INFORMS Journal on Computing
Articles in Advance, pp. 1–17
ISSN 1091-9856 | EISSN 1526-5528

http://dx.doi.org/10.1287/ijoc.1110.0482

Dynamic Appointment Scheduling of a Stochastic Server with Uncertain Demand

S. Ayca Erdogan
Graduate Program in Operations Research, North Carolina State University, Raleigh, North Carolina, 27695
saerdoga@ncsu.edu

Brian Denton
Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina, 27695, bdenton@ncsu.edu

We formulate and solve two new stochastic linear programming formulations of appointment scheduling problems that are motivated by the management of health services. We assume that service durations and the number of customers to be served on a particular day are uncertain. In the first model, customers may fail to show up for their appointments ("no-show"). This model is formulated as a two-stage stochastic linear program. In the second model, customers are scheduled dynamically, one at a time, as they request appointments. This model is formulated as a multistage stochastic linear program with stages defined by customer appointment requests. We analyze the structure of the models and adapt decomposition-based algorithms to solve the problems efficiently. We present numerical results that illustrate the impact of uncertainty on dynamic appointment scheduling, and we identify useful insights that can be applied in practice. We also present a case study based on real data for an outpatient procedure center.

Key words: appointment scheduling; stochastic programming; health care
History: Accepted by Allen Holder, Area Editor for Applications in Biology, Medicine, and Health Care; received May 2010; revised January 2011, June 2011; accepted July 2011. Published online in *Articles in Advance*.

1. Introduction

The problem of appointment scheduling to a stochastic server is well known and widely studied in the literature (Welch and Bailey 1952, Mercer 1960, Ho and Lau 1992). It is commonly assumed that service times are random, and a deterministic schedule of appointment times is selected to optimize competing performance criteria, including expected customer waiting time, server idle time, and overtime. This problem differs from typical single server queuing models in two important ways. First, the scheduling horizon is finite, typically limited by the number of customers seen on a particular day. Second, customers arrive deterministically according to a defined schedule of appointment times. Thus, the focus is on the transient behavior as opposed to steady state and stochastic arrival assumptions that are common in the queuing literature.

In this article we relax two common assumptions in the appointment scheduling literature. First, we assume customers may fail to show up ("no-show") at their assigned time. This is motivated by the common occurrence of no-shows in outpatient health care environments (Lee et al. 2005). In outpatient clinics, no-shows have been reported to range from 12% to 42% of all appointments, making efficient management of outpatient clinics' resources difficult (Deyo and Inui 1980, Moore et al. 2001). Second, we assume some customers request appointments dynamically over time, and the exact number to be scheduled for a particular day is not known with certainty until the day of service. This is arguably the case for most appointment-based service systems.

Our study is motivated in part by problems faced by health-care providers who schedule a nominal number of routine appointments in advance of a given day and then must accommodate some high-priority add-on patients that may arrive on short notice. In surgery delivery systems, for instance, urgent add-on cases arise on short notice and create the need to dynamically update schedules to accommodate these high-priority cases (Gerchak et al. 1996, Dexter et al. 2004). Other applications of this problem have been identified in the literature, including material handling, scheduling cargo ports, and outpatient services. It is frequently the case that appointments are quoted dynamically to customers with imperfect knowledge of total demand.

The main contributions of this paper are as follows: the models we propose include considerations that

are representative of many types of health-care environments and that have not yet been well studied. First, we propose a two-stage stochastic linear programming (2-SLP) model for static appointment scheduling in the presence of no-shows. Next, we present a novel formulation of a multistage stochastic linear program (M-SLP) that considers dynamic scheduling of uncertain add-on customers that may request appointments. We present insights into optimal scheduling policies in the presence of uncertain demand for services, including results based on a real problem involving scheduling of an outpatient practice.

From a methodological perspective, we discuss the structural properties of the model we propose and novel adaptations of decomposition methods for solving it. We show that relaxations of the M-SLP provide easy-to-compute valid inequalities that can be used to accelerate decomposition methods. We also show that the M-SLP model can be decomposed into a set of two-variable linear programs (LPs) that can be solved efficiently. We further show that the structure of the M-SLP is well suited to a customized multicut implementation of the nested decomposition method. Finally, numerical experiments are used to compare the performance of several alternative decomposition-based methods. We also perform a series of numerical experiments to illustrate important insights regarding the influence of uncertainty in customer load on the server.

The remainder of this article is organized as follows: Section 2 is a brief review of related literature. In §3, the model formulations for no-shows and dynamic scheduling are presented. Section 4 discusses the structure of our models and the methodology used to solve them. In §5, we present the results of our computational experiments. Finally, in §6, we summarize our main conclusions.

2. Literature Review

Scheduling customers to a stochastic server has been widely studied. Many of the studies have been in the context of outpatient clinics and other appointment-based health-care environments (therefore in the following review, we use "patients" and "customers" interchangeably). Appointment scheduling is a challenging problem for many reasons, including the uncertainty in arrival times and service durations, preferences of the patients and the providers, and the presence of multiple and competing criteria. Gupta and Denton (2008) pointed out the complexity of the appointment scheduling problem by clarifying recent issues and challenges in primary and specialty care as well as surgery settings. The authors referred to several complicating factors related to uncertainty in patient arrivals and/or requests for appointments. Cayirli and Veral (2003) also provided a comprehensive literature survey in outpatient appointment scheduling, classifying the research by methodology. Erdogan and Denton (2011) provided a recent review of the literature related to surgery scheduling. All of these studies point out that uncertainty in patient demand, such as no-shows, urgent patients, and emergencies, is an important consideration.

Because of the difficulty of finding an analytical solution for problems with more than two customers, much of the existing literature on appointment scheduling is based on either queueing theory or discrete event simulation. Queueing studies generally require restrictive assumptions, including equal appointment intervals, independent and identical service times, and an infinite number of customers (Mercer 1960, 1973; Jansson 1966).

Studies based on discrete event simulation models, on the other hand, relax these assumptions. For example, Vissers and Wijngaard (1979) were among the first to use a simulation model to study outpatient clinics. They studied the experimental design of a simulation model for an outpatient clinic that aims to minimize the patient waiting time and doctor idle time. They proposed a simulation model with five variables: mean consultation time, coefficient of variation of the consultation time, mean system earliness, standard deviation of patient punctuality, and the number of appointments. Ho and Lau (1992) used a simulation model to evaluate many different scheduling rules for various scheduling environments characterized by different combinations of patient no-show probability, coefficients of variation of service times, and numbers of patients per day.

Some researchers have studied the appointment scheduling problem with the goal of optimizing some weighted combination of expected customer waiting and server idle time. Weiss (1990) provided a closed-form solution to find the optimal estimated appointment times for two patients. Wang (1993) used *phase-type* distributions to obtain closed-form expressions for expected customer waiting time and server idle time for problems involving more than two customers. Using these closed-form expressions he computed schedules that minimize a weighted sum of expected waiting and idling. In addition to the queuing literature, Bosch and Dietz (2001) proposed an efficient algorithm for optimizing the appointment schedule that fathoms the solution space of the possible schedules using the piecewise convex structure of the cost function. Denton and Gupta (2003) formulated a 2-SLP model and exploited the problem structure to develop an algorithm that provides bounds on the optimal solution. Robinson and Chen (2003) used

Monte Carlo integration to find approximate optimal appointment times for a stochastic server.

Many researchers have considered customer no-shows in appointment scheduling. For example, Brahimi and Worthington (1991) studied the problem in the context of outpatient appointment systems. They applied a queuing model that considers no-shows as well as uncertain punctuality of patients. Hassin and Mendel (2008) studied the effects of no-shows on the performance of a single-server with exponential service times. Kaandorp and Koole (2007) developed a local search procedure that they showed converges to the optimal solution due to the multimodularity property of the outpatient appointment scheduling problem with no-shows. In their model they assumed homogeneous no-show probabilities. Zeng et al. (2009) extended their work by considering heterogenous no-show probabilities.

Wang (1993) investigated a dynamic scheduling problem in which an unscheduled customer must be added in to the schedule. However, this model is not truly dynamic in that the initial schedule does not anticipate the possibility of an additional customer arrival. Klassen and Rohleder (1996) also studied the dynamic nature of appointment scheduling systems by considering urgent arrivals. They developed a simulation model that leaves open slots in the schedule for possible urgent customers. They concluded that leaving open slots at the beginning of the day for urgent customers decreases customer waiting time but also decreases the percentage of urgent customers served. However, leaving open slots at the end of the day improves both the percentage of the urgent customers served and the server idle time.

This article differs from the aforementioned literature in the following ways: Although several authors have considered heuristics (see, for example, Wang 1993, Muthuraman and Lawley 2008, Robinson and Chen 2003) for appointment scheduling problems with an uncertain number of customers, to our knowledge we present the first formulation of a stochastic programming model of the dynamic appointment scheduling problem to compute optimal appointment times. We present insights into optimal scheduling policies in the presence of no-shows and optimal dynamic scheduling policies illustrating the differences relative to their static counterparts. We also propose and evaluate several new methods that take advantage of the structure of these problems. These methods may also be applicable to future extensions.

3. Model Formulation

We begin by presenting an extension to the 2-SLP model proposed by Denton and Gupta (2003) to incorporate customer no-shows. Next, we present our M-SLP formulation of the dynamic appointment scheduling problem. The objective function in each of the two models is to minimize a weighted sum of costs of expected customer waiting time and overtime with respect to a defined length of day (referred to as *session length* below). Both models assume punctual arrivals for those customers that do show up. Furthermore, both models assume a fixed sequence of arrivals and a first-come-first-served (FCFS) queue discipline.

3.1. Appointment Scheduling in the Presence of No-Shows

The problem addressed here is finding the optimal arrival times for n customers to visit a stochastic server. Service times are assumed to be random variables, and the objective is to minimize a weighted sum of expected customer waiting time and expected overtime with respect to an established session length, d. Customers, $i = 1, \ldots, n$, have no-show probabilities, p_i. We use the following additional notation, where uppercase indicates random variables and boldface is used to denote vectors.

Model Parameters

n: number of customers to be scheduled.
ω: index for service duration and no-show scenarios.
$\mathbf{A}(\omega)$: random vector of indicators for customer arrival (1) or no-show (0), $\mathbf{A}(\omega) \in \Re^n$.
$\mathbf{Z}(\omega)$: vector of random service durations for n customers, $\mathbf{Z}(\omega) \in \Re^n$.
d: session length to complete all customers before overtime is incurred, $d \in \Re$.
$\mathbf{p}$: vector of probabilities of no-show, $\mathbf{p} \in \Re^n$.
$\mathbf{c^w}$: vector of waiting time cost coefficients for n customers, $\mathbf{c^w} \in \Re^n$.
c^ℓ: cost coefficient for overtime, $c^\ell \in \Re$.

Decision Variables

$\mathbf{x}$: vector of time allowances for the first $n-1$ customers (interarrival times for n customers), $\mathbf{x} \in \Re^{\mathbf{n-1}}$.
$\mathbf{w}(\omega)$: vector of customer waiting times, $\mathbf{w}(\omega) \in \Re^n$.
$\mathbf{s}(\omega)$: vector of server idle times between consecutive customers, $\mathbf{s}(\omega) \in \Re^n$.
$\ell(\omega)$: overtime with respect to session length d, $\ell(\omega) \in \Re$.

The vector of time allowances $\mathbf{x} \in \Re^{n-1}$ denotes first stage decisions made in advance of the observation of random service durations and no-shows (note that $\mathbf{x}$ is $n-1$ dimensional because x_i denotes interarrival times between the n customers). The scheduled appointment time for customer i is the sum of job allowances from 1 to $i-1$. Thus, customer 1

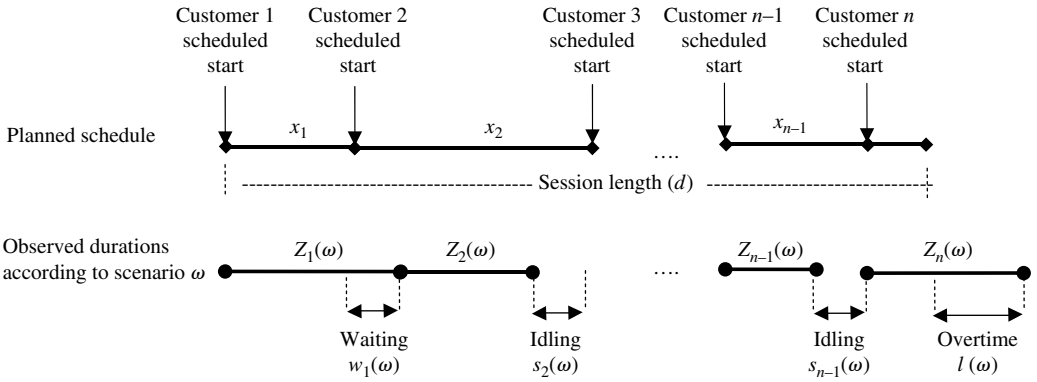

Figure 1 Planned Schedule and Observed Schedule for a Single Scenario Problem

arrives at time 0, customer 2 at time x_1, customer 3 at time $x_1 + x_2$, and so on. The random service time durations vector, $\mathbf{Z}(\omega)$, has support $\Xi \in \Re^n$, and the possible collective outcomes of service times (scenarios) are indexed by $\omega \in \Omega$. The vectors $\mathbf{w}(\omega), \mathbf{s}(\omega) \in \Re^n$, $\ell(\omega) \in \Re$, denote the second-stage (recourse) decisions made after the observation ω of random service durations. The parameters $\mathbf{c^w} \in \Re^n$ and $c^\ell \in \Re$ denote the cost per unit time for waiting and overtime, respectively. We assume that $\mathbf{c^w} \geq 0$ and $c^\ell \geq 0$. Figure 1 depicts the decision variables and the parameters on a sketch of planned schedule with allowances (x_i) and observed service durations ($Z_i(\omega)$), waiting ($w_i(\omega)$), idling ($s_i(\omega)$), and overtime ($\ell(\omega)$) for a single scenario ω.

Commonly considered criteria for determining optimal time allowances include customer waiting time, server idle time, and overtime, which can be written as follows:

$$w_i(\omega) = (w_{i-1}(\omega) + Z_{i-1}(\omega) - x_{i-1})^+, \quad i = 2, \ldots, n, \quad (1)$$

$$s_i(\omega) = (-w_{i-1}(\omega) - Z_{i-1}(\omega) + x_{i-1})^+, \quad i = 2, \ldots, n, \quad (2)$$

$$\ell(\omega) = \left(w_n(\omega) + Z_n(\omega) + \sum_{i=1}^{n-1} x_i - d \right)^+, \quad (3)$$

where $(\cdot)^+$ indicates $\max(\cdot, 0)$. The waiting time and server idle time associated with the first customer is zero ($w_1(\omega) = s_1(\omega) = 0, \forall \omega$); i.e., the first customer receives service as soon as he or she arrives. The optimal appointment schedule is defined by the following unconstrained minimization problem:

$$\min_{\mathbf{x}} \left\{ \sum_{i=1}^{n} c_i^w E_\omega[w_i(\omega)] + c^\ell E_\omega[\ell(\omega)] \right\}. \quad (4)$$

Some authors have considered a weighted sum of expected server idle time, overtime, and customer waiting time as the objective function; for simplicity, we consider only customer waiting time and overtime as in (4). This is justified by the following proposition adapted from Denton and Gupta (2003).

Proposition 1. *When $d = 0$, expected idle time is equal to expected overtime minus expected total service time*:

$$\sum_{i=1}^{n} E[s_i] = E[\ell] - \sum_{i=1}^{n} \mu_i$$

where μ_i is the expected value of the service time distribution.

The proposition in Denton and Gupta (2003, p. 1007) states that "expected idle time is equal to the difference between two sums: the sum of expected tardiness and the session length, and the sum of average job durations and expected earliness"; i.e.,

$$\sum_{i=1}^{n} E[s_i] = (E[\ell] + d) - \left(E[G] + \sum_{i=1}^{n} \mu_i \right),$$

where $E[G]$ is defined as expected earliness (please see Denton and Gupta 2003 for more detailed proof). Without loss of generality, by setting $d = 0$, which corresponds to minimizing makespan, the earliness variable disappears, and the expression in Propostion (1) is obtained.

Formulation (4) can be modified to account for no-shows as follows. Define random service durations as $\hat{Z}_i(\omega) = A_i(\omega) Z_i(\omega)$, where

$$A_i(\omega) = \begin{cases} 0 & \text{with probability } p_i, \\ 1 & \text{with probability } 1 - p_i. \end{cases} \quad (5)$$

$\hat{Z}_i(\omega)$ is a random variable representing the service duration for customer i, given the customer shows up for his appointment, which occurs with probability $1 - p_i$. Thus, this random variable differs from the original service duration variable $Z_i(\omega)$. In general, there is no closed-form expression for the solution to (4). Denton and Gupta (2003) discuss the properties of an equivalent 2-SLP formulation that can be used

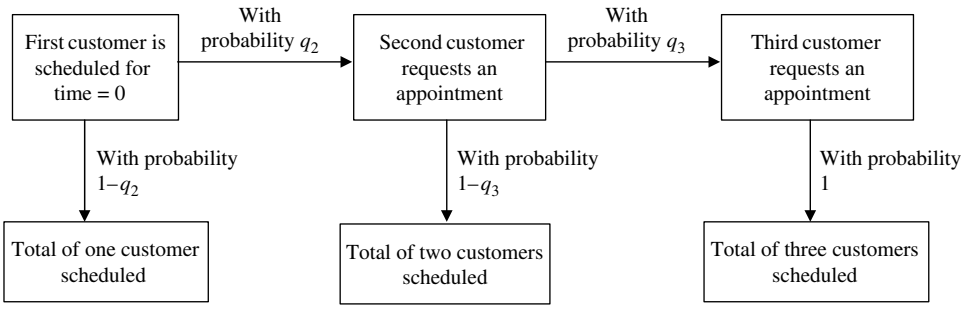

Figure 2 **Illustration of the Scheduling Problem with Probabilistic Arrival of Customers Given $n^U = 3$**

to achieve significant computational advantages. Similarly, our model can also be formulated as a 2-SLP as follows:

(NS-ASP)

$$\begin{aligned}
\min\ & E_\omega\left[\sum_{i=2}^{n} c_i^w A_i(\omega) w_i(\omega) + c^\ell \ell(\omega)\right] \\
\text{s.t.}\ & w_2(\omega) \geq \hat{Z}_1(\omega) - x_1 \quad \forall\, \omega \\
& -w_2(\omega) + w_3(\omega) \geq \hat{Z}_2(\omega) - x_2 \quad \forall\, \omega \\
& \qquad \ddots\ \ddots\ \vdots \\
& -w_{n-1}(\omega) + w_n(\omega) \geq \hat{Z}_{n-1}(\omega) - x_{n-1} \quad \forall\, \omega \\
& -w_n(\omega) + \ell(\omega) \geq \hat{Z}_n(\omega) + \sum_{i=1}^{n-1} x_i - d \quad \forall\, \omega \\
& \mathbf{x} \geq \mathbf{0}, \quad \mathbf{w}(\omega), \quad \ell(\omega) \geq \mathbf{0}, \quad \forall\, \omega.
\end{aligned}$$

We refer to the above model as the no-show appointment scheduling problem (NS-ASP). Note that (NS-ASP) has *complete recourse* because the second stage is feasible for any $\mathbf{x} \in \Re^{n-1}$.

3.2. Dynamic Appointment Scheduling

The second model we propose assumes customers are scheduled dynamically as they call to request an appointment. Appointment requests are probabilistic, i.e., the total number to be scheduled is not known with certainty, and there is a maximum of n^U customers that will be scheduled (n^U as an upper bound on the capacity of the system). Let q_i be the probability of an appointment request by customer i given that customer $i-1$ is scheduled. We assume that customers are scheduled based on FCFS in the sequence of their appointment requests.

3.2.1. Simple Examples. To illustrate the nature of our problem, we consider two simple examples. We assume that there is at least one customer in the system.

Example 1 ($n^U = 2, q_2 = 1$). This represents the case where two customers will be scheduled with certainty. This problem corresponds to the newsvendor problem when $d = 0$.

Example 2 ($n^U = 3, q_2 > 0, q_3 > 0$). In this case one customer will certainly be scheduled. With conditional probabilities q_2 and q_3, customers 2 and 3 may request appointments. For this example there are three customer arrival scenarios:

1. The first customer is scheduled. The second and third customers do not request appointments.
2. The second customer requests an appointment after the first customer is scheduled. The third customer does not request an appointment.
3. The second customer requests an appointment after the first customer is scheduled. The third customer requests an appointment after the second customer is scheduled.

The sequential nature of the uncertainty in the customer requests in Example 2 is illustrated in Figure 2. In contrast to Example 1, a closed-form expression for the solution to this problem is not easily obtained.

In Example 2 uncertainty is resolved sequentially as appointment requests arise, and appointments must be scheduled with imperfect information about the number of customers and their service times. Therefore each request is treated as an additional stage in the decision-making process. At each stage, j, the time allowance decision, x_j, is made for customer j without perfect knowledge of the number of additional future appointment requests. To formulate our model we use similar notation to that of (NS-ASP) with an additional index, $j = 1, \ldots, n^U$, to denote the stage. Thus, $w_{j,i}(\omega)$ is the waiting time of the ith customer on the day of the service, given j customers request appointments. We let ω_j index service duration scenarios for stage j. Similarly, we let $\ell_j(\omega_j)$ denote the overtime given j customers are scheduled. Thus for Example 2, customer arrival scenario 1 and service duration scenario ω_1 can be written as

$$\begin{aligned}
w_{1,1}(\omega_1) &= 0, \\
\ell_1(\omega_1) &= (Z_1(\omega_1) - d)^+.
\end{aligned} \tag{6}$$

For customer arrival scenario 2 and service duration scenario ω_2,

$$\begin{aligned} w_{2,1}(\omega_2) &= 0, \\ w_{2,2}(\omega_2) &= (Z_1(\omega_2) - x_1)^+, \\ \ell_2(\omega_2) &= (x_1 + w_{2,2}(\omega_2) + Z_2(\omega_2) - d)^+. \end{aligned} \tag{7}$$

For customer arrival scenario 3 and service duration scenario ω_3,

$$\begin{aligned} w_{3,1}(\omega_3) &= 0, \\ w_{3,2}(\omega_3) &= (Z_1(\omega_3) - x_1)^+, \\ w_{3,3}(\omega_3) &= (w_{3,2}(\omega_3) + Z_2(\omega_3) - x_2)^+, \\ \ell_3(\omega_3) &= (w_{3,3}(\omega_3) + Z_3(\omega_3) + x_1 + x_2 - d)^+. \end{aligned} \tag{8}$$

Thus the three arrival schedules define the waiting time and overtime associated with one, two, and three scheduled customers, respectively. The indices $\omega_1 \in \Omega_1$, $\omega_2 \in \Omega_2$, and $\omega_3 \in \Omega_3$ refer to service time scenarios given one, two, and three customers are scheduled, respectively.

3.2.2. Dynamic Appointment Scheduling Model. The appointment request scenarios for the dynamic appointment scheduling process are represented by the tree in Figure 3. In the figure, nodes represent the number of customers in the system. Solid nodes denote the number of scheduled customers and a state in which the schedule is waiting for future appointment requests. The dashed nodes define the day of service given that a certain number of customers requested appointments and the system terminated without another appointment request. Our model starts with two customers because the solution of the one customer problem is trivial. Starting with two customers, customer 3 requests an appointment with probability q_3, and with probability $1 - q_3$, no additional customers are scheduled. Given a third customer requests an appointment, a fourth customer will request an appointment with probability q_4, and so on.

We formulate this model as the following unconstrained optimization problem:

$$\min_{x_1}\left\{(1-q_3)Q_1(x_1) + \min_{x_2}\left\{q_3(1-q_4)Q_2(x_2) + \cdots + \min_{x_{n^U-1}}\left\{\prod_{i=3}^{n^U}(q_i Q_{n^U-1}(x_{n^U-1}))\right\}\cdots\right\}\right\}, \tag{9}$$

where $Q_j(x_j) = E_{\omega_j}[Q_j(x_j, \omega_j)]$ denotes the expected cost given that $j+1$ customers request appointments (note that $Q_j(x_j)$ corresponds to $j+1$ customers because x_j is the interarrival time between customers j and $j+1$). We refer to $Q_j(x_j, \omega_j)$ as the *terminal subproblem* for stage j under service duration scenario ω_j (represented by dashed nodes in Figure 3). We refer to this as terminal because it represents the case in which no additional customers beyond $j+1$ request appointments. Although not explicitly denoted in the formulation, it is implied that decision x_j is made prior to knowledge of whether customer $j+1$ (or additional customers) will request an appointment. We use this implicit definition to simplify the notation rather than explicitly write a series of linking constraints between stages. The terminal subproblem for stage j can be written as

$$Q_j(x_j, \omega_j) = \min_{\mathbf{w},\mathbf{s},\ell}\left\{\sum_{i=2}^{j+1} c_i^w w_{j,i}(\omega_j) + c^\ell \ell_{j+1}(\omega_j)\right\}$$

s.t.

$$\begin{aligned} w_{j,2}(\omega_j) &\geq Z_1(\omega_j) - x_1, \\ -w_{j,2}(\omega_j) + w_{j,3}(\omega_j) &\geq Z_2(\omega_j) - x_2, \\ &\ddots \\ -w_{j,j}(\omega_j) + w_{j,j+1}(\omega_j) &\geq Z_j(\omega_j) - x_j, \\ -w_{j,j+1}(\omega_j) + \ell_{j+1}(\omega_j) &\geq Z_{j+1}(\omega_j) + \sum_{i=1}^{j} x_i - d, \\ w_{j,i}(\omega_j) \geq 0 \quad \forall i,\ \ell_j(\omega_j) &\geq 0. \end{aligned} \tag{10}$$

Formulation (9) can be formulated recursively, with $R_j(x_j)$ denoting the expected *cost-to-go* given that

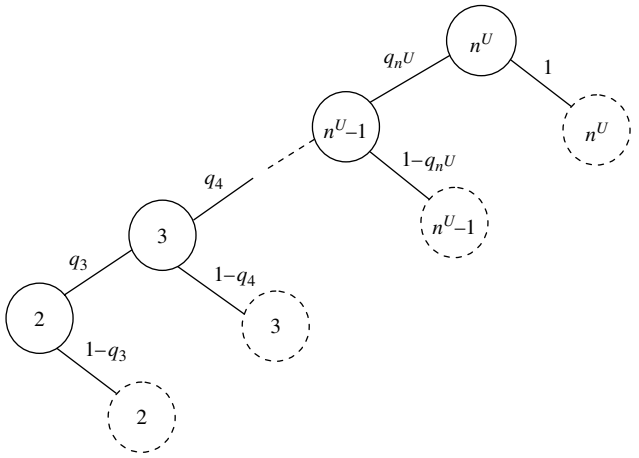

Figure 3 Tree of Scenarios for M-SLP Problem with n^U Customers

Note. Solid nodes denote stages in which additional customer appointment requests are pending, and dashed nodes define the day of service given a certain number of customer arrivals.

additional customers may arrive as follows:

$$R_j(x_j) = \min_{x_{j+1}} \{(1-q_{j+2})Q_j(x_j) + q_{j+2}R_{j+1}(x_{j+1})\}. \quad (11)$$

The recursion terminates at the last stage, $n^U - 1$, with $R_{n^U-1}(x_{n^U-1}) = Q_{n^U-1}(x_{n^U-1})$. Thus (9) can be expressed as

$$\text{(D-ASP)} \qquad \min_{x_1} R(x_1).$$

We refer to the above M-SLP as the dynamic appointment scheduling problem (D-ASP). We discuss several ways to take advantage of the recursive structure of (D-ASP) in §4.

It is worth noting some special cases of (D-ASP) that correspond to specific applications. First, in some applications it may be appropriate to assume a certain minimum number of customers arrives with certainty, which is equivalent to defining a lower bound on the number of customers that will be scheduled. This assumption is motivated by health-care applications such as hospital-based colonoscopy practices, where a certain minimum number of patients is scheduled in advance (outpatients) and some uncertain number of urgent add-on cases is scheduled on short notice (inpatients). It is also representative of a common primary care appointment scheduling process called *advanced access* (Murray and Tantau 2000), in which some patients are booked in advance and some urgent patients call for appointments on the day they want to be seen.

For simplicity, in (D-ASP) we have not considered no-shows; however, (NS-ASP) and (D-ASP) could easily be integrated to include the possibility of no-shows, which are common in both of the dynamic scheduling applications described above. It is also worth noting that NS-ASP is a special case of D-ASP when $q_i = 1$ and $\forall\, i$ and Z_is are as defined in (5).

Thus, some of the methods we develop to take advantage of the structure of (D-ASP) are also directly applicable to (NS-ASP) and the standard static appointment scheduling problem (Denton and Gupta 2003).

3.3. Motivation for FCFS Assumption

In this section we provide motivation, based on a stylized example, for the assumption that patients are scheduled FCFS in order of their appointment requests.

Proposition 2. *For $n^U = 2$ with i.i.d. service durations, if the second customer requests an appointment with probability q, FCFS is optimal.*

Proof. Let the optimal solutions for FCFS and scheduling the second customer first (last come, first serve, or LCFS) be x_1^F and x_1^L, respectively, and the optimal objective function values, $z_1(x_1^F)$ and $z_2(x_1^L)$, respectively. By convexity of the expectation of waiting and overtime costs, it follows that

$$\begin{aligned} z_1(x_1^F) &\le z_1(x_1^L) \\ &= (1-q)c^\ell E[(Z(\omega)-d)^+] + q(c^w E[(Z(\omega)-x_1^L)^+] \\ &\quad + c^\ell E[(Z(\omega)-x_1^L)^+ + Z(\omega) + x_1^L - d)^+]) \quad (12) \\ &\le (1-q)c^\ell E[(Z(\omega)+x_1^L-d)^+] \\ &\quad + q(c^w E[(Z(\omega)-x_1^L)^+] \\ &\quad + c^\ell E[((Z(\omega)-x_1^L)^+ + Z(\omega) + x_1^L - d)^+]) \quad (13) \\ &= z_2(x_1^L). \quad \square \quad (14) \end{aligned}$$

Thus if the two customers are identical in their service distributions and the waiting time cost coefficients, the appointment request sequence should be FCFS order. We have not been able to prove this for $n > 2$, but numerical experiments and intuition suggest that FCFS ordering is also optimal for those cases.

4. Structural Properties and Solution Methodology

Because of the potentially large size of the stochastic programs we propose, taking advantage of the problem structure is important. Furthermore, our initial study motivates additional problems where (NS-ASP) and (D-ASP) are subproblems, such as problems with multiple servers involving patient to server assignment decisions. In this section we concentrate on (D-ASP), the more computationally challenging of the two models, but many of the properties can also be exploited to solve (NS-ASP).

Nested decomposition is a common approach to take advantage of the recursive structure of M-SLPs (Birge 1985). It is based on outer linearization of the recourse function, $R_j(x_j)$, at each stage j. At each stage a solution, x_j, is generated by solving a relaxed master problem, which is a linear program representing the expected waiting and overtime cost for scheduled customers and the expected cost-to-go for future stages. At each (terminal) stage, subproblems based on a number of service time scenarios, indexed by ω_j, are solved given the solution to the master problem, x_j. The dual solutions to the subproblems are used to generate supporting hyperplanes (called *optimality cuts*) for the stage j recourse function. These cuts are added sequentially at each stage until the relaxed master problem converges to the optimal solution. The following subsections describe a number of opportunities to improve efficiency of the nested decomposition method (ND) for formulation (D-ASP).

4.1. Subproblem Structure

The dual solution to terminal subproblems, $Q_j(x_j, \omega_j)$, can be computed efficiently using the following backward recursion:

$$\pi_{j,i}(x_j,\omega_j)=\begin{cases}0 & w_{j,i+1}(\omega_j)=0,\\ c^w_{i+1}+\pi_{j,i+1}(x_j,\omega_j) & \\ & w_{j,i+1}(\omega_j)>0,\end{cases} \quad (15)$$

$$\text{for } i=1,\dots,j-1, \quad \text{and}$$

$$\pi_{j,j}(x_j,\omega_j)=\begin{cases}0 & \ell_j(\omega_j)=0,\\ c^\ell & \ell_j(\omega_j)>0.\end{cases} \quad (16)$$

This closed-form expression for the dual allows efficient generation of optimality cuts at each stage of the ND algorithm (Denton and Gupta 2003). The master problem of each stage (except the last stage) includes another subproblem that represents the expected cost-to-go, $R_j(x_j)$, for the remaining future stages. Each master problem is based on the following equivalent (outer linearization) formulation:

$$\min\{\theta_j \mid \theta_j \ge R_j(x_j)\}, \quad (17)$$

in which the decision variables are x_j and θ_j. Thus, the master problem for stage j is a two-variable LP with optimality cuts at stage j of the form:

$$\theta_j + E_j x_j \ge e_j - \sum_{k=1}^{j-1} E_k x_k,$$

where E_j are the cut coefficients and e_j is the right-hand-side value of the optimality cut generated at each iteration of the decomposition algorithm. These values are calculated using the dual solution to the subproblems at each iteration. The reader is referred to Birge and Louveaux (1997, §7.1, pp. 234–237) for more information on generating the optimality cuts. Substituting the known values for $x_1, x_2, \dots, x_{j-1}$, determined in stages $1, 2, \dots, j$, the cut then takes the following general form for each stage, j, at each iteration, ν, of the ND method:

$$\theta_j \ge \alpha_\nu x_j + \beta_\nu. \quad (18)$$

The master problem at iteration ν has the following general form:

$$\min\{\theta_j \mid \theta_j \ge \alpha_k x_j + \beta_k, k=1,\dots,\nu\}. \quad (19)$$

A linear time method was developed by Dyer (1984) to solve two-variable LPs with this special structure. We adapt the algorithm to incorporate nonnegativity constraints on the decision variables (the algorithm is summarized in the Online Supplement available at http://joc.pubs.informs.org/ecompanion.html).

4.2. Valid Inequalities

The standard ND method is attractive for (D-ASP) given the special structure of the subproblems discussed in §4.1. However, slow convergence of outer linearization methods such as this one has been observed by several authors (see, for example, Magnanti and Wong 1981). This results because little information is available in the form of optimality cuts at early stages of the algorithm, and significant degeneracy in subproblems results from the outer linearization process (Birge 1985). We examine opportunities to overcome this problem using lower bounding inequalities based on the *mean value problem*.

Batun et al. (2011) first used the mean value problem to generate valid inequalities for accelerating convergence of the L-shaped method for two-stage stochastic programs. We propose some variants of these valid inequalities that are suited to our M-SLP formulation. The valid inequalities are derived from the mean value problem using Jensen's inequality, $Q(x_j) \ge Q(x_j, \mu_j)$ (Jensen 1906). Thus, $\theta \ge Q(x_j, \mu_j)$ is a valid inequality that can be added to the master problems at stage j. We begin by providing the following property of the mean value solution to (D-ASP) that is central to the development of our valid inequalities.

Proposition 3. *The optimal solution to the mean value problem for* (D-ASP) *is* $\bar{x}_i = \mu_i$.

Proof. Replacing all random variables, $Z_i, i = 1, 2, \dots, n^U$, in (11) with their mean, μ_i, it is straightforward to show that $x_i = \mu_i$ results in $w_{i,j} = 0$ $\forall i$. Because $w_{i,j} \ge 0$, then clearly $x_i = \mu_i$ minimizes $w_{i,j}$. Furthermore, $x_i = \mu_i$ results in overtime $\ell_{j+1} = \sum_{i=1}^{j+1} \mu_i - d$. Substituting $w_{j,2} = (\mu_1 - x_1)$ in (10) gives the lower bound $w_{j,3} \ge \mu_1 - x_1 + \mu_2 - x_2$. Following the same substitutions for all $w_{j,i}$ and finally for ℓ_{j+1}, we obtain the lower bound $\ell_{j+1} \ge \sum_{i=1}^{j+1} \mu_i - d$. Thus $x_i = \mu_i$ simultaneous achieves lower bounds on $w_{j,i}$ and ℓ_{j+1} and is therefore the optimal solution to the mean value problem. □

We denote the objective function for the mean value problem for stage j at the optimum as $\bar{R}_j(\boldsymbol{\mu})$. From Proposition 3, $\bar{x}_i = \mu_i$ $\forall i$ minimizes the stage j mean value problem. Therefore, in the absence of uncertainty in service durations, it is optimal to allocate μ_i to customer i, independent of whether there is uncertainty in the number of arrivals.

Lemma 1. *The following is a lower bound on* $Q_j(x_j, \omega_j)$ *where* ω_j *denotes the duration scenario for stage* j:

$$\sum_{i=2}^{j} c^w_i (Z_i(\omega_j) - x_i)^+ + c^\ell \left(Z_{j+1}(\omega_j) + \sum_{i=1}^{j} x_i - d \right)^+. \quad (20)$$

Proof. From (1),

$$w_{j,i}(\omega_j) = (w_{j,i-1}(\omega_j) + Z_{i-1}(\omega_j) - x_{i-1})^+, \quad i = 2, \ldots, j+1 \tag{21}$$

$$\geq (Z_{i-1}(\omega_j) - x_{i-1})^+, \quad i = 2, \ldots, j+1. \tag{22}$$

From (3),

$$\ell_j(\omega_j) = \left(w_{j,j}(\omega_j) + Z_j(\omega_j) + \sum_{i=1}^{j} x_i - d\right)^+ \tag{23}$$

$$\geq \left(Z_j(\omega_j) + \sum_{i=1}^{j} x_i - d\right)^+, \tag{24}$$

which completes the proof. □

We now use Proposition 3 and Lemma 1 to develop valid inequalities for (D-ASP).

Proposition 4. *The following is a valid inequality* (Valid-1) *for outer linearization of* (11):

$$\text{(Valid-1)} \qquad \theta_j \geq (1 - q_{j+2})(c^\ell - c^w_{j+1})x_j + k_1(\mu_1, \ldots, \mu_{j-1}),$$

where $k_1(\mu_1, \ldots, \mu_{j-1}) = (1-q_{j+2})c^\ell(\mu_{i+1} + \sum_{i=1}^{j-1}\mu_i - d) + q_{j+2}\bar{R}_{j+1}(\boldsymbol{\mu})$ *is a constant*.

Proof. The following is a lower bound on $R_j(x_j)$:

$$R_j(x_j) \geq (1-q_{j+2})\left(\sum_{i=1}^{j} c^w_{i+1}(\mu_i - x_i)^+ + c^\ell\left(\mu_{j+1} + \sum_{i=1}^{j} x_i - d\right)^+\right) + q_{j+2}\bar{R}_{j+1}(\boldsymbol{\mu}), \tag{25}$$

which follows directly from Lemma 1, Jensen's inequality, and substitution of $\bar{R}_{j+1}(\boldsymbol{\mu})$ into (11). Replacing all decision variables x_i with μ_i in (25), excluding the current-stage decision variable x_j will result in the following lower bound:

$$R_j(x_j) \geq (1 - q_{j+2})\left(c^w_{j+1}(\mu_j - x_j)^+ + c^\ell\left(\mu_{j+1} + \sum_{i=1}^{j-1}\mu_i + x_j - d\right)^+\right) + q_{j+2}\bar{R}_{j+1}(\boldsymbol{\mu}). \tag{26}$$

Relaxing the nonnegativity functions in the first two terms of the right-hand side in (26), we obtain the following:

$$\theta_j \geq (1 - q_{j+2})(c^\ell - c^w_{j+1})x_j + (1 - q_{j+2}) \cdot c^\ell\left(\mu_{i+1} + \sum_{i=1}^{j-1}\mu_i - d\right) + q_{j+2}\bar{R}_{j+1}(\boldsymbol{\mu}), \tag{27}$$

$$\theta_j \geq (1 - q_{j+2})(c^\ell - c^w_{j+1})x_j + k_1(\mu_1, \ldots, \mu_{j-1}). \quad \square$$

Note that $k_1(\mu_1, \ldots, \mu_{j-1})$ in Proposition 4 is a constant, and thus ((Valid-1)) is a linear constraint in two variables, x_j and θ. We note the importance of this in preserving the two-variable master problem structure discussed in §4.1.

Proposition 5. *The following is a set of valid inequalities for outer linearization of* (*D-ASP*):

$$\text{(Valid-2)} \qquad \begin{aligned} \theta_j \geq{}& (1 - q_{j+2})(c^w_{j+1}\hat{w} + c^\ell\hat{\ell}) \\ &+ k_2(\mu_1, \ldots, \mu_{j-1}) \\ \hat{w} \geq{}& \mu_j - x_j, \\ \hat{\ell} \geq{}& \mu_{j+1} + \textstyle\sum_{i=1}^{j} x_i - d, \\ \hat{w} \geq{}& 0, \quad \hat{\ell} \geq 0, \end{aligned}$$

where $k_2(\mu_1, \ldots, \mu_{j-1}) = q_{j+2}\bar{R}_{j+1}(\boldsymbol{\mu})$.

Proof. The proof follows from Proposition 3 and Lemma 1 and the use of two new variables, $\hat{w}$ and $\hat{\ell}$, to linearize the first and second terms in (25), which correspond to the waiting time of the last customer and the overtime at stage j, respectively. □

Note that $k_2(\mu_1, \ldots, \mu_{j-1})$ in Proposition 5 is a constant, and (Valid-2) is a linear set of constraints in three decision variables, $\hat{w}$, $\hat{\ell}$, and x_j. The following is the final valid inequality based on the mean value problem.

Proposition 6. *The following is a set of valid inequalities for outer linearization of* (D-ASP):

$$\begin{aligned} \text{(Valid-3)} \quad \theta_j \geq{}& (1 - q_{j+2})\left(\sum_{i=2}^{j+1} c^w\hat{w}_i + c^\ell\hat{\ell}_j\right) \\ &+ q_{j+2}(1 - q_{j+3})\left(\sum_{i=2}^{j+2} c^w\hat{w}_i + c^\ell\hat{\ell}_{j+1}\right) \\ &+ \cdots + \left(\prod_{k=j+2}^{n^U} q_k\right)\left(\sum_{i=2}^{n^U} c^w\hat{w}_i + c^\ell\hat{\ell}_{n^U-1}\right), \\ \hat{w}_{i+1} \geq{}& \hat{w}_i + \mu_i - x_i \quad \forall\, i \leq j, \\ \hat{w}_{i+1} \geq{}& \hat{w}_i + \mu_i - \hat{x}_i \quad \forall\, i > j, \\ \hat{\ell}_j \geq{}& \mu_{j+1} + \sum_{i=1}^{j} x_i - d, \\ \hat{\ell}_k \geq{}& \mu_{k+1} + \sum_{i=1}^{j} x_i + \sum_{i=j+1}^{n^U-1} \hat{x}_i - d \quad \forall k = j+1, \ldots, n^U - 1, \\ x_i \geq{}& 0 \quad \forall i = 1, \ldots, j,\ \hat{x}_i \geq 0 \ \ \forall i = j+1, \ldots, n^U - 1, \\ \hat{w}_i \geq{}& 0 \quad \forall i = 2, \ldots, j+1,\ \hat{\ell}_k \geq 0 \ \ \forall k = j+1, \ldots, n^U - 1. \end{aligned}$$

Proof. (Valid-3) follows directly from Proposition 1 of Batun et al. (2011), based on adding several new

auxiliary variables, $\hat{x}_i$, $\hat{w}_i$, and $\hat{\ell}_k$, that define the mean value problem. The first constraint follows because the objective of the mean value problem is a lower bound on the optimal solution from Jensen's inequality. We let $\hat{w}_i$ denote the waiting time for customer i in the mean value problem, $\hat{\ell}_k$ the overtime in the mean value problem, and $\hat{x}_i$ the time allowance for customer i in the mean value problem. □

Because (Valid-1), (Valid-2), and (Valid-3) are based on progressively weaker relaxations of the mean value problem, they are increasingly stronger valid inequalities. However, the number of constraints in each set is increasing, causing greater computational effort in solving the master problem at each stage of ND. Furthermore, (Valid-1) includes only two variables and therefore retains the computational advantage of a two-variable master problem at each stage.

4.3. Multicut Outer Linearization

We use a two-cut adaptation of the multicut L-shaped method proposed by Birge and Louveaux (1988). Based on the structure of (D-ASP), we generate one cut for each of the two terms in the objective function, i.e., the terminal subproblem and the expected cost-to-go. Thus the outer linearization problem is of the form

$$\min \quad (1-q_{j+2})\theta_j^1 + (q_{j+2})\theta_j^2 \tag{28}$$

$$\text{s.t.} \quad \theta_j^1 \geq Q_j(x_j), \tag{29}$$

$$\theta_j^2 \geq \min_{x_{j+1}}\{R_{j+1}(x_{j+1})\}, \tag{30}$$

where the right-hand side of the cuts is replaced with supporting hyperplanes. In other words, we separately outer linearize the right-hand side of the two terms in (11) using two variables, θ_j^1 and θ_j^2, at each stage j. Thus, we add two optimality cuts to the master problems simultaneously at each iteration.

4.4. Nested Decomposition

The ND algorithm proceeds by iteratively improving the approximation of each stage's convex objective function by adding supporting hyperplanes. Master problems at each stage approximate the expected value of all future stages. (NS-ASP) and (D-ASP) both have complete recourse; therefore, decisions made in a given stage have feasible completion in future stages. Thus, we do not need to consider feasibility cuts in our implementation. In summary we propose the following opportunities to improve efficiency of the ND algorithm: (a) addition of valid inequalities, (b) a fast method for solving two-variable LPs, and (c) multicut outer linearization. The various implementations of our algorithm are summarized as follows:

Nested Decomposition Algorithm

1. $\nu = 1$, $j = 1$, $k = 1$
2. Start with an arbitrary solution x_j
3. **While** (current bound$_j - \theta_j >$ tolerance) **do**
4. Direction ← Forward
5. $\nu \leftarrow \nu + 1$
6. **for** $j = 1$ to $n^U - 1$
7. **if** valid inequality = True
8. Add valid inequality (Valid-1), (Valid-2), or (Valid-3) to the master problem
9. **end if**
10. Solve master problem j
11. Solve subproblem j for each k
12. **end for**
13. Direction ← Backward
14. $\nu \leftarrow \nu + 1$
15. **for** $j = n^U - 1$ to 1
16. **if** Standard ND = True
17. Add single optimality cut (18) to master problem
18. **else** (Multicut ND = True)
19. Add optimality cuts (29), (30) to the master problem
20. **end if**
21. Solve master problem
22. **end for**
23. **end while**

In our implementation, master problems were solved with either CPLEX 11.0 or with our implementation of the two-variable LP algorithm of Dyer (1984). Note that the two-variable algorithm cannot be used for the multicut ND procedure because it has three decision variables (θ_j^1, θ_j^2, x_j). It can only be used in combination with (Valid-1) because it is the only set of valid inequalities that maintains the two-variable structure of the master problems.

The ND algorithm is implemented using the *fast-forward-fast-back* strategy proposed by Wittrock (1983), which explores all scenarios at stage j before moving forward to stage $j+1$ or backward to stage $j-1$. That is, starting from the first stage, all problems at future stages are solved sequentially as the information gathered from solved problems is passed to the future stages. Upon reaching the last stage, the direction is reversed, and optimality cuts are added to the master problems at each stage. The cycle repeats until no new cuts can be generated. Motivation for the efficiency of this particular strategy is provided by Gassmann (1990).

5. Results

In this section we provide the results of numerical experiments to illustrate the structure of optimal

schedules and to evaluate the proposed methods. All experiments were performed with 10,000 randomly generated service duration scenarios, which we have found sufficient to achieve tight confidence intervals on the optimal solution. The methods proposed in §4 were implemented in C++ with the CPLEX 11.0 callable library (except where noted) to solve the linear subproblems and master problems. The problem instances are solved to optimality within the tolerance of 10^{-6}. All experiments were performed on an Intel Core2 Quad CPU Q6600, with 2.39 GHz and 3.25 GB of RAM.

We present the results of a series of numerical experiments illustrating the solution time for the various methods proposed, as well as relevant insights related to the value of the stochastic solution (VSS), and sensitivity of the optimal solution to model parameters. We begin by providing specific examples that illustrate the structure of the optimal solution and its sensitivity to changes in cost parameters, c^w and c^ℓ, for (NS-ASP) and (D-ASP). We present numerical experiments to evaluate the performance of our algorithms. Next, we compute the VSS for a series of randomly generated model instances. Finally, we present the results of a case study based on a real problem faced at Mayo Clinic in Rochester, Minnesota. We use the case study to illustrate insights about the optimal solution to both (NS-ASP) and (D-ASP).

5.1. Structure of the Optimal Schedule

In our numerical experiments we used uniform and lognormal distributions for service durations. These were motivated in part by applications to primary care (Zeng et al. 2009) and specialty care (Berg et al. 2010). We begin by presenting examples that illustrate the structure of the optimal schedule with respect to changes in relative cost of waiting, c^w, and overtime, c^ℓ, defined by cost ratio $\alpha = c^\ell/c^w$.

Figure 4 illustrates the optimal allowances for a 10-customer problem instance of (NS-ASP) for varying no-show probabilities. The first plot on the top left corner of Figure 4 indicates the allowances between 10 customers when $\alpha = 0.1$ and $p_i = 0$. According to this optimal schedule, for instance, the allowances for the first three customers are as follows: $x_1 = 38.36$, $x_2 = 38.27$, and $x_3 = 38.40$. That is, the first customer is scheduled at time 0, the second customer is scheduled to arrive at 38.36, the third customer is scheduled to arrive at $38.36 + 38.27 = 77.63$, and so on. When no-shows are not allowed ($p_i = 0\ \forall i$), the optimal schedule preserves the *dome shape*, i.e., shorter allowances for patients early and late in the session and larger allowances for the patients in the middle. This pattern has been observed for static scheduling problems (Denton and Gupta 2003). When no-shows are present, with probability $p = 0.2$ and $p = 0.3$, the optimal allowances are reduced to hedge against high idling that is caused by customers who do not show up. When $\alpha = 10$, we observe double-booking for the first two customers; i.e., the first two customers are scheduled to arrive at the same time. As α increases, we observe additional double-bookings for the customers early in the schedule. Double-booking is very common in practice when scheduling patients in the presence of no-shows. Indeed, we find that it is optimal for certain choices of cost coefficients. As the cost ratio α increases, consistent with intuition we observe more double-bookings for the customers early in the schedule; i.e., for $\alpha = 10{,}000$ and $\alpha = 1{,}000$, customers 1, 2, 3, and 4; for $\alpha = 100$, customers 1, 2, and 3; and for $\alpha = 10$, customers 1 and 2 are double-booked.

We used (D-ASP) to evaluate the optimal schedule for an endoscopy suite at the Mayo Clinic in Rochester, MN. Endoscopy procedure durations are reported to have a shifted lognormal distribution $3 +$ Lognormal(23.55, 11.89) by Berg et al. (2010). Based on the analysis of a historical data set for a six-month period during 2006, five routine patients are scheduled for colonoscopy for a given session prior to the day of the procedure. Physicians may request additional appointments for up to three more patients. Based on observational data, the conditional probabilities for the appointment requests for these patients are approximately 0.8, 0.5, and 0.3.

Figure 5 depicts the optimal allowances for each customer for different choices of the cost ratio, $\alpha = c^\ell/c^w$. The optimal allowances for routine patients form a dome shape as observed in static scheduling problems. As α decreases from 10 to 1, the allowances between patient arrivals increase, and for $\alpha = 0.1$, allowances are nearly identical. In other words, as the cost of overtime increases, patient interarrival times decrease; as the cost of waiting time increases, patient interarrival times increase.

Figure 6 illustrates the results of an experiment to observe the effects of the changes in the number of add-on patients on the optimal schedule. The numbers in parentheses in the legend denote routine patients and add-on patients, respectively. For this experiment we used uniformly distributed service times. The appointment request probabilities, q_i, for add-on customers are selected to have decreasing values. Our results show that the optimal schedule is sensitive to the number of routine versus add-on patients. As the expected number of patients in the system decreases from 12 routine and 0 add-on $(12, 0)$ to 3 routine and 9 add-on $(3, 9)$, the optimal time allowances for patients early in the day increase and those for patients later in the day decrease (for most patients). Note that the time allowances in the presence of add-on patients are not monotonic and in some cases do not exhibit the dome shape observed for the static appointment scheduling problems.

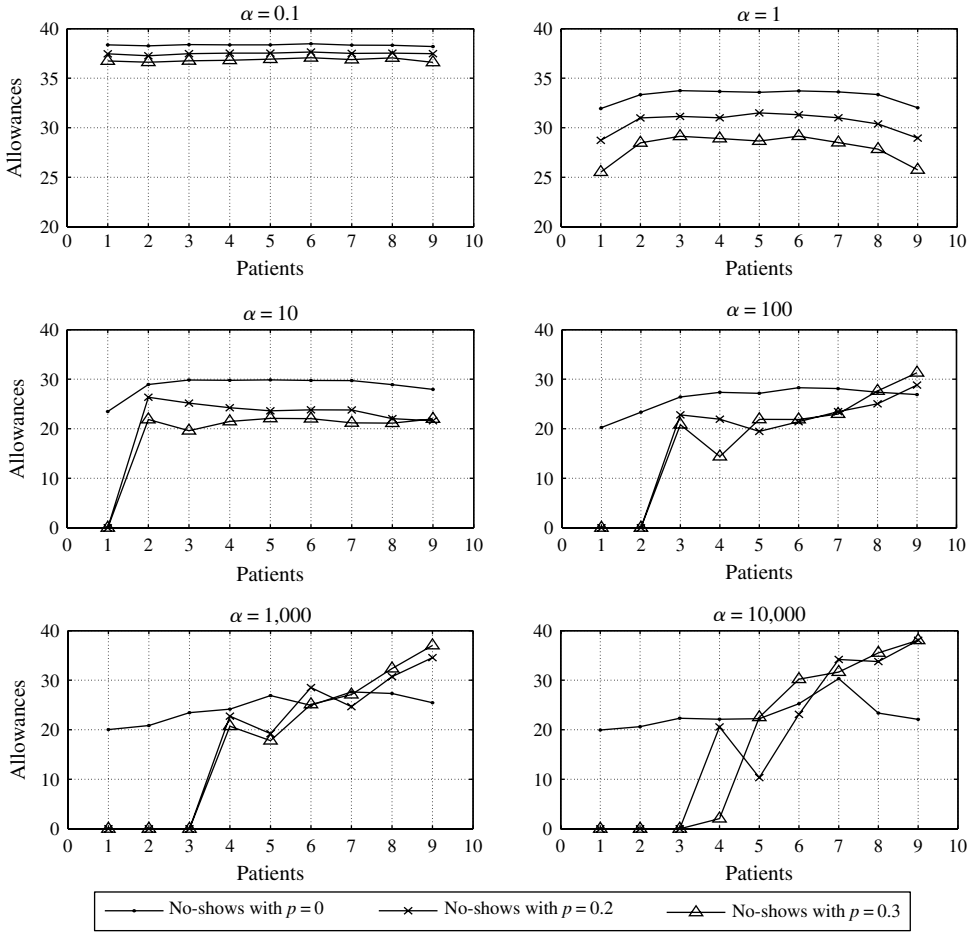

Figure 4 **Effects of Cost Ratio $\alpha = c^{\ell}/c^{w}$ and No-Show Probability p on an Optimal Schedule for the 10-Customer Problem Compared to the Case in Which All Customers Arrive** $(p = 0)$, $(Z_i \sim \mathsf{U}(20, 40), d = 200)$

5.2. Numerical Experiments

5.2.1. Computational Performance of Proposed Methods. We test the algorithms we propose on (D-ASP) because it is the more computationally challenging of the two models. We solve (D-ASP) with variants of the ND algorithm combined with our multicut approach, two-variable algorithm of Dyer (1984), and the valid inequalities described in §4.2. We report the solution times and the number of iterations for three instances ($n^U = 10, 20, 30$). The appointment

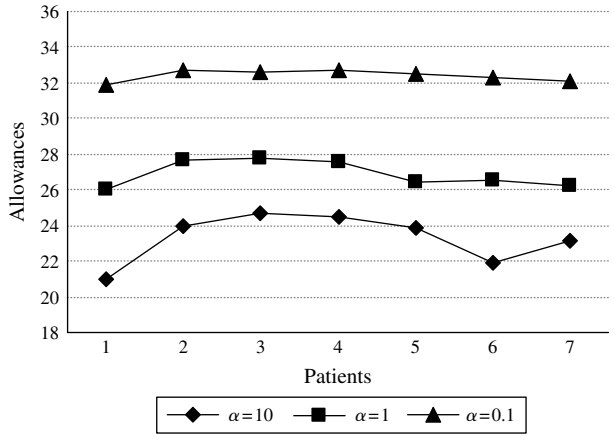

Figure 5 **Structure of the Optimal Solution of Five Routine and Three Add-on Patients with Different Cost Ratio $\alpha = c^{\ell}/c^{w}$,** $Z_i \sim 3 + \mathsf{Lognormal}(23.55, 11.89)$, $d = 150$

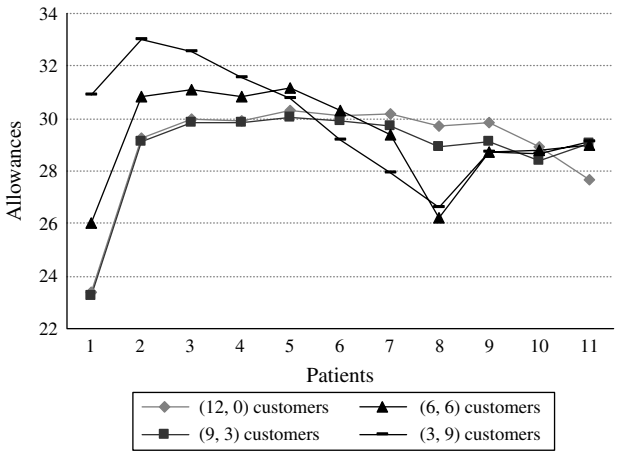

Figure 6 **Effects of the Number of Add-on Customers on the Optimal Schedule,** $Z_i \sim \mathsf{U}[20, 40]$, $d = 250$, $\alpha = 10$

Table 1 Computational Performance of Standard ND, Multicut Version of ND, and the Two-Variable Algorithm Implemented Within ND ($\alpha = 10$)

			Number of iterations			CPU time (in seconds)		
			$n^U = 10$ ($d = 200$)	$n^U = 20$ ($d = 400$)	$n^U = 30$ ($d = 600$)	$n^U = 10$ ($d = 200$)	$n^U = 20$ ($d = 400$)	$n^U = 30$ ($d = 600$)
U(20, 40)	$\alpha = 10$	ND	244	432	438	3.42	23.26	49.68
		Multicut ND	186	244	202	2.63	13.52	23.21
		Two-variable ND	254	406	362	3.56	24.06	43.59
U(20, 40)	$\alpha = 1$	ND	192	330	392	2.75	16.77	42.46
		Multicut ND	106	184	174	1.55	9.81	19.85
		Two-variable ND	186	290	284	2.54	16.32	31.82
U(20, 40)	$\alpha = 0.1$	ND	190	302	422	2.55	14.54	43.48
		Multicut ND	96	176	162	1.33	8.79	17.45
		Two-variable ND	186	290	384	2.37	15.49	42.95
LogN(3.34, 0.325)	$\alpha = 10$	ND	238	436	594	3.48	23.39	69.02
		Multicut ND	182	320	330	2.58	18.05	40.53
		Two-variable ND	254	466	534	3.56	27.37	69.89
LogN(3.34, 0.325)	$\alpha = 1$	ND	180	432	486	2.59	22.29	54.11
		Multicut ND	120	230	254	1.73	12.43	29.68
		Two-variable ND	202	370	454	2.73	21.11	55.98
LogN(3.34, 0.325)	$\alpha = 0.1$	ND	200	392	520	2.67	18.99	54.65
		Multicut ND	112	216	236	1.53	10.99	25.93
		Two-variable ND	188	362	466	2.42	19.54	54.92

Table 2 Computational Performance of Standard ND and Standard ND with (Valid-1), (Valid-2), and (Valid-3) ($Z_i \sim \mathrm{U}(20, 40)$)

	Number of iterations			CPU time (seconds)		
	$n^U = 10$ ($d = 200$)	$n^U = 20$ ($d = 400$)	$n^U = 30$ ($d = 600$)	$n^U = 10$ ($d = 200$)	$n^U = 20$ ($d = 400$)	$n^U = 30$ ($d = 600$)
$\alpha = 10$						
ND	244	432	438	3.42	23.26	49.68
ND with (Valid-1)	224	456	412	3.18	23.69	43.29
ND with (Valid-2)	264	460	556	3.82	24.64	62.70
ND with (Valid-3)	232	370	442	3.65	20.83	51.79
$\alpha = 1$						
ND	210	334	392	3.42	17.09	42.70
ND with (Valid-1)	228	344	398	3.15	17.62	44.72
ND with (Valid-2)	210	410	398	3.00	21.15	67.51
ND with (Valid-3)	188	306	364	2.98	16.89	42.50
$\alpha = 0.1$						
ND	190	302	422	3.56	14.56	43.83
ND with (Valid-1)	180	304	440	2.31	14.68	48.67
ND with (Valid-2)	170	344	638	2.40	16.73	68.37
ND with (Valid-3)	174	284	412	2.62	14.86	45.70

request probabilities, q_i are assumed to be decreasing as the number of customers gets larger, which is a natural attribute of a scheduling system with add-on customers. The daily session length d is chosen to be less than the product of the mean service duration and n^U so that the expected overtime is nonzero (to represent the potentially congested nature of appointment scheduling systems).

From Table 1, we conclude that the multicut version of the ND algorithm performs best in average computation time and in total number of iterations of the ND algorithm for all problem instances. The two-variable algorithm also shows promising performance. Based on our experiments it provides similar results to CPLEX 11.0, indicating that perhaps CPLEX includes an implementation of Dyer's algorithm or a similar algorithm to take advantage of the structure of two-variable LPs. In general, for large problems it takes more computational effort to solve models based on lognormal service times than uniform service times.

Table 2 shows the effects of the valid inequalities (Valid-1), (Valid-2), and (Valid-3) added to the master problems in the ND algorithm for varying

Table 3 Computational Performances of ND and ND with (Valid-1), (Valid-2), and (Valid-3) ($d = 0$, $Z_i \sim \mathrm{U}(20, 40)$)

	Number of iterations			CPU time (seconds)		
	$n^U = 10$ ($d = 200$)	$n^U = 20$ ($d = 400$)	$n^U = 30$ ($d = 600$)	$n^U = 10$ ($d = 200$)	$n^U = 20$ ($d = 400$)	$n^U = 30$ ($d = 600$)
$\alpha = 10$						
ND	192	370	472	2.65	18.76	51.98
ND with (Valid-1)	172	328	408	2.4	16.48	44.23
ND with (Valid-2)	178	332	414	2.61	17.04	45.65
ND with (Valid-3)	172	350	428	2.64	18.65	49.37
$\alpha = 1$						
ND	212	434	492	3.12	23.94	60.32
ND with (Valid-1)	212	408	462	3.09	22.31	55.62
ND with (Valid-2)	186	412	470	2.89	22.79	57.17
ND with (Valid-3)	182	378	476	3.06	21.78	60.23
$\alpha = 0.1$						
ND	202	390	496	2.75	19.25	53.29
ND with (Valid-1)	190	376	436	2.59	18.45	46.28
ND with (Valid-2)	208	348	462	2.86	17.25	49.37
ND with (Valid-3)	184	368	490	2.72	19.15	55.46

cost ratios, α. Adding valid inequalities (Valid-1) and (Valid-2) did not result in significant changes in total computation time. On the contrary, including more cuts in the master problems decreased the efficiency of the solution procedure for many of the test instances. Adding valid inequality set (Valid-3) generally provided modest improvement in computational performance of ND.

Table 3 presents the results for similar numerical experiments with $d = 0$. By Proposition 1, this is consistent with the objective of minimizing a weighted sum of expected customer waiting time and expected server idle time. As can be seen from Table 3, the valid inequalities (Valid-1) and (Valid-2) provide greater benefit than observed in $d > 0$ experiments. (Valid-1) performs better than does (Valid-2) except for two cases ($n^U = 10, \alpha = 1$; and $n^U = 20, \alpha = 0.1$). There may be two reasons for this. First, (Valid-1) and (Valid-2) relax the dependence of overtime on the time allowances in early stages when $d > 0$. Second, (Valid-1) preserves the two-variable structure while at the same time involving fewer additional inequalities than does (Valid-2).

5.2.2. Convergence of the Lower Bound. In this section we provide results of an experiment illustrating the convergence of ND (standard ND and ND with (Valid-3)). Figure 7 depicts how quickly the bound improves with respect to the number of iterations, ν. Although adding (Valid-3) did not improve the solution performance significantly, it improved the bound on the optimal solution considerably at early iterations. With the addition of valid inequalities (Valid-3), ND method reaches approximately 70% of the optimum lower bound within approximately 50 iterations as opposed to 70 iterations with standard ND. We note that this is encouraging for solving problems in which (D-ASP) is a subproblem and efficient computation of lower bounds is important.

5.2.3. Value of the Stochastic Solution. The value of the stochastic solution represents the relative benefit of solving the stochastic problem as opposed to the deterministic problem with mean value of the random parameter. We compute the value of the stochastic solution (VSS) for (D-ASP) for several test instances. Tables 4 and 5 show the VSS values for a 30-customer problem with varying numbers of routine and add-on customers and with varying cost ratios, α. Note that the VSS is typically increasing as α decreases, meaning that solving the stochastic problem is increasingly important because the customer waiting cost increases with respect to overtime cost. This is consistent with the findings of Denton and Gupta (2003) for the static

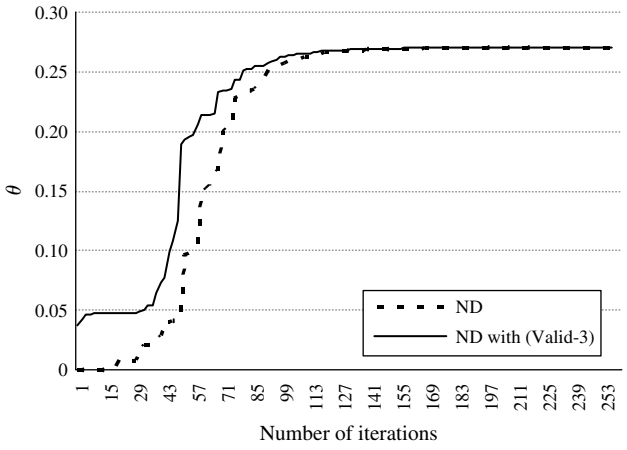

Figure 7 Comparison of Convergence for ND and ND with Added (Valid-3) ($n^U = 40$, $Z_i \sim \mathrm{U}[20, 40]$, $d = 450$, $q_i = 0.5$)

Table 4 **Value of the Stochastic Solution for Several Test Instances with $Z_i \sim \mathrm{U}(20, 40)$ and $q_i = 0.5$ for Add-on Requests**

Number of patients (routine, add-on)	VSS (%)					
	$d = 0$			$d = 200$		
	$\alpha = 10$	$\alpha = 1$	$\alpha = 0.1$	$\alpha = 10$	$\alpha = 1$	$\alpha = 0.1$
(0, 30)	2.33	1.4	39.87	9.63	65.59	95.15
(10, 30)	0.61	10.65	66.95	1.40	19.63	79.41
(20, 30)	0.38	18.45	75.46	0.50	23.63	80.33

scheduling problem. Tables 4 and 5 show that the changes in VSS are not monotonic with respect to the number of add-on requests.

A common approach of scheduling in health-care environments is to use the mean value of the random service durations and schedule the patients within equally spaced intervals. The VSS results also show that solving the stochastic program as opposed to commonly used mean value solution is very beneficial, especially when all of the appointment requests are dynamic and uncertainty in customer load is high (i.e., the case of (0, 30) customers).

5.3. Outpatient Procedure Center

In this section, we evaluate optimal schedules generated by (D-ASP) in a more realistic outpatient procedure center setting. For this purpose, we use a discrete event simulation model of an outpatient endoscopy suite that was previously developed by Berg et al. (2010). The endoscopy suite is typical of a single provider practice. It includes two intake rooms for patient preparation, two parallel endoscopy rooms in which procedures are performed, and four recovery rooms for postprocedure recovery. The service time distributions for intake, procedure, and recovery were assumed to be Lognormal(14.63, 7.24), 3+ Lognormal(23.55, 11.89), and Lognormal(59.18, 18.18), similar to Berg et al. (2010). As in Berg et al. (2010), we assume that patients arrive punctually (the authors note that patients are typically punctual or they arrive early).

The optimal schedules found using (D-ASP) in Table 5 are compared to a schedule based on equally spaced mean service time intervals. The latter is typical of schedules used in practice for endoscopy scheduling. The results based on 10,000 replications are included in Table 6. According to the results, scheduling patients with intervals equal to the mean of the endoscopy procedure distribution is near optimal for the service environments in which service overtime and customer waiting time are considered equally costly ($\alpha = 1$). As α decreases, patient waiting time becomes the dominant performance criteria, and schedules provided by (D-ASP) produce better waiting time results on the average than does the mean value schedule. Similarly, as α increases, server overtime becomes the dominant criteria, and (D-ASP) schedules produces better overtime results on average than does the mean value schedule. According to the total cost values, (D-ASP) produces better schedules with lower total cost than does the mean value schedule in all three environments ($\alpha = 10, 1, 0.1$). The benefits are most pronounced for low values of α, i.e., when the cost of patient waiting is high relative to the cost of overtime.

Table 5 **Value of the Stochastic Solution for Several Test Instances with $Z_i \sim \mathrm{Lognormal}(3.34, 0.325)$ and $q_i = 0.5$ for Add-on Requests**

Number of patients (routine, add-on)	VSS (%)					
	$d = 0$			$d = 200$		
	$\alpha = 10$	$\alpha = 1$	$\alpha = 0.1$	$\alpha = 10$	$\alpha = 1$	$\alpha = 0.1$
(0, 30)	3.99	0.99	36.60	6.42	55.02	89.66
(10, 30)	1.49	7.01	61.41	3.42	12.76	72.17
(20, 30)	0.65	16.07	72.78	0.91	20.26	76.85

Table 6 **Expected Waiting Time and Overtime According to Different Schedules**

	Mean value schedule			Stochastic programming schedule		
	$\alpha = 10$	$\alpha = 1$	$\alpha = 0.1$	$\alpha = 10$	$\alpha = 1$	$\alpha = 0.1$
Expected total cost	975.19	111.72	253.71	878.03	104.58	162.65
Expected waiting time		15.78		16.28	10.54	5.06
Expected overtime		95.94		86.17	94.05	111.97

6. Conclusions and Future Research

In this article we proposed models for scheduling a stochastic server in the presence of uncertainty in demand for appointment requests. Our models aim to find the optimal appointment times given that (a) some patients may fail to show up for their appointment (no-show) and (b) some additional patients may request appointments after an initial schedule has been created. The objective in both of our models is to minimize the total expected cost of patient waiting and overtime.

For (NS-ASP) in which no-shows may occur, our results in some cases indicate a dome shape such as that observed in Denton and Gupta (2003) in the absence of no-shows. The presence of no-shows generally causes the optimal interarrival times (allowances) between customer appointments to decrease. Furthermore, as the probability of no-show increases, it may be optimal to double book customers. Double-booking is common in practice, and our results show that it is optimal in some cases where overtime or idling costs are high or no-show probabilities are high.

For (D-ASP), the more computationally challenging of the two models, we proposed several methods based on ND that take advantage of the underlying structure of the problem, including additional valid inequalities, a fast method for solving two-variable LPs, and an adaptation of multicut outer linearization. We conducted a series of numerical experiments with varying cost ratios, service time distribution types, and appointment request probabilities. Computational improvements were observed, and they are particularly encouraging for future applications that involve solving problems in which (D-ASP) is a subproblem and efficient computation of lower bounds is important (e.g., a branch-and-bound implementation for problems that consider assignment of customers among multiple servers or multiple days). Our results indicate that the multicut implementation of ND for (D-ASP) also gives significant computational advantage.

For (D-ASP), our numerical experiments show that the dome shape observed in the static scheduling case is preserved for routine customer appointments. However, as the relative number of routine customers decreases and of add-on customers increase, the interarrival times increase for the routine patients scheduled to arrive early in the day and decrease for the add-on patients scheduled to arrive later in the day. Finally, we observe very high VSS for some problem instances, indicating that the solution of the stochastic program is important in the dynamic scheduling context.

The models that are presented in this paper are motivated by scheduling problems that are common to many service systems. The models we present are realistic representations of many scheduling environments. Nevertheless, our models have some limitations. First, we assumed that the decision maker is risk neutral, i.e., that his or her goal is to minimize expected cost. However, the true behavior may vary from person to person and from institution to institution. Second, the FCFS assumption, though widely used in many service systems, may not be appropriate in certain environments, especially in the presence of patients with varying priorities. Third, our models do not take into account the possibility of rescheduling, i.e., changing the schedule partway through the day. This is not an unrealistic assumption in outpatient scheduling environments because most service systems avoid rescheduling due to additional costs and negative effects on customer satisfaction. However, the potential benefits of rescheduling remain to be determined.

Acknowledgments

This project was funded in part by CMMI-0620573 and CMMI-550047 from the National Science Foundation. The authors are grateful for the help of Bjorn Berg in evaluating schedules with the endoscopy suite simulator. They are also grateful for the time of anonymous reviewers and the associate editor, whose comments helped to improve this manuscript.

References

Batun, S., B. Denton, A. Schaefer, T. Huschka. 2011. Operating room pooling and parallel surgery processing under uncertainty. *INFORMS J. Comput.* **23**(2) 220–237.

Berg, B., B. Denton, H. Nelson, H. Balasubramanian, A. Rahman, A. Bailey, K. Lindor. 2010. A discrete event simulation model to evaluate operational performance of a colonoscopy suite. *Medical Decision Making* **30**(3) 380–387.

Birge, J. R. 1985. Decomposition and partitioning methods for multistage stochastic linear programs. *Oper. Res.* **33**(5) 989–1007.

Birge, J. R., F. V. Louveaux. 1988. A multicut algorithm for two-stage stochastic linear programs. *Eur. J. Oper. Res.* **34**(3) 384–392.

Birge, J. R., F. Louveaux. 1997. *Introduction to Stochastic Programming*. Springer, New York.

Bosch, P. M. V., D. C. Dietz. 2001. Scheduling and sequencing arrivals to an appointment system. *J. Service Res.* **4**(1) 15–25.

Brahimi, M., D. J. Worthington. 1991. Queuing models for outpatient appointment systems—A case study. *J. Oper. Res. Soc.* **42**(9) 773–746.

Cayirli, T., E. Veral. 2003. Outpatient scheduling in health care: A review of literature. *Production Oper. Management* **12**(4) 519–549.

Denton, B., D. Gupta. 2003. A sequential bounding approach for optimal appointment scheduling. *IIE Trans.* **35**(11) 1003–1016.

Dexter, F., R. H. Epstein, R. D. Traub, Y. Xiao. 2004. Decision making on the day of surgery based on operating room efficiency and patient waiting. *Anesthesiology* **101**(6) 1444–1453.

Deyo, R. A., T. S. Inui. 1980. Dropouts and broken appointments, a literature review and agenda for future research. *Medical Care* **18**(11) 1146–1157.

Dyer, M. E. 1984. Linear time algorithms for two- and three-variable linear programs. *SIAM J. Comput.* **13**(1) 32–45.

Erdogan, S. A., B. T. Denton. 2011. Surgery planning and scheduling. J. J. Cochran, L. A. Cox, P. Keskinocak, J. P. Kharoufeh, J. C. Smith, eds. *Wiley Encyclopedia of Operations Research and Management Science.* John Wiley & Sons, Hoboken, NJ,

Gassmann, H. I. 1990. MSLip: A computer code for the multistage stochastic linear programming problem. *Math. Programming* **47**(1–3) 407–423.

Gerchak, Y., D. Gupta, M. Henig. 1996. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Sci.* **42**(3) 321–334.

Gupta, D., B. Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE Trans.* **40**(9) 800–819.

Hassin, R., S. Mendel. 2008. Scheduling arrivals to queues: A single-server model with no-shows. *Management Sci.* **54**(3) 565–572.

Ho, C., H. Lau. 1992. Minimizing total cost in scheduling outpatient appointments. *Management Sci.* **38**(12) 1750–1764.

Jansson, B. 1966. Choosing a good appointment system—A study of queues of the type $(D, M, 1)$. *Oper. Res.* **14**(2) 292–312.

Jensen, J. L. 1906. Sur les fonctions convexes et les inégalités entre les valeurs moyanes. *Acta Mathematica* **30**(1) 175–193.

Kaandorp, G., G. Koole. 2007. Optimal outpatient appointment scheduling. *Health Care Management Sci.* **10**(3) 217–229.

Klassen, K. J., T. R Rohleder. 1996. Scheduling outpatient appointments in a dynamic environment. *J. Oper. Management* **14**(2) 83–101.

Lee, V. J., A. Earnst, M. I. Chen, B. Krishnan. 2005. Predictors of failed attendances in a multi-specialty outpatient center using electronic databases. *BMC Health Services Res.* **5** doi:10.1186/1472-6963-5-51.

Magnanti, T. L., R. T. Wong. 1981. Accelerating Benders decomposition: Algorithmic enhancements and model selection criteria. *Oper. Res.* **29**(3) 464–484.

Mercer, A. 1960. A queuing problem in which the arrival times of the customers are scheduled. *J. Roy. Statist. Soc. Ser. B* **22**(1) 108–113.

Mercer, A. 1973. Queues with scheduled arrivals: A correction, simplification and extension. *J. Roy. Statist. Soc. Ser. B* **35**(1) 104–116.

Moore, C. G., P. Wilson-Witherspoon, J. C. Probst. 2001. Time and money: Effects of no-shows at a family practice residency clinic. *Family Medicine* **33**(7) 522–527.

Murray, M., C. Tantau. 2000. Same-day appointments: Exploding the access paradigm. *Family Practice Management* **7**(8) 45–50.

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

Robinson, L., R. Chen. 2003. Scheduling doctors' appointments: Optimal and empirically-based heuristic policies. *IIE Trans.* **35**(3) 295–307.

Vissers, J., J. Wijngaard. 1979. The outpatient appointment system: Design of a simulation study. *Eur. J. Oper. Res.* **3**(6) 459–463.

Wang, P. P. 1993. Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Res. Logist.* **40**(3) 345–360.

Weiss, E. N. 1990. Models for determining estimated start times and case orderings in hospital operating rooms. *IIE Trans.* **22**(2) 143–150.

Welch, J. D., N. T. J. Bailey. 1952. Appointment systems in hospital outpatient departments. *Lancet* **259**(6718) 1105–1108.

Wittrock, R. J. 1983. Advances in a nested decomposition algorithm for solving staircase linear programs. Technical Report SOL 83-2, Systems Optimization Laboratory, Stanford University, Stanford, CA.

Zeng, B., A. Turkcan, J. Lin, M. Lawley. 2009. Clinic scheduling models with overbooking for patients with heterogenous no-show probabilities. *Ann. Oper. Res.* **178**(1) 121–144.