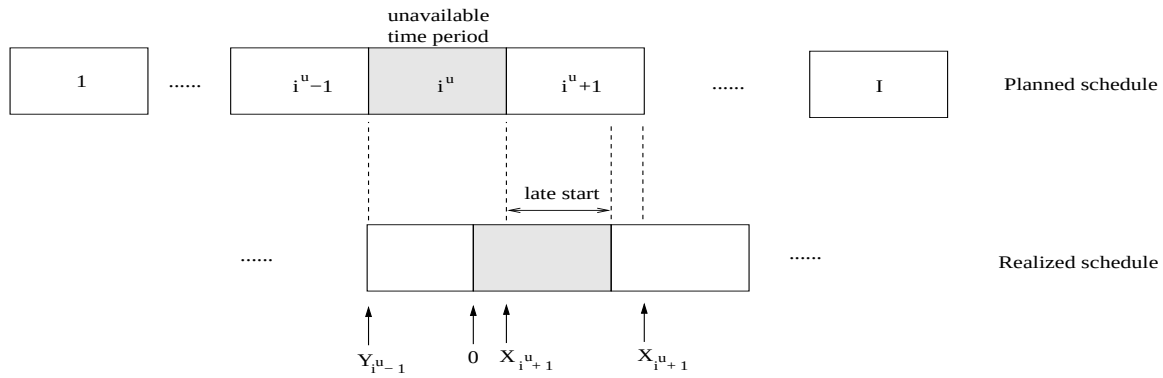


Figure 11: The unavailable period where the patients in the system at the beginning of unavailable period are served, but the patients scheduled after the unavailable period start to be served late

from slot $i^u - 1$ to slot i^u and the duration of unavailable time period. The duration of provider lateness is equal to $\mu \cdot Y_{i^u-1}$. If we just ignore the unavailable period, and calculate the arrival and overflow matrices for the available periods assuming that the unavailable period length is zero, then we can calculate expectation and variance of waiting time and number of patients in the queue/system at each slot without making any changes in our calculations. The only performance criterion that will change is the expected overtime, which will be calculated as $E[Y_I]\mu + t_{i^u}$, where t_{i^u} is the length of unavailable time period. The second case do not add any additional complexity to the execution of the proposed algorithm.

In both cases, there is no change in the basic steps of the proposed constraint-based approach. However, new constraints might be added to the model and additional calculations might be necessary to calculate the moment-based criteria.

6. Conclusion

In this study, we considered the clinical scheduling problem with stochastic service times and patient no-shows. The main contribution of the study is using unfairness and congestion measures as primary objectives. We proposed new unfairness measures in order to find uniform schedules for the patients assigned to different slots. The proposed measures are the minimization of the difference between maximum and minimum expected waiting times at each slot and the number of patients in the system at the beginning of each slot. In order to calculate these measures, we derive analytical expressions of the criteria for each slot. The derived expressions are exact expressions for the exponential service time distribution and approximations for the general distribution.

We also considered several other moment-based criteria such as expectation and variance of waiting

times and number of patients in the system/queue. The variance of waiting time and number of patients in the system/queue can also be considered as fairness measures. Instead of minimizing a weighted linear combination of all the criteria, we use a constraint-based approach. We believe that it is easier for the decision maker to give an upper bound on these criteria rather than assigning a cost value. We derived the properties of the criteria and showed that the no-show rate of the patients is actually important in accepting or rejecting the patient for a given session when a constraint-based approach is used. We also showed the flexibility of constraint-based approach in handling unavailable time periods. The basic steps of the proposed algorithm do not change. There would be additional constraints and calculations to calculate the moment-based criteria and find a slot satisfying all constraints.

We proposed a sequential scheduling algorithm in order to minimize the unfairness and maximize the revenue. We performed a computational study in order to show the trade-off between multiple criteria. According to the results, unfairness can be minimized when the upper bounds are tight. However, this reduces the number of patients scheduled, which might have adverse effects on unscheduled patients' health due to reduced access to care. As the upper bounds increase, the expected revenue increases because of overbooking. The unfairness measures, especially the variance of waiting times and expected number of patients in the system, become worse for the patients assigned to later slots. In our computational study, we also showed the effect of slot lengths on the performance measures. As the slot length decreases, the expected revenue increases, and the expected waiting time and number of patients in the system decreases. The unfairness increases due to empty slots, where no patient can be scheduled.

Future research directions include considering other constraints such as service level constraints, unexpected provider unavailability (due to emergency) and restrictions on the assignment of patient types (such as new and return patients) with different service time requirements to slots. The multiple patient types can have different revenue values and constraints in the model. Another challenging future research direction might be finding the worst case performance bounds for the objectives based on bounds on the constraints.

References

[1] AHRQ. National Quality Health Report, 2007. http://www.ahrq.gov/qual/nhqr07/nhqr07.pdf

[2] U.S. Census Bureau. Ambulatory health care services: 2002.

[3] N. Bailey. A study of queues and appointment systems in hospital outpatient departments with special reference to waiting times. *Journal of the Royal Statistical Society*, 14, 185–199, 1952.

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

[5] T. Cayirli, E. Veral and H. Rosen Assessment of patient classification in appointment system design. *Production and Operations Management*, 17 (3), 338–353, 2008.

[6] S. Chakraborty, M.A. Lawley, K. Muthuraman. Sequential clinical scheduling with patient no-shows and general service time distributions. To appear in *IIE Transactions*, 2009a.

[7] S. Chakraborty, K., Muthuraman, M.A. Lawley. Sequential clinical scheduling under patient no-show in absence of fixed appointment slots. *Submitted*, 2009b.

[8] T.F. Cox, J.F. Birchall and H. Wong. Optimizing the queuing system for an ear, nose and throat outpatient clinic. *Journal of Applied Statistics*, 12, 113–126, 1985.

[9] J. Daggy, M.A. Lawley, D., Willis, D. Thayer, C. Suelzer, L.P. Sands, P.C. DeLaurentis, S. Chakraborty. The effect of accurate no-show modeling on clinical scheduling outcomes. *Submitted*, 2009.

[10] L.V. Green and S. Savin. Reducing delays for medical appointments: A queueing approach. *Operations Research*, 56 (6), 1526–1538, 2008.

[11] D. Gupta and B. Denton. Appointment scheduling in healthcare: Challenges and opportunities. *IIE Transactions*, 40, 800–819, 2008.

[12] R. Hassin and S. Mendel. Scheduling Arrivals to Queues: A Single-Server Model with No-Shows. *Management Science*, 54 (3), 565–572, 2008.

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

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

[15] 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.

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

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

[18] K. Muthuraman and M.A. Lawley. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Transactions*, 40 (9), 820–837, 2008.

[19] S. Savin. Managing patient appointments in primary care. *Patient Flow: Reducing Delay in Healthcare Delivery*, ed. Randolp W. Hall, Springer, New York, 2006.

[20] K.K. Yang, M.L. Lu and S.A. Quek. A new appointment rule for a single-server, multiple customer service system. *Naval Research Logistics*, 45, 313–326, 1998.

[21] B. Zeng, A. Turkcan, J. Lin and M.A. Lawley. Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. To appear in *Annals of Operations Research*, 2010.

Appendix A. Derivations of service criteria

Expected waiting time of the patients assigned to slot i

Expected waiting time of a patient assigned to slot i ($E[W_i]$) is calculated by conditioning on the number of patients in the system who will be served before that patient, δ_i.

$$E[W_i] = E[E[W_i|\delta_i]] = E[\delta_i]\mu \text{ (service times are i.i.d. with mean } \mu)$$

$E[\delta_i]$ is calculated by conditioning on the number of patients in the system at the beginning of slot i.

$$\begin{aligned} E[\delta_i] &= E[E[\delta_i|Y_{i-1}, X_i]] = \sum_k \sum_m R_{i-1,k} Q_{im} \frac{1}{m} \sum_{s=0}^{m-1} (k+s) \\ &= \sum_k \sum_m R_{i-1,k} Q_{im} (k + (m-1)/2) \\ &= E[Y_{i-1}](1 - Q_{i0}) + \frac{E[X_i] - 1 + Q_{i0}}{2} \end{aligned}$$

Then the expected waiting time becomes

$$E[W_i] = (E[Y_{i-1}](1 - Q_{i0}) + \frac{E[X_i] - 1 + Q_{i0}}{2})\mu$$

Expected queue length at the beginning of slot i

The queue length at the beginning of a slot (QL_i) is the sum of overflow from previous slot (Y_{i-1}) and the number of arrivals for the current slot (X_i) minus 1. The expected queue length at the beginning of slot i is calculated as:

$$\begin{aligned} E[QL_i] &= E[E[QL_i|Y_{i-1}, X_i]] \\ &= \sum_{k \geq 1} \sum_{m \geq 0} R_{i-1,k} Q_{im} (k + m - 1) + \sum_{m \geq 1} R_{i-1,0} Q_{im} (m - 1) \\ &= E[Y_{i-1}] + (E[X_i] - 1)(1 - R_{i-1,0}) + R_{i-1,0}(E[X_i] - 1 + Q_{i0}) \\ &= E[Y_{i-1}] + E[X_i] - 1 + R_{i-1,0} Q_{i0} \end{aligned}$$

Expected number of patients in the system at the beginning of slot i

The number of patients in the system at the beginning of a slot (NS_i) is the sum of overflow from previous slot (Y_{i-1}) and the number of arrivals for the current slot (X_i). The expected number of patients in the system at the beginning of slot i is:

$$E[NS_i] = E[Y_{i-1}] + E[X_i] \text{ (since } Y_{i-1} \text{ and } X_i \text{ are independent)}$$

Variance of waiting time for the patients assigned to slot i

The variance of waiting time of the patients assigned to slot i ($Var[W_i]$) is calculated as

$$Var[W_i] = E[Var[W_i|\delta_i]] + Var[E[W_i|\delta_i]] \text{ (conditional variance)}$$

where δ_i is the number of patients in the system who will be served before a patient.

$$\begin{aligned} Var[W_i|\delta_i] &= \delta_i\sigma^2(\text{service times are i.i.d. with variance } \sigma^2) \\ E[W_i|\delta_i] &= \delta_i\mu(\text{service times are i.i.d. with mean } \mu) \end{aligned}$$

The variance of waiting time ($Var[W_i]$) becomes

$$Var[W_i] = E[Var[W_i|\delta_i]] + Var[E[W_i|\delta_i]] = E[\delta_i]\sigma^2 + Var[\delta_i]\mu^2$$

$Var[\delta_i] = E[\delta_i^2] - (E[\delta_i])^2$ and $E[\delta_i^2]$ is calculated as:

$$\begin{aligned} E[\delta_i^2] &= E[E[\delta_i^2|Y_{i-1}, X_i]] \\ &= \sum_k \sum_m R_{i-1,k} Q_{im} \frac{1}{m} \sum_{s=0}^{m-1} (k+s)^2 \\ &= \sum_k \sum_m R_{i-1,k} Q_{im} (k^2 + k(m-1) + \frac{(m-1)(2m-1)}{6}) \\ &= E[Y_{i-1}^2](1-Q_{i0}) + E[Y_{i-1}](E[X_i] - 1 + Q_{i0}) + \frac{2E[X_i^2] - 3E[X_i] + 1 - Q_{i0}}{6} \end{aligned}$$

The variance of waiting time for patients assigned to slot i is

$$\begin{aligned} Var[W_i] &= E[\delta_i]\sigma^2 + Var[\delta_i]\mu^2 \\ &= \left(E[Y_{i-1}](1-Q_{i0}) + \frac{E[X_i] - 1 + Q_{i0}}{2}\right)\sigma^2 \\ &\quad + \left(E[Y_{i-1}^2](1-Q_{i0}) + E[Y_{i-1}](E[X_i] - 1 + Q_{i0}) + \frac{2E[X_i^2] - 3E[X_i] + 1 - Q_{i0}}{6}\right)\mu^2 \\ &\quad - \left(E[Y_{i-1}](1-Q_{i0}) + \frac{E[X_i] - 1 + Q_{i0}}{2}\right)^2 \mu^2 \end{aligned}$$

Variance of queue length and number of patients in the system at the beginning of slot i

The variance of queue length and the variance of number of patients in the system at the beginning of slot i are:

$$\begin{aligned} Var[QL_i] &= Var[Y_{i-1}] + Var[X_i] \\ Var[NS_i] &= Var[Y_{i-1}] + Var[X_i] \end{aligned}$$

Appendix B. Proof of properties

Proof of Property 1 To calculate the expected overtime and the expected service time for the patients overflowing from slot $i-1$ to slot i, we use the memoryless property of exponential distribution. One of the overflowing patients will be with the provider at the beginning of the slot. For the exponential

distribution, the time spent in the previous slot will not have any effect on the expected service time in the current slot due to the memoryless property. If the service time distribution is not exponential, the time spent in the the previous slot must be considered. For example, if the service time has a general distribution with mean μ for all patients, then the expected service time of the overflowing patients will be calculated as follows:

$$\begin{aligned}
& \sum_{k\geq 1}\left(E[t_{i-1}^{rs}]+\sum_{1<k'\leq k}E[t_{i-1,k'}^{s}]\right)Pr\{Y_{i-1}=k\} \\
= & \sum_{k\geq 1}(E[t_{i-1}^{rs}]+(k-1)\mu)R_{i-1,k} \\
= & \sum_{k\geq 1}E[t_{i-1}^{rs}]R_{i-1,k}+\sum_{k\geq 1}k\mu R_{i-1,k}-\sum_{k\geq 1}\mu R_{i-1,k} \\
= & E[t_{i-1}^{rs}]\sum_{k\geq 1}R_{i-1,k}+\mu\sum_{k\geq 1}kR_{i-1,k}-\mu\sum_{k\geq 1}R_{i-1,k} \\
= & E[t_{i-1}^{rs}](1-R_{i-1,0})+\mu E[Y_{i-1}]-\mu(1-R_{i-1,0}) \\
= & (E[t_{i-1}^{rs}]-\mu)(1-R_{i-1,0})+\mu E[Y_{i-1}]
\end{aligned}$$

where $E[t_{i-1}^{rs}]$ is the expected remaining service time of the patient being served at the end of slot $i-1$ and $E[t_{i-1,k'}^{s}]$ is the expected service time for all other patients overflowing from slot $i-1$ to i. Because $E[t_{i-1}^{rs}]\leq\mu$, we have

$$\mu E[Y_{i-1}]+(E[t_{i-1}^{rs}]-\mu)(1-R_{i-1,0})\leq\mu E[Y_{i-1}],$$

implying that total expected service time for overflowing patients under general service time is not greater than that under exponential. □

Proof of Property 2 a. The expected number of arrivals to slot i in schedule $S=S^{n-2}+\Delta^{i1}+\Delta^{i4}$ is $E[X_i^n]=E[X_i^{n-2}]+p_1+p_4$ and is $E[X_i^n]=E[X_i^{n-2}]+p_2+p_3$ for schedule $S'=S^{n-2}+\Delta^{i2}+\Delta^{i3}$, which are equal to each other.

b. The expected increase of overflow from slot i in schedule $S=S^{n-2}+\Delta^{i1}+\Delta^{i4}$ is:

$$\begin{aligned}
E[Y_i^n-Y_i^{n-2}] = & \ 2p_1p_4Pr\{Y_i^n-Y_i^{n-2}=2|\mathcal{A}_{14}\}+p_1p_4Pr\{Y_i^n-Y_i^{n-2}=1|\mathcal{A}_{14}\} \\
& +p_1(1-p_4)Pr\{Y_i^n-Y_i^{n-2}=1|\mathcal{A}_1\}+(1-p_1)p_4Pr\{Y_i^n-Y_i^{n-2}=1|\mathcal{A}_4\}
\end{aligned}$$

where $Pr\{Y_i^n-Y_i^{n-2}=2|\mathcal{A}_{14}\}$ is the probability that the arrival of patients of types 1 and 4 leads to two more patients overflowing from slot i. This occurs when the number of patients served is less than or equal to the number of patients in slot i in schedule S^{n-2}. Thus,

$$Pr\{Y_i^n-Y_i^{n-2}=2|\mathcal{A}_{14}\}=Pr\{L_i\leq X_i^{n-2}+Y_{i-1}^{n-2}\}.$$

Similarly, $Pr\{Y_i^n - Y_i^{n-2} = 1|\mathcal{A}_1\}$ ($Pr\{Y_i^n - Y_i^{n-2} = 1|\mathcal{A}_4\}$) is the probability that the arrival of patients of type 1 (4) lead to one more patient overflowing from slot i, which is equal to $Pr\{L_i \leq X_i^{n-2} + Y_{i-1}^{n-2}\}$. So, we have

$$\begin{aligned}
E[Y_i^n - Y_i^{n-2}] &= 2p_1p_4Pr\{L_i \leq X_i^{n-2} + Y_{i-1}^{n-2}\} + p_1p_4Pr\{Y_i^n - Y_i^{n-2} = 1|\mathcal{A}_{14}\} \\
&\quad + p_1(1-p_4)Pr\{L_i \leq X_i^{n-2} + Y_{i-1}^{n-2}\} + (1-p_1)p_4Pr\{L_i \leq X_i^{n-2} + Y_{i-1}^{n-2}\} \\
&= (2p_1p_4 + p_1(1-p_4) + (1-p_1)p_4)Pr\{L_i \leq X_i^{n-2} + Y_{i-1}^{n-2}\} \\
&\quad + p_1p_4Pr\{Y_i^n - Y_i^{n-2} = 1|\mathcal{A}_{14}\} \\
&= (p_1 + p_4)Pr\{L_i \leq X_i^{n-2} + Y_{i-1}^{n-2}\} + p_1p_4Pr\{Y_i^n - Y_i^{n-2} = 1|\mathcal{A}_{14}\}
\end{aligned}$$

Similarly, the expected increase of overflow from slot i in schedule $S' = S^{n-2} + \Delta^{i2} + \Delta^{i3}$ is:

$$E[Y_i'^n - Y_i^{n-2}] = (p_2 + p_3)Pr\{L_i \leq X_i^{n-2} + Y_{i-1}^{n-2}\} + p_2p_3Pr\{Y_i'^n - Y_i^{n-2} = 1|\mathcal{A}_{23}\}$$

Since $p_1 + p_4 = p_2 + p_3$ and $p_1p_4 \leq p_2p_3$, $E[Y_i^n - Y_i^{n-2}] \leq E[Y_i'^n - Y_i^{n-2}]$. Thus, schedule S gives smaller expected increase in overflow than schedule S', implying $E[Y_i^n] \leq E[Y_i'^n]$.

c. By equation (2), the expected waiting time of schedule S is calculated as:

$$E[W_i] = \frac{(2E[Y_{i-1}^{n-2}] - 1)(1 - Q_{i0}^{n-2}(1-p_1)(1-p_4)) + E[X_i^{n-2}] + p_1 + p_4}{2}\mu$$

The expected waiting time of schedule S' is:

$$E[W_i'] = \frac{(2E[Y_{i-1}^{n-2}] - 1)(1 - Q_{i0}^{n-2}(1-p_2)(1-p_3)) + E[X_i^{n-2}] + p_2 + p_3}{2}\mu$$

Since $p_1 + p_4 = p_2 + p_3$ and $(1-p_1)(1-p_4) \leq (1-p_2)(1-p_3)$, $E[W_i] > E[W_i']$ when $2E[Y_{i-1}^{n-2}] - 1 > 0$.
$\Rightarrow$ If $E[Y_{i-1}^{n-2}] < 1/2$, then $E[W_i] < E[W_i']$. □

Proof of Property 3 Let S^{n-1} be the initial schedule and n^{th} patient of type j is assigned to slot i^*.

a. The expected overflow for slot i ($i \geq i^*$) is calculated as follows:

$$E[Y_i^n] = E[Y_i^{n-1}] + E[Y_i^n - Y_i^{n-1}] = E[Y_i^{n-1}] + p_j\left\{\prod_{i^* \leq l \leq i} Pr\{L_l \leq X_l^{n-1} + Y_{l-1}^{n-1}\}\right\}$$

When a patient is scheduled for slot i^*, the expected overflow from slot i ($i \geq i^*$) increases as p_j increases.

b. The expected waiting time for slot i^* is calculated as follows:

$$\begin{aligned}
E[W_{i^*}^n] &= \frac{(2E[Y_{i^*-1}^{n-1}] - 1)(1 - Q_{i^*0}^{n-1}(1-p_j)) + E[X_{i^*}^{n-1}] + p_j}{2}\mu \\
&= E[W_{i^*}^{n-1}] + \frac{(2E[Y_{i^*-1}^{n-1}]Q_{i^*0}^{n-1} - Q_{i^*0}^{n-1} + 1)p_j}{2}\mu
\end{aligned}$$

The expected waiting time at slot i^* increases as p_j increases. For slots $i > i^*$, the expected waiting time ($E[W_i^n]$) increases because $E[Y_{i-1}^n]$ increases (due to Property 3.a).

c. The expected queue length at the beginning of slot i^* is calculated as follows:

$$E[QL_{i^*}^n] \quad = \quad E[Y_{i-1}^{n-1}] + E[X^{n-1}] + p_j - 1 + R_{i-1,0}^{n-1} Q_{i0}^{n-1}(1-p_j)) = E[QL_{i^*}^{n-1}] + p_j(1 - R_{i-1,0}^{n-1} Q_{i0}^{n-1})$$

Since $(1 - R_{i-1,0}^{n-1} Q_{i0}^{n-1}) \geq 0$, $E[QL_{i^*}^n]$ increases as p_j increases. For slots $i > i^*$, the expected queue length at the beginning of slot i is $E[Y_{i-1}^n] + E[X^{n-1}] - 1 + R_{i-1,0}^n Q_{i0}^{n-1}$. As p_j increases, $E[Y_{i-1}^n]$ increases (due to Property 3.a) and $R_{i-1,0}^n$ decreases. Therefore, $E[QL_i^n]$ increases.

d. The expected number of patients in the system at the beginning of slot i^* is calculated as follows:

$$E[NS_{i^*}^n] \quad = \quad E[Y_{i-1}^{n-1}] + E[X^{n-1}] + p_j = E[NS_{i^*}^{n-1}] + p_j$$

$E[NS_{i^*}^n]$ increases as p_j increases. For slots $i > i^*$, the expected number of patients in the system at the beginning of slot i increases, because $E[Y_{i-1}^n]$ increases (Property 3.a) □

Proof of Property 5 Let S^{n-1} be the initial schedule and n^{th} patient of type j is assigned to slot i^*. We define $\mathcal{A}_j$ as the arrival of patient j to slot i^*. When a patient of type j is assigned to slot i_1, the expected overflow at the end of the day is calculated as:

$$\begin{aligned} E[Y_I^n] &= E[Y_I^{n-1}] + E[Y_I^n - Y_I^{n-1}] \\ &= E[Y_I^{n-1}] + p_j Pr\{Y_I^n - Y_I^{n-1} = 1 | \mathcal{A}_j\} \\ &= E[Y_I^{n-1}] + p_j \left\{ \prod_{i_1 \leq i \leq I} Pr\{L_i \leq X_i^{n-1} + Y_{i-1}^{n-1}\} \right\} \end{aligned}$$

When the patient is assigned to slot i_2, the expected overflow at the end of the day is:

$$E[Y_I^{n'}] \quad = \quad E[Y_I^{n-1}] + p_j \left\{ \prod_{i_2 \leq i \leq I} Pr\{L_i \leq X_i^{n-1} + Y_{i-1}^{n-1}\} \right\}$$

Since $i_1 < i_2$, $\left\{ \prod_{i_1 \leq i \leq I} Pr\{L_i \leq X_i^{n-1} + Y_{i-1}^{n-1}\} \right\} \leq \left\{ \prod_{i_2 \leq i \leq I} Pr\{L_i \leq X_i^{n-1} + Y_{i-1}^{n-1}\} \right\}$. Thus, $E[Y_I]$ is smaller when patient is assigned to slot i_1 rather than slot i_2. □