Computers & Operations Research 50 (2014) 24–37

Contents lists available at ScienceDirect

Computers & Operations Research

journal homepage: www.elsevier.com/locate/caor

Optimal booking and scheduling in outpatient procedure centers

Bjorn P. Berg [a,*], Brian T. Denton [b,1], S. Ayca Erdogan [c,2], Thomas Rohleder [d,3], Todd Huschka [d,3]

[a] Department of Systems Engineering & Operations Research, George Mason University, 4400 University Drive, MS4A6, Fairfax, VA 22030, United States
[b] Industrial & Operations Engineering, University of Michigan, 2893 IOE Building, 1205 Beal Avenue, Ann Arbor, MI 48109-2117, United States
[c] Daniel J. Epstein Department of Industrial & Systems Engineering, University of Southern California, 3715 McClintock Ave. GER 240, Los Angeles, CA 90089, United States
[d] Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, United States

ARTICLE INFO

Available online 26 April 2014

Keywords:
OR in health services
Stochastic programming
Scheduling
Simulation

ABSTRACT

Patient appointment booking, sequencing, and scheduling decisions are challenging for outpatient procedure centers due to uncertainty in procedure times and patient attendance. We extend a previously developed appointment scheduling model to formulate a model based on a two-stage stochastic mixed integer program for optimizing booking and appointment times in the presence of uncertainty. The objective is to maximize expected profit. Analytical insights are reported for special cases and experimental results show that they provide useful rules of thumb for more general problems. Three solution methods are described which take advantage of the underlying structure of the stochastic program, and a series of experiments are performed to determine the best method. A case study based on an endoscopy suite at a large medical center is used to draw a number of useful managerial insights for procedure center managers.

1. Introduction

Outpatient procedure centers (OPCs) are a growing trend for providing specialty health care procedures (surgical or non-surgical) in the United States. From 1996 to 2006, the rate of visits to OPCs in the United States increased 300% while the rate of similar visits to surgery centers in a hospital setting remained constant [11]. The increase in visit frequency is in part due to the patient benefits for surgery in an OPC setting including lower costs, appointment systems that are often more amenable to patient preferences, the ability to recover at home, lower complication rates, lower infection rates, and shorter procedure durations.

Patient appointment planning decisions are challenging for OPCs due to uncertainty in procedure times and patient attendance. The reasons why patients do not attend their appointments (no-show) vary widely. Contributing reasons for no-shows include a high co-pay, incomplete preparation for the procedure, transportation problems, and forgetfulness. No-shows cause poor resource utilization and unanticipated loss of revenue for the provider. This problem is particularly acute for OPCs because they often have very high fixed costs of staff and physical resources. Due to the nature of preparation process for procedures on the patient's behalf, OPCs have little flexibility in changing their schedule as the day of the planned procedures approaches. Thus, OPCs have little recourse when in terms of adjusting their schedules for no-shows or late cancellations. A review by Macharia et al. [27] reported that no-show rates range from 6 to 92% in outpatient settings. For the endoscopy suite considered in this article, no-show rates ranged between 13 and 24%, depending on the type of procedure.

Recent attention has focused on interventions to prevent no-shows including appointment reminders or educational material to help the patient prepare for their appointment. While interventions may help reduce the no-show rate (at a cost to the provider), high no-show rates are still reported among providers [16]. Therefore, we consider ways to mitigate the effects of no-shows through a combination of booking decisions and optimal scheduling decisions that account for uncertain procedure time and patient attendance. Note that consistent with the appointment scheduling literature we use the term *sequence* for the order in which patients arrive and *schedule* to refer to patient interarrival times given a fixed sequence.

Overbooking has been successful in other service industries including airline scheduling [34], hotel booking [33], and car rental [17]. However, overbooking in OPCs presents some unique challenges. For instance, while an airline can pay an overbooked passenger to take

* Corresponding author. Tel.: +1 703 995 6562.
E-mail addresses: bberg2@gmu.edu (B.P. Berg), btdenton@umich.edu (B.T. Denton), serdogan@stanford.edu (S. Ayca Erdogan), rohleder@mayo.edu (T. Rohleder), huschka.todd@mayo.edu (T. Huschka).
[1] Tel.: +1 734 763 2060.
[2] Tel.: +1 213 821 0705.
[3] Tel.: +1 507 284 2511.

http://dx.doi.org/10.1016/j.cor.2014.04.007
0305-0548/

a later flight, it is not appropriate for a patient that has undergone preparation for a procedure to be delayed to a later day. Furthermore, most perishable asset problems study the use of a discrete physical asset (e.g. airplane seat and hotel room). The problem faced by health care providers is unique in that the asset being reserved (time with a provider or resource) is continuous in nature and there is not a one-to-one allocation of resources to customers. Since patients may have different procedure times and no-show rates, the choice of how many patients to book on a particular day is closely related to the sequencing and scheduling decisions.

In this article, we begin by presenting a generic model for a stochastic server. The objective of our model is to maximize profit which is defined as the difference in per patient reimbursements (revenue) and the weighted sum of costs associated with patient waiting time, provider idle time, and overtime. Decision variables include the number of patients to book on a given day, the sequence of patient arrivals during the day, and interarrival times between patients. The sequencing and scheduling decisions are made for a particular booking decision (number of patients); however, by iteratively evaluating a range of booking decisions, the optimal booking, sequencing, and scheduling decisions are formulated as a single optimization model. Procedure times and patient attendance are random in nature and observed on the day of service, after the appointment decisions are made. Patients are organized into *classes* according to characteristics that influence their no-show probabilities and procedure duration distributions. We consider the perspective of the manager that needs to decide on an appointment sequence and schedule prior to receiving appointment requests. Given the historical demand of classes, a certain number of appointments for each class are reserved. As appointment requests arrive over time patients are allocated the first available appointment corresponding to their class and procedure type. This process is consistent with several practices we have observed, including that which motivates the case study in Section 7.

The appointment sequencing and scheduling aspect of our model is formulated as a two-stage stochastic mixed-integer program. We analyze the structure of the stochastic program and we identify several properties that can be exploited to achieve computational advantages. We consider some special cases for which it is possible to provide sufficient conditions for the optimal sequence and schedule, and we show that these provide useful insights to the general problem. These insights are used to motivate heuristics that are shown to be useful for finding the optimal sequences. Three different solution methods are evaluated which include two decomposition-based approaches and a primal heuristic with which to begin the traditional branch-and-bound algorithm. The decomposition-based solution methods take advantage of the special structure of the underlying stochastic program. The results of our experiments are used to draw managerial insights for OPCs.

This article seeks to answer a number of important research questions. From the perspective of an OPC manager we investigate the following questions: How many patients should be booked for a given day and how is the number influenced by an OPC's costs? What is the optimal arrival pattern to mitigate the risk of no-shows? What is the potential benefit of overbooking in an OPC environment? We evaluate the effectiveness of a single-server approximation, which is commonly studied in the operations research literature, in the context of a real OPC. We present a case study based on the Division of Gastroenterology and Hepatology at Mayo Clinic in Rochester, MN. The division performs minimally invasive, endoscopic procedures for preventive, diagnostic, and therapeutic reasons. We use historical data to generate a set of realistic problem instances. Although our results are presented in the context of an endoscopy practice, the managerial and methodological insights are applicable to many other contexts.

The remainder of this article is organized as follows. In Section 2 we review the relevant literature on appointment scheduling systems and no-shows. In Section 3 we present our model formulation. In Section 4 we present analytical insights about the structure of the optimal sequence and schedule for a special case. In Section 5 we discuss three solution methods that are applicable to our problem, and which take advantage of the underlying structure of our model. In Sections 6 and 7 we present numerical results comparing the proposed solution methods, and the results of the case study, respectively. Finally, in Sections 8 we summarize the most important managerial insights that can be drawn from our study.

2. Literature review

In this section we review some of the relevant literature on appointment scheduling and no-shows. More comprehensive reviews of outpatient appointment scheduling are provided by Cayirli and Veral [9] and Gupta and Denton [19]. Although this article focuses on OPCs, the terms procedure and service, and patient and customer, are used interchangeably in referencing the literature.

Appointment scheduling in the outpatient setting has received considerable attention beginning with Bailey [1], who used a queuing model to compare schedules of customer arrivals at a single server. In addition to queuing models there has been a long history of development of heuristics for appointment scheduling. Soriano [35] was among the first. The author studied a *two-at-a-time* heuristic motivated by increasing provider utilization and mitigating the effects of tardy patients. Although not explicitly motivated by no-shows, this heuristic is an early reference to the use of double booking to attempt to mitigate the impact of uncertainty in the patient arrival process. However, for appointment scheduling settings such as OPCs, the queue does not reach a steady state.

The articles cited above assume a fixed sequence of patient arrivals to the stochastic server. Some studies relaxed this assumption. For example, Weiss [39] provided analytical results for a two-patient sequencing and scheduling problem, establishing sufficient conditions for the optimal schedule to follow a convex ordering of service time distributions. Similarly, Klassen and Rohleder [25] explored the use of service time distribution information and reported that sequencing patients by increasing duration variance works well when patients are dynamically scheduled and there is uncertainty in future demand classification. Similar conclusions are reported by Rohleder and Klassen [32] for realistic clinic settings with patient preferences and scheduler behavior uncertainty. Dexter and Ledolter [14] developed a Bayesian method for calculating prediction bounds for surgical durations that can be used to sequence surgeries. Vanden Bosch and Dietz [37] used a pairwise swap based heuristic to sequence patients and subsequently determine the optimal appointment schedule. Denton et al. [12] formulated a model to include decisions about the sequence of arrivals where total enumeration was applied to small problems (fewer than 5 customers) and compared to simple heuristics for larger problems. The authors concluded that the heuristics worked well for small test cases but there were cases in which the heuristics performed very poorly. Extending this line of research to multiple operating rooms, Mancilla and Storer [28] developed a decomposition-based algorithm for allocating and sequencing a single surgeon's procedures in two parallel operating rooms. Considering the problem of developing surgery schedules that absorb the added uncertainty of emergency surgeries, Bruni et al. [8] develop a stochastic program to find optimal scheduling strategies. A rolling horizon heuristic is presented to evaluate rescheduling and overtime use strategies.

Recently, several authors have considered models that consider the challenge of patient no-shows. Hassin and Mendel [20]

explored the impact of no-shows in the context of a single server queuing model. The authors minimized the sum of expected customer waiting and server availability costs using sequential quadratic programming. Kim and Giachetti [24] developed a stochastic booking model to determine the optimal number of patients to book based on conditional probabilities of no-shows and walk-ins during the day. LaGanga and Lawrence [26] developed a utility function that considers the need to balance patient waiting and overtime with the goal of serving additional patients. They used discrete event simulation and regression analysis in their experiments and concluded that overbooking is appropriate in certain cases such as when there is a high volume of patients, high no-show rates, and low service variability. Erdogan and Denton [15] use a multi-stage stochastic program to formulate the problem of dynamically assigning appointment times to patients when the future demand, service duration, and attendance are uncertain. Dynamic booking decisions were also considered by Muthuraman and Lawley [29] where patient waiting, overtime, and revenue are the objectives in their queuing model, which assumes exponential service times. Zeng et al. [40] extended the work of Muthuraman and Lawley [29] to consider a heuristic for overbooking patients with heterogeneous no-show probabilities. Cayirli et al. [10] used simulation and nonlinear regression to develop an appointment system that focuses on the *dome shape* appointment schedule and can be parametrized for individual practices based on service duration characteristics and attendance rates. Begen and Queyranne [3] took advantage of the special structure of the appointment scheduling problem with discrete service times, no-shows, and walk-ins to find optimal appointment schedules in polynomial time; the authors specifically identify the challenge of simultaneously making sequencing and scheduling decisions as important future research.

This article contributes to the literature in the following ways. First, our model combines several aspects of the literature referenced above including simultaneously determining the number of patients to book, and the patient sequence and schedule on the day of service, in the presence of procedure time and attendance uncertainty. Unlike queuing models, our model requires no special assumptions regarding the distributions of procedure times, no-show probabilities, or other model parameters. Second, instead of studying only simple heuristics using simulation models we study optimization methods to find exact solutions, or tight optimality gaps when limited by computation time. While the model formulation presented in this article has similarities to that in [12], significant model enhancements are included in order to solve larger instances with exact methods. Third, we present new theoretical results for small problems that give insights into the special structure of optimal solutions including simple sequencing rules and the use of double booking to mitigate the risk of no-shows. We show that theoretical results are useful in providing heuristics that lead to optimal solutions, based on results from the exact solution methods. Substantial focus has been given to heuristically determining patient sequences, the performance of many heuristics when compared to provably optimal sequences is undetermined. Finally, we apply our model to a OPC to evaluate the effectiveness of using a single server approximation for a more complicated multi-server system.

3. Model formulation and structure

We start by formulating the booking decision problem. We assume that the goal is to maximize the difference between the expected revenue generated from booking n patients, $R(n)$, and the expected variable cost associated with booking n patients, $C(n)$. It is important to note that there is an implicit cost of no-shows reflected in the lost revenue. The booking decision problem can be defined as follows:

$$\max_n \{R(n) - C(n)\}, \tag{1}$$

where $R(n) = (\text{marginal revenue}) \times \sum_{i=1}^{n}(1-p_i)$ and $C(n)$ is the expected weighted sum of patient waiting time, provider idling time, and overtime.

Since the revenue, $R(n)$, is straightforward to compute, the remainder of this section focuses on the cost, $C(n)$, which is determined by the optimal appointment sequence and schedule that minimizes the expected costs. For a fixed n, $C(n)$ can be formulated as a two-stage stochastic mixed-integer program. The first stage decisions include sequencing patients, and determining the interarrival times for each patient in a given sequence. The second stage decisions are patient waiting time, provider idling time, and overtime under each possible scenario. These are determined after the sequencing and scheduling decisions are made, and patient attendance and procedure time durations are observed. In our model we implicitly assume that there is no opportunity to modify the schedule on the day of service, i.e., rescheduling during the day. This is reasonable since such changes are very uncommon for most service systems including OPCs. Although patients may fail to attend their appointments, we assume those patients that do attend are punctual. This is consistent with our observations of several OPCs in which patients are generally observed to be on time or early.

Our stochastic programming model is defined using the following notation where bold face is used to denote vectors throughout:

Indices

i, i': indices for patients
j: index for appointment sequence slot assignments
ω: index for scenarios

Fixed model parameters

n: number of patients
$c^w_{ii'}$: sequence dependent waiting cost for patient i' following patient i
$c^s_{ii'}$: sequence dependent idling cost for idling time between patients i' and i
c^l: overtime cost
d: planned length of clinic day
M_1 and M_2 : upper bounds for patient waiting and provider idling, respectively

Scenario dependent model parameters

$z_i(\omega)$: procedure duration for patient i in scenario ω
$A_i(\omega)$: attendance indicator for patient i in scenario ω ($A_i(\omega) = 1$ if the patient attends,
$A_i(\omega) = 0$ otherwise)
$\xi(\omega)$: random vector containing scenario dependent parameters, $\xi(\omega) = (z_1(\omega), \ldots, z_n(\omega)$,
$A_1(\omega), \ldots, A_n(\omega))$ where n is the number of patients, $\boldsymbol{A} \in \mathbb{B}^n$, and $\boldsymbol{z} \in \mathbb{R}^n_+$

First stage decision variables

$o_{ii'}$:binary precedence variable defining whether patient i is followed by patient i' ($o_{ii'} = 1$)
or not ($o_{ii'} = 0$)
q_{ij}: binary assignment variable defining whether patient i is assigned to appointment slot j
($q_{ij} = 1$) or not ($q_{ij} = 0$)
x_i: time allotted to patient i's procedure

Second stage decision variables

$w_{ii'}(\omega)$: sequence dependent waiting time for patient i' when preceded by patient i in scenario ω

$s_{ii'}(\omega)$: sequence dependent idle time between patients i and i' in scenario ω

$l(\omega)$: total overtime for scenario ω with respect to the planned length of clinic day d

$g(\omega)$: total earliness of the completion of the last procedure for scenario ω with respect to the planned length of clinic day d

Note that for simplicity we suppress dependence of the second stage decision variables on ω in the mathematical formulation that follows, but the dependence of the second stage random parameters on ω is maintained.

The length of day is the sum of procedure times for all patients plus the idle times between each procedure. However, since the day ends when the last procedure ends, or is a no-show, we exclude the idle time following the last patient. To do this, a *dummy* patient is introduced, who is always the final patient. The dummy patient has zero procedure time and defines the completion time of the final real patient as the completion time of the clinic day. The dummy patient and the associated final appointment slot are denoted in the indices by $n+1$. Using the notation defined above, the mathematical formulation can be written as follows.

First stage problem:

$$C(n) = \min \quad \mathcal{Q}(\boldsymbol{o}, \boldsymbol{q}, \boldsymbol{x}) \tag{2a}$$

$$\text{s.t.} \quad \sum_{i'=1}^{n+1} o_{ii'} \le 1 \quad \forall i \tag{2b}$$

$$\sum_{i=1}^{n+1}\sum_{i'=1}^{n+1} o_{ii'} = n \tag{2c}$$

$$q_{ij} + q_{i'j+1} - 1 \le o_{ii'} \quad \forall (i, i', j \le n) \tag{2d}$$

$$\sum_{i=1}^{n+1} q_{ij} = 1 \quad \forall j \tag{2e}$$

$$\sum_{j=1}^{n+1} q_{ij} = 1 \quad \forall i \tag{2f}$$

$$\sum_{i=1}^{n+1} o_{i,n+1} = 1 \tag{2g}$$

$$\sum_{i=1}^{n+1} o_{n+1,i} = 0 \tag{2h}$$

$$q_{n+1,n+1} = 1 \tag{2i}$$

$$o_{ii'}, q_{ij} \in \{0, 1\} \quad \forall (i, i', j) \tag{2j}$$

$$x_i \ge 0 \quad \forall i, \tag{2k}$$

where

$$\mathcal{Q}(\boldsymbol{o}, \boldsymbol{q}, \boldsymbol{x}) = E_{\xi}[Q(\boldsymbol{o}, \boldsymbol{q}, \boldsymbol{x}, \boldsymbol{\xi})]. \tag{3}$$

Second stage recourse problem:

$$Q(\boldsymbol{o}, \boldsymbol{q}, \boldsymbol{x}, \boldsymbol{\xi}) = \min \quad \sum_{i=1}^{n+1}\sum_{i'=1}^{n} c_{ii'}^{w} A_{i'}(\omega) w_{ii'} + \sum_{i=1}^{n+1}\sum_{i'=1}^{n+1} c_{ii'}^{s} s_{ii'} + c^{l} l \tag{4a}$$

$$\text{s.t.} \quad w_{ii'} \le M_1 o_{ii'} \quad \forall (i, i', \omega) \tag{4b}$$

$$s_{ii'} \le M_2 o_{ii'} \quad \forall (i, i', \omega) \tag{4c}$$

$$-\sum_{i'=1}^{n+1} w_{i'i} + \sum_{i'=1}^{n+1} w_{ii'} - \sum_{i'=1}^{n+1} s_{ii'} = A_i(\omega) z_i(\omega) - x_i \quad \forall (i : i \ne n+1, \omega) \tag{4d}$$

$$\sum_{i=1}^{n+1}\sum_{i'=1}^{n} s_{ii'} - l + g = d - \sum_{i=1}^{n+1} A_i(\omega) z_i(\omega) \quad \forall (\omega) \tag{4e}$$

$$w_{ii'}, s_{ii'}, l, g \ge 0 \quad \forall (i, i', \omega). \tag{4f}$$

The formulation in (2a)–(2k) minimizes the expected costs of patient waiting, server idling, and overtime costs over all scenarios. Note that there are no direct costs associated with the first stage decisions. Constraint (2b) ensures that each patient precedes at most one other patient. Constraint (2c) ensures that every patient, except for the dummy patient and the first patient, is included in exactly two precedence relationships. Constraint (2d) states that a precedence relationship can only exist if that same relationship is defined by the appointment slot assignment decisions. Constraints (2e) and (2f) require that one patient is assigned to every appointment sequence slot and every patient is assigned to one appointment sequence slot. Constraints (2g)–(2i) ensure that the dummy patient will be the last patient as defined by the binary precedence variables and the appointment slot assignment variables. Binary and non-negativity restrictions on the first stage decision variables are defined by (2j) and (2k), respectively.

If patient i does not precede patient i', the associated sequence dependent waiting and idling times will be 0 by constraints (4b) and (4c) as enforced by M_1 and M_2, respectively. Constraint (4d) calculates the waiting and idling times associated with each patient based on the waiting time for the preceding patient. Note that for a given patient, either the associated waiting time or idling time can be positive, but not both. The clinic's overtime and earliness are defined by (4e) with respect to the planned length of the clinic day, d. Non-negativity restrictions on the second stage decision variables are defined by (4f).

The attendance indicator, $A_i(\omega)$, in (4a), (4d), and (4e) is assigned according to the probability of no-show for that patient, p_i, and can be written as

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

Thus, if a patient does not show up for an appointment the attendance indicator is 0, and the procedure duration is 0 in (4a), (4d), and (4e). Note that the model allows for the possibility that individual no-show probabilities differ among patients.

The formulation presented in Denton et al. [12] is extended in this formulation first through the inclusion of heterogeneous no-show probabilities, and second through the use of both precedence and assignment variables in order to strengthen the formulation. The use of assignment decision variables has the advantage of not requiring sub-tour elimination constraints that would otherwise be necessary with the binary precedence variables alone. Waiting and idling time decisions are sequence dependent since different patients may have different no-show probabilities. Thus, binary precedence variables are included in the formulation.

4. Analytical insights

In this section we present properties of special cases that provide insights into the general problem. The results of this section are important for two reasons. First the propositions discussed provide some intuition behind the properties of optimal solutions that we observe for larger practical problems. Second, they are used to motivate easy-to-implement heuristics that we evaluate in Section 6.

We consider the special case where $n = 2, c^{l} = 0$, and $z_i(\omega)$ are independent and identically distributed for $i = 1,2$. Note that this is equivalent to minimizing the expected cost of waiting and server

idling, which is a common formulation in the appointment scheduling literature. In comparing sequences, we assume that the time allotted to the first patient in each of the sequences is represented by x. In the case of two patients, the end of the day is defined as when the last patient finishes, regardless of how much time is allotted for it. Thus, the only time allotment decision that influences the waiting of the second patient or the idle time between the patients is the allocation of the first patient. The expected value of the cost function for the sequence $\{1,2\}$, denoted by Z^{12}, is

$$Z^{12} = E_\omega[(1-p_1)(1-p_2)(c^w(z_1(\omega)-x)^+) + (1-p_1)(c^s(x-z_1(\omega))^+) + (p_1)(c^s(x))]. \quad (5)$$

Similarly, the expected value of the cost function for the sequence $\{2,1\}$ is

$$Z^{21} = E_\omega[(1-p_2)(1-p_1)(c^w(z_2(\omega)-x)^+) + (1-p_2)(c^s(x-z_2(\omega))^+) + (p_2)(c^s(x))]. \quad (6)$$

With this definition of a two patient problem we state the following propositions that define optimal sequencing decisions based on no-show probabilities and procedure duration distribution conditions. Proofs for each proposition can be found in the appendix.

Proposition 4.1. *If $p_1 < p_2$ and $z_1(\omega) \leq_{cx} z_2(\omega)$, then the sequence $\{1,2\}$ is optimal $\leq_{cx}$ denotes a convex ordering.*

Proof. Let x_1^* and x_2^* be the optimal solutions for the sequences $\{1,2\}$ and $\{2,1\}$, respectively. Then,

$$Z^{12}(x_1^*) \leq Z^{12}(x_2^*) = (E_\omega[c^s((x_2^*-z_1(\omega))^+ + p_1(x_2^*-(x_2^*-z_1(\omega))^+))]) \leq (E_\omega[c^s((x_2^*-z_2(\omega))^+ + p_2(x_2^*-(x_2^*-z_2(\omega))^+))]) = Z^{21}(x_2^*).$$

The second inequality follows from the convex ordering, the convexity of overtime, and the assumption that $p_1 < p_2$. □

Intuitively, this can be explained as sequencing the patient with greater uncertainty at the end of the day where they are less likely to disrupt the rest of the schedule. Next, we present the optimal allotment of time for the first patient, x_1^*, for the sequence $\{1,2\}$ where $p_1 < p_2$ and $z_i(\omega)$ are i.i.d for $i=1,2$, and x_1^*. x_1^* is defined as the following based on (5)

$$x_1^* = \arg\min_{x_1}\{E_\omega[(1-p_1)(1-p_2)(c^w(z_1(\omega)-x_1)^+) + (1-p_1)(c^s(x_1-z_1(\omega))^+) + (p_1)(c^s(x_1))]\}.$$

Taking the derivative of (5) with respect to x_1 and setting it equal to 0 yields

$$-(1-p_1)(1-p_2)c^w P(z_1(\omega) > x_1) + (1-p_1)c^s P(z_1(\omega) < x_1) + (p_1)c^s = 0,$$

which is a convex function of x_1. From which we can solve for $P(z_1(\omega) < x_1)$,

$$P(z_1(\omega) < x_1) = \frac{c^w(1-p_1)(1-p_2) - c^s(p_1)}{c^w(1-p_1)(1-p_2) + c^s(1-p_1)}.$$

Thus, the expected cost in (5) is minimized when the allotted time for the first patient in the sequence $\{1,2\}$ is

$$x_1^* = \inf\left\{x_1 \geq 0 : F(x_1) \geq \frac{c^w(1-p_1)(1-p_2) - c^s(p_1)}{c^w(1-p_1)(1-p_2) + c^s(1-p_1)}\right\}. \quad (7)$$

The case for sequence $\{2,1\}$ for x_2^* can be derived similarly from (6). The next proposition relates the above optimal decisions in (7) to double booking, which we define as scheduling the simultaneous arrival of two patients.

Proposition 4.2. *If $z_1(\omega)$ and $z_2(\omega)$ are i.i.d., and $p_1 < p_2$, then it is optimal to double book if $(1-p_1)(1-p_2)/p_1 \leq c^s/c^w$.*

Proof. Double booking implies that $x_1^* = 0$ for $\{1,2\}$. Since $P(z_1(\omega) < 0) = 0$ and $F(x_1 = 0) = 0$ we have the following from (7)

$$F(0) = 0 \geq \frac{c^w(1-p_1)(1-p_2) - c^s(p_1)}{c^w(1-p_1)(1-p_2) + c^s(1-p_1)}.$$

Since the denominator is always positive, we know

$$0 \geq c^w(1-p_1)(1-p_2) - c^s(p_1)$$

and it must hold that

$$\frac{(1-p_1)(1-p_2)}{p_1} \leq \frac{c^s}{c^w}.$$

Thus, if $(1-p_1)(1-p_2)/p_1 \leq c^s/c^w$, $z_i(\omega)$ are i.i.d. and $p_1 < p_2$, then it is optimal to double book. □

Proposition 4.2 provides a sufficient condition based on waiting and idling cost parameters, c^w and c^s, and no-show probabilities, p_1 and p_2, for double booking patients to be optimal. Intuitively, as p_1 grows large and/or the ratio of server idling to waiting cost becomes large, double booking becomes optimal. This provides theoretical support for double booking, which is commonly done in practice.

5. Solution methods

The most computationally challenging part of our booking decision model in (1) is the two-stage stochastic mixed-integer program in (2a)–(2k). In this section we briefly summarize three alternative methods that are suited to the underlying structure of the problem. The first two methods are decomposition methods based on the classic L-shaped method suggested by [36]. The third uses another well known stochastic programming method, *progressive hedging*, suggested by [31], as a primal heuristic, to accelerate branch and bound to solve the extensive form of (2a)–(2k).

5.1. Model structural properties

For the solution methods we exploited properties of the stochastic mixed-integer program in (2a)–(2k). First, valid inequalities derived from the mean value problem have been used as a lower bound in the L-shaped method to accelerate convergence of two-stage stochastic mixed-integer programs such as ours by Batun et al. [2]. From the multivariate version of Jensen's inequality [5] it follows that $\mathcal{Q}(o,q,x) \geq Q(o,q,x,\bar{\xi}(\omega))$ where $\bar{\xi}(\omega)$ represents the mean value scenario. Jensen's inequality applies to functions that are convex in the random variables. Note that the waiting time for patient i is $w_{i'i} = (A_{i'}z_{i'} - x_{i'})^+ + w_{i''i'}$, which is then multiplied by A_i in (4.4a). The idling time before patient i is $s_{i'i} = (x_{i'} - A_{i'}z_{i'} + w_{i''i'})$. Thus, waiting and idling are both convex in $\boldsymbol{A}$ and $\boldsymbol{z}$. The overtime can be written as $l = (\sum_i A_i z_i - d + \sum_i \sum_{i'} s_i i')^+$, and is thus convex in $\boldsymbol{A}$ and $\boldsymbol{z}$. Thus, (4.4a) is a sum of convex functions, and is itself convex in $\boldsymbol{A}$ and $\boldsymbol{z}$. We use the mean value problem inequalities as a lower bound to accelerate our proposed solution methods by adding the following constraints:

$$\theta \geq \sum_{i=1}^{n+1}\sum_{i'=1}^{n} c_{ii'}^w \overline{A}_{i'}\overline{w}_{ii'} + \sum_{i=1}^{n+1}\sum_{i'=1}^{n+1} c_{ii'}^s \bar{s}_{ii'} + c^l \bar{l} \quad (8a)$$

$$\overline{w}_{ii'} \leq M_1 o_{ii'} \quad \forall (i,i') \quad (8b)$$

$$\bar{s}_{ii'} \leq M_2 o_{ii'} \quad \forall (i,i') \quad (8c)$$

$$-\sum_{i'=1}^{n+1}\overline{w}_{i'i} + \sum_{i'=1}^{n+1}\overline{w}_{ii'} - \sum_{i'=1}^{n+1}\bar{s}_{ii'} = \overline{A}_i\overline{z}_i - x_i \quad \forall (i : i \neq n+1) \quad (8d)$$

$$\sum_{i=1}^{n+1}\sum_{i'=1}^{n}\overline{s}_{ii'}-\overline{l}+\overline{g}=d-\sum_{i=1}^{n+1}\overline{A}_i\overline{z}_i \tag{8e}$$

$$\overline{w}_{ii'},\overline{s}_{ii'},\overline{l},\overline{g}\geq 0 \quad \forall(i,i'). \tag{8f}$$

In the above inequalities $\overline{A}_i$ and $\overline{z}_i$ are the mean values for the no-show indicators and procedure times respectively, $\overline{w}_{ii'},\overline{s}_{ii'},\overline{l}$, and $\overline{g}$ are auxiliary decision variables for the mean value scenario, and $o_{ii'}$ and x_i are the first stage decision variables previously defined.

The second property relates to the strength of the formulation. For each scenario, ω, an upper bound for $w_{ii'}$ can be defined as the sum of all the procedure durations. That is, the waiting time associated with each patient will be no greater than the total procedure time for all patients. An upper bound for idling time can be defined as the upper bound for x_i, where we define the upper bound for x_i to be the length of the clinic day, d, since all scheduling is planned to be within the given scheduling horizon. Thus, to strengthen our formulation in Eqs. (4b) and (4c) for each scenario, ω, we define M_1 and M_2 as the following:

$$M_1(\omega)=\sum_{i=1}^{n}z_i(\omega) \quad \forall\omega, \tag{9}$$

$$M_2=d. \tag{10}$$

Note that M_1 now has argument ω because the upper bound is dependent on the particular scenario, ω.

The third property relates to symmetry in the patient sequencing decisions. When similar patients can be aggregated into a class based on procedure duration distributions and no-show probabilities, we use the following symmetry breaking constraints to enforce an arbitrary sequence within a class:

$$q_{i,j}-\sum_{k>j}^{n}q_{i+1,k}\leq 0 \quad \forall j=1,\ldots,n \text{ and } (i,i+1\in C), \tag{11}$$

where C defines the set of patients in the class. Antisymmetry constraints such as this have been shown to have a significant impact on computation time [13,30].

5.2. Exact methods

The L-shaped method is a classic decomposition method for two-stage stochastic programs (see [36] for an early reference, and [6] for a general review) that takes advantage of an alternative formulation of (2a)–(2k) using outer linearization. At each iteration, ν, of the L-shaped method, a *master problem* is solved to obtain a feasible solution, $(\boldsymbol{o}^\nu,\boldsymbol{q}^\nu,\boldsymbol{x}^\nu,\theta^\nu)$, where θ^ν lower bounds the recourse function, $\mathcal{Q}(\boldsymbol{o},\boldsymbol{q},\boldsymbol{x})$. The first stage solution is passed to the second stage subproblems which are solved independently to obtain the dual solutions. Optimality cuts, which are lower bounding hyperplanes of the recourse function, $\mathcal{Q}(\boldsymbol{o},\boldsymbol{q},\boldsymbol{x})$, are generated from the second stage dual solutions and added to the master problem. The algorithm proceeds until the optimal solution is found or some predefined tolerance is satisfied.

In our model the subproblems are very easy to solve and *feasibility cuts*, which induce feasibility in the master solution at each iteration, are not required since every sequence and schedule generated by the master problem is feasible for the second stage subproblems, i.e., the problem has *complete recourse*.

In addition to the classic L-shaped method, we implemented a hybrid multicut version of the algorithm. The original multicut L-shaped method [7] generates an optimality cut for every scenario in the second stage. Our initial computational experience revealed that while more information is passed to the first stage problem through additional cuts, the size of the master problem MIP grows quickly, and the master problem becomes computationally expensive to solve due to the large number of additional cuts. Instead, we aggregated cuts by ranking scenarios into groups based on total procedure time. For example, in a two cut implementation, one cut is generated from scenarios with the lowest 50% of total procedure times, and the other is generated from scenarios with the highest 50% of total procedure times.

In our third method, we solve the extensive form of (2a)–(4f) directly by branch and bound, using a primal heuristic to generate an initial feasible solution. Progressive hedging, a scenario-based decomposition method for multi-stage stochastic programs proposed by [31], is well suited as a primal heuristic. It is based on an alternative formulation of the two-stage stochastic program in which the first stage decision variables are indexed according to scenarios. Thus, the first stage decisions depend on the observed outcome of the random variables, and different decisions may be made for each scenario in the first stage in anticipation of the random outcome. In progressive hedging, the *nonanticipativity* constraints are relaxed and scenario subproblems for each ω are solved independently. Lagrangian multipliers are introduced to enforce nonanticipativity across all scenarios.

5.3. Heuristics

In addition to the exact methods discussed above, easy-to-implement heuristics motivated by the properties in Section 4 were evaluated. The heuristics exploit the fact that fixing the sequence of patients allows the binary decision variables to be fixed. The resulting stochastic linear program is very easy to solve. The proposed heuristics operate in the following way:

Algorithm 1. Heuristics for sequencing and scheduling procedures.

input: Sorting Method: $s=1,2,3,4,5,6$
output: Sequence and arrival schedule for procedures

1 *Sort patients* $i=1,\ldots,n$ *using Sorting Method* s
2 *Fix sequence resulting from Step*1 *for first stage binary decision variables* $\boldsymbol{o}$ *and* $\boldsymbol{q}$
3 *Solve resulting fixed sequence stochastic linear program using the L-shaped method*

The Sorting Methods in Step 1 are based on two procedure parameters related to uncertainty: procedure duration standard deviation (σ_i) and no-show rate (p_i) for each procedure i. Six sorting methods were evaluated and are defined in Table 1.

The sorting methods that use the product of a procedure's standard deviation and no-show rate capture both types of uncertainty in a single parameter. While Sorting Methods 4–6 have no intuitive motivation, these *pessimistic* Sorting Methods are included in order to evaluate the impact of the sequencing decisions.

Table 1
Six sorting methods were evaluated using the heuristic algorithm.

Sorting method	Resulting sequence
1	Increasing σ_i
2	Increasing p_i
3	Increasing $\sigma_i \times p_i$
4	Decreasing σ_i
5	Decreasing p_i
6	Decreasing $\sigma_i \times p_i$

6. Results

In this section we present results based on a series of numerical experiments. First, we describe how parameters were estimated from historical data for a particular OPC. Next, we present results to evaluate sensitivity of optimal decisions to model properties, the value of the stochastic solution, and an analysis of the effectiveness of the heuristics in Section 5. Finally, we present results illustrating sensitivity of optimal booking decisions to cost and revenue estimates for a particular OPC.

6.1. Parameter estimation

Data from the Division of Gastroenterology and Hepatology at Mayo Clinic in Rochester, MN, were used for parameter estimation and numerical experimentation. The division is made up of over 70 faculty providers, conducting procedures ranging from routine colorectal cancer screenings to more complex cases where the patient requires procedures for further diagnostic and therapeutic purposes primarily related to colorectal cancer. Because there is a high variety of procedures on a given day, any provider may perform multiple types of procedures, each requiring different amounts of time and having different historical no-show rates. The heterogeneous daily patient demand and operational structure of the practice lend itself well as a case study for our model formulation.

Procedure time analysis was performed using 6 months worth of procedure data from January, 2010 through June, 2010 with over 10,000 observations covering three GI locations throughout Mayo Clinic. There were five basic procedures performed at these sites: Colonoscopy, EGD (Esophagogastroduodenoscopy), ERCP (Endoscopic Retrograde Cholangiopancreatography), EUS (Endoscopic Ultrasound) Colonoscopy, and EUS–EGD. Expert Fit 7.0 was used to fit probability distributions for each procedure type. Visual and numerical results confirmed very good fits across all procedure types with Expert Fit providing scores above 90 on a 0–100 scale. Procedure descriptive statistics are presented in Table 2.

We assumed that the cost for patient waiting is the average hourly wage in the United States, as this is common practice in the health services research literature [22]. We defined d as the expected duration for the n patients. Overtime cost was defined through the ratio c^l/c^w. We consider several choices of c^l/c^w and we include $c^l/c^w = 33$ in our experiments as this represents an estimated ratio provided by administrators. The parameter c^s was set to 0 since there is no direct cost of idling, and idle time for providers is generally filled with administrative, research, and dictation activities. While not presented, we note experiments where $c^s \neq 0$ proved to be computationally easier to solve. Problem instances were generated for $n=5$ and $n=10$ based on the 4–11 procedures being allocated to each procedure room in the OPC studied.

Table 2
Parameter were estimated based on historical data from the Division of Gastroenterology and Hepatology at Mayo Clinic in 2010. The mean, variance, and the distribution fit for each procedure type is presented along with the no-show probabilities for each procedure type.

Procedure type	Mean	Variance	Distribution fit	No-show probability
Colonoscopy	30.96	188.57	Weibull	0.18
EGD	12.05	58.75	Weibull	0.14
ERCP	38.63	598.25	Weibull	0.13
EUS colonoscopy	28.38	210.66	Weibull	0.16
EUS–EGD	29.59	249.99	Log-logistic	0.24

6.2. Numerical results

Implementation of the methods in Section 5 was done with IBM ILOG Optimization Programming Language using CPLEX 12.2. Experiments were run on a Dell Linux server with 2 Quad-Core Intel Xeon E5420 2.5 GHz CPUs and 16GB shared RAM. To evaluate computational performance of the three methods we propose, we used test cases based on a single provider and procedure room where 55% of procedures are colonoscopies and 45% are EGD's. This scenario is based on a typical day for the endoscopy suite we studied.

For the decomposition-based methods, when a tolerance of 1% was achieved, the algorithm was terminated. A maximum of 15,000 CPU seconds was allowed for each instance and the optimality gap is reported in cases where the problem did not solve to optimality within the CPU allowance. For the branch and bound implementation, the progressive hedging algorithm was terminated at 1000 CPU seconds and branch and bound was allowed a maximum of 15,000 CPU seconds. The upper bounds on M_1 and M_2 in (8) and (9), and symmetry breaking constraints in (10) discussed in Section 5.1 were implemented in these results.

The results presented in Table 3 are based on averages from 10 problem instances with 1000 scenarios each. The optimality gap is defined by the difference between the upper and lower bounds as a percentage of the lower bound. We observed that the solution times and optimality gaps are much better, for each of the methods, for the instances where the overtime to waiting time cost ratio is high (these are likely to be the most realistic scenarios for OPCs). Furthermore, while each of the three proposed methods are computationally competitive for smaller instances, only the hybrid multicut method was able to solve the larger problems to the specified tolerance.

Table 3
Computational results for the classic L-shaped method (L-shaped), hybrid multicut L-shaped method (Multicut), and the progressive hedging primal heuristic method (PH) are presented for instances varying in size and parameter estimates. The minimum, average, and maximum CPU times and average optimality gap, across 10 randomly generated instances, are presented for each method.

n	c^l/c^w	CPU time (sec)									Average gap (%)		
		L-shaped			Multicut			PH					
		Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	L-shaped	Multicut	PH
5	1	244	315.6	370	114	327.9	429	268	327.1	428	< 1.0	< 1.0	0.0
	10	268	342.7	386	201	331.7	595	285	399.9	630	< 1.0	< 1.0	0.0
	33	228	270.9	324	246	282.7	325	236	326.9	433	< 1.0	< 1.0	0.0
10	1	[a]	[a]	15k	[a]	[a]	15k	[a]	[a]	15k	83.1	7.2	186.2
	10	[a]	[a]	15k	9339	10,842.7	13,564	[a]	[a]	15k	43.8	< 1.0	82.9
	33	[a]	[a]	15k	6756	7726.9	9786	[a]	[a]	15k	76.6	< 1.0	32.8

[a] Some instances reached the limit of 15,000 CPU seconds.

6.2.1. Heuristics

Given the significant computational challenges posed by this problem, the heuristic in Algorithm 1 was evaluated for each sorting method for $n=5$ problem instances where procedure durations were assumed to be log-normally distributed with a mean of 30 min. This mean represents the expected duration of the procedures in Table 2 and the log-normal distribution has been identified as an appropriate distribution for procedure durations in similar OPCs [4]. Four test scenarios were generated based on the relationship between procedure duration standard deviation and no-show rate in order to measure the effects of each parameter on the sequence performance. Procedure duration standard deviation ranged from 6 to 30 min and no-show rates ranged from 5% to 45%. The choice of standard deviation range represents a significant variety in procedure duration coefficients of variation (20–100%). While no-show rates for endoscopy suites have been reported between 13% and 24%, the range was broadened for the heuristic experiments in order to have managerial implications for general outpatient clinic settings where higher no-show rates have been reported. The four test scenarios are detailed in Table 4. We note that the stochastic linear programs in Step 3 of Algorithm 1 were solved to within 1% optimality in less than a minute.

The heuristic was evaluated for each sorting method in Table 1 and each test scenario in Table 4 where the ratio c^l/c^w was set to 1, 10, and 33. The results present the average optimality gap for each scenario and sorting method resulting from the heuristic using 10 random seed instances.

In Fig. 1, one graph is presented for each test scenario. In general, the optimality gaps resulting from the heuristic decrease as overtime estimates increase, illustrating that the sequence becomes less important to minimizing costs as overtime costs are more highly valued relative to patient waiting costs. This can be seen most clearly in Scenario 3 where optimality gaps for Sorting Methods 4–6 decrease from approximately 37% to 2% when overtime estimates increase. However, this trend is less accentuated for Sorting Methods 1–3 where the heuristic provides small optimality gaps. This indicates that the sorting methods motivated by the propositions may be generalizable to larger instances with more diverse patient classes.

Table 4
Four test scenarios were used to evaluate the heuristic. Scenarios differ based on the relationship between procedure duration and attendance uncertainties.

Scenario	Description	Standard deviation for procedure $i=1,\ldots,5$ (min)	No-show rate for procedure $i=1,\ldots,5$ (%)
1	Constant standard deviation and increasing no-show rate	18, 18, 18, 18, 18	5, 15, 25, 35, 45
2	Increasing standard deviation and constant no-show rate	6, 12, 18, 24, 30	25, 25, 25, 25, 25
3	Increasing standard deviation and increasing no-show rate	6, 12, 18, 24, 30	5, 15, 25, 35, 45
4	Increasing standard deviation and decreasing no-show rate	6, 12, 18, 24, 30	45, 35, 25, 15, 5

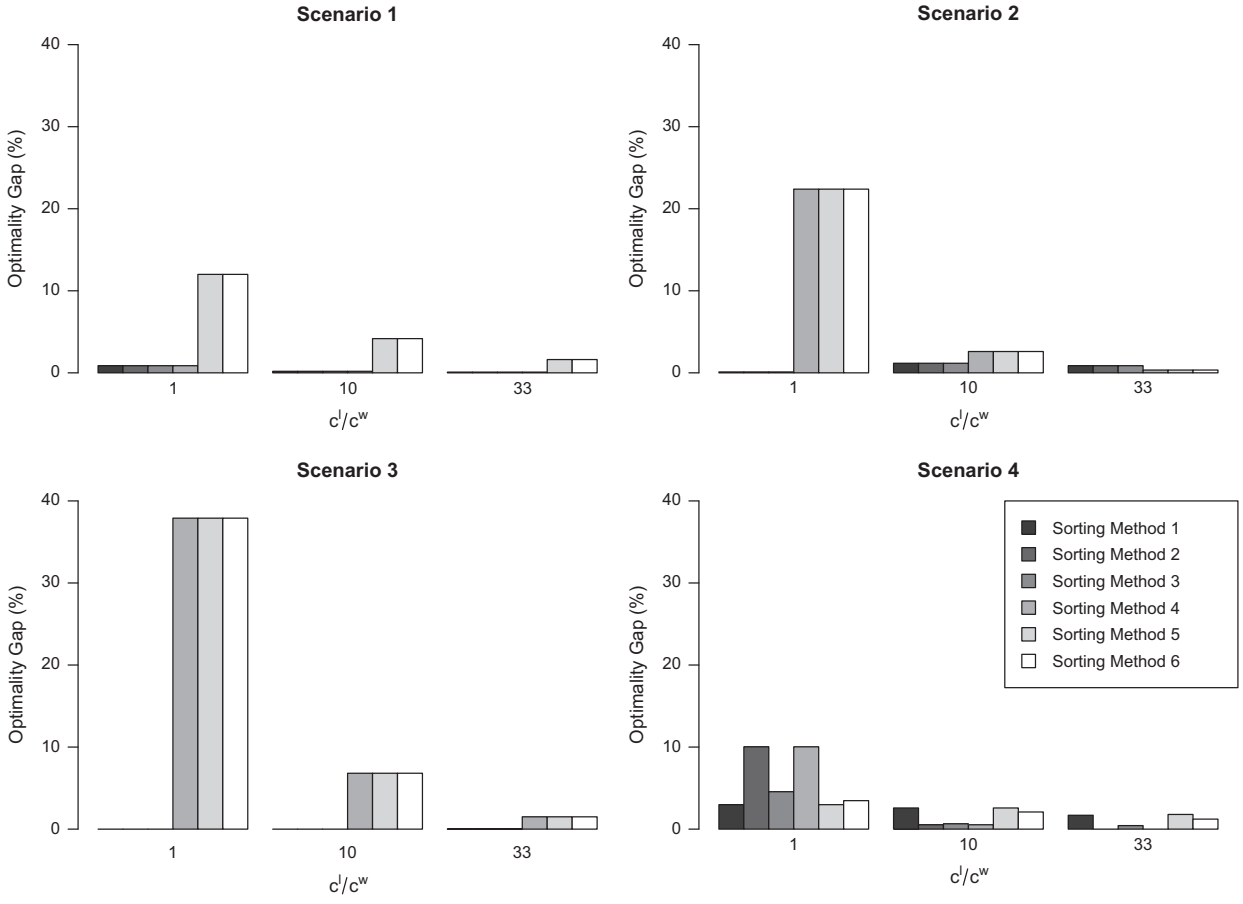

Fig. 1. The optimality gaps for the four scenarios and six sorting methods are compared for three values of c^l/c^w.

The results for Scenario 4 indicate that when procedure duration standard deviation increases and no-show rates decrease, the sorting methods are less differentiated due to the two uncertainties "canceling" each other out in a given sequence. However, when procedure duration standard deviation and no-show rates exhibit a nondirect relationship, sorting by procedure duration standard deviation may take precedence since Sorting Methods 3 and 5 provided the highest optimality gaps. In other words, not including procedure duration standard deviation in the heuristic, and only focusing on no-show rates, resulted in higher optimality gaps.

6.3. Value of the stochastic solution

Table 5 presents the value of the stochastic solution (VSS). VSS is the difference between the expected costs for using the solution of the mean value problem (EEV) and the optimal solution value, $\mathcal{Q}(\boldsymbol{o}^*, \boldsymbol{q}^*, \boldsymbol{x}^*)$. The VSS results were generated using the multicut method since this method tended to have the best solution at the time of termination in the larger instances. We present the VSS in two contexts. First is the VSS for the cost portion of our problem. This represents the traditional interpretation of VSS as it compares the optimal solution value with that of the mean value problem. We also present the VSS within the context of profit. The VSS for the profit represents the improvement in $P(n)$ by using the optimal solution value for $C(n)$ compared to the mean value problem solution.

For the $n=5$ instances in Table 5, the VSS for costs range from 16.80% to 24.80%. The corresponding VSS for profits are lower, ranging from 0.45% to 14.12%, and are most significant for the instances where overtime costs are significantly higher than waiting costs (14.12% improvements). For the $n=10$ instances that did not terminate with an optimal solution, the best solution at termination is used. This represents a lower bound on the VSS. It is interesting to note that for $n=10$, and high ratios of c^l/c^w, the VSS is slightly lower indicating that the value of the stochastic program may be lower for large problems in which the overtime cost is higher than the patient waiting time cost. However, the VSS is significantly higher for $n=10$ and $c^l/c^w=1$. In general, Table 5 shows that the VSS is high for this problem.

6.4. Sensitivity to no-show probabilities

To examine the sensitivity of the optimal schedule to no-show probabilities, we considered a single patient class for which the no-show probability, p_i, is the same for all patients, $i=1,\ldots,n$. The procedure time distribution for colonoscopies in Table 2 was used for all patients. Fig. 2 presents the optimal time allowances for the scenario where $n=10$ colonoscopy procedures (no sequencing decisions), $c^l/c^w=33$, and p is varied from 0.0 to 0.3. When no-show probabilities are $p=0.15$ and $p=0.3$, double booking is observed at the beginning of the day. This is consistent with theoretical results for the $n=2$ case in Section 4 in that it is optimal to double book, and increasingly so as the no-show probability, p, increases. This is also consistent with commonly employed practice, as noted in Section 2. However, the costs associated with the optimal schedule increases if the patients who are double booked at the beginning of the day both show up. For example, in the $p=0.15$ case where it is optimal to double book the first two patients the costs for days where both patients attend are approximately 17% higher than the expected value. In comparison, the costs for days where both patients fail to attend their appointments are approximately 57% less than the expected value. In general, the dome shape is observed early in the schedule, but interarrival times increase throughout the day.

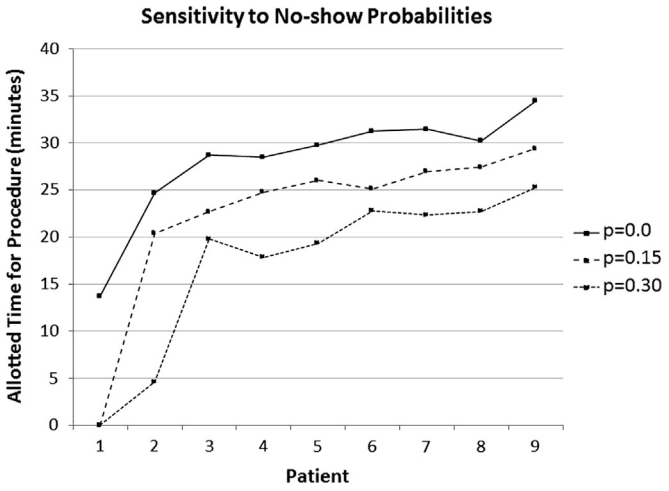

Fig. 2. The interarrival time allowances are presented for a single patient class scenario where no-show probabilities range from 0.0 to 0.3 and the procedure time distribution are for colonoscopies in Table 2. Double booking is observed at the beginning of the day for no-show probabilities of 0.15 and 0.3, respectively. Solution values are averaged over 10 random seed instances.

Two-way sensitivity analysis was also conducted in order to evaluate the structure of the optimal schedule with respect to no-show rates and c^l/c^w ratio estimates. For this analysis, 10 patients with procedure durations based on the colonoscopy procedure distribution were considered. Half of the patients were assigned a no-show rate of 0.15 and the other half was assigned a no-show rate of 0.05, 0.10, 0.20, and 0.25. The c^l/c^w ratio was evaluated for the values of 10, 33, and 100. The results are presented in Fig. 3. As compared to the results in Fig. 2, having multiple patient classes, and thus including a sequencing decision, appears to add volatility to the schedule structure. In general, however, the pattern of scheduling earlier patients closer together is still observed. Further, as the no-show rate and c^l/c^w ratio increase, the optimal schedules tend to allocate less time to each patient. With regard to sequencing decisions for the two patient classes, Fig. 4 illustrates that the patient sequence may be less important as the c^l/c^w ratio increases. That is, when c^l/c^w is low (the value of patient waiting time is high) the optimal schedules tend to sequence the lower no-show rate patients earlier in the day to minimize costs. However, as c^l/c^w increases the portion of low no-show rate patients scheduled earlier in the day tends toward 50%, indicating that the sequence does not matter as overtime costs significantly outweigh patient waiting costs.

6.5. Booking results

The above results have focused on the sequencing and scheduling element of the booking problem, for a fixed number of

Table 5
The value of the stochastic solution (VSS) for instances varying in size and parameter estimates. Results are based on $\mathcal{Q}(\boldsymbol{o}^*, \boldsymbol{q}^*, \boldsymbol{x}^*)$, or the current upper bound at the CPU time limit (noted by *). The value of the stochastic solution is presented in the context of costs (VSS $C(n)$) and profits (VSS $P(n)$).

n	c^l/c^w	Average EEV	Average $\mathcal{Q}(\boldsymbol{o}^*, \boldsymbol{q}^*, \boldsymbol{x}^*)$	VSS $C(n)$ (% improvement from EEV)	VSS $P(n)$ (% improvement from EEV)
5	1	21.91	18.23	16.80	0.45
	10	100.60	83.50	17.00	2.31
	33	305.47	229.72	24.80	14.12
10	1	66.55	45.18*	32.11*	1.32*
	10	198.02	168.75	14.78	1.97
	33	540.33	410.46	24.03	11.35

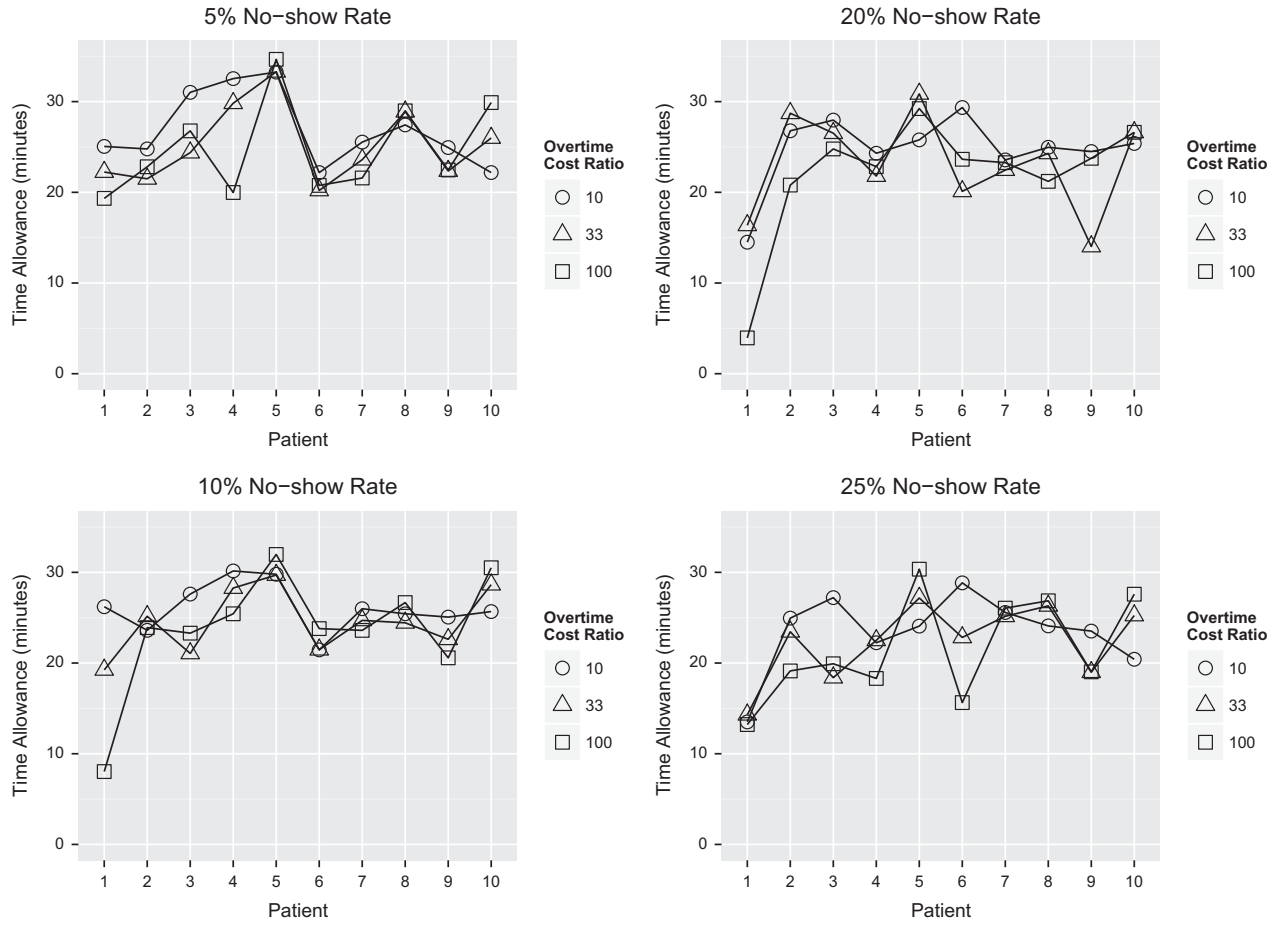

Fig. 3. The interarrival time allowances are presented for a two patient classes where no-show probabilities are 0.15 for one class and range from 0.05 to 0.25 for the other class, with 5 patients in each class. The procedure time distribution is based on colonoscopies in Table 2. Solutions are averaged over 10 random seed instances.

patients. Next, we present results for optimal booking decisions. Recall that the booking decision problem is

$$\max_n \{R(n) - C(n)\}.$$

To find the optimal n we enumerated over the range of feasible values of n. Vargo et al. [38] assumed that 70% of endoscopy suite reimbursements are used to pay the fixed costs of running the suite, leaving 30% ($201.45) of each colonoscopy performed as the remaining revenue based on CMS (Centers for Medicare & Medicaid Services) reimbursement rates. Thus, after accounting for fixed costs, as a baseline we assume $R(n) = \$201.45 \times \sum_{i=1}^{n}(1-p_i)$. We assume $c^s = 0$ and $c^l/c^w = 33$ as previously defined. Procedure duration distributions and no-show probabilities are the same as defined in the computational experiments of the previous subsections. In light of the fact that reimbursements vary across providers, we assume a fixed cost estimate of 70% as a lower bound and compare results for 80% and 90% in Fig. 5. We observe that the optimal number of patients to book decreases as the fixed costs estimate increases. Further, Fig. 5 illustrates a non-smooth trend surrounding the peaks of the curves. This is explained by the fact that each additional patient that is booked is not necessarily of the same class as the previous additional patient. For example, when an additional EGD is added to the booking schedule the resulting increase in expected costs is less than if an additional colonoscopy were added due to the conoloscopy procedures having longer procedure durations with higher variance. For each number of booked patients in Fig. 5, the case mix was held constant with respect to the long-run demand case mix.

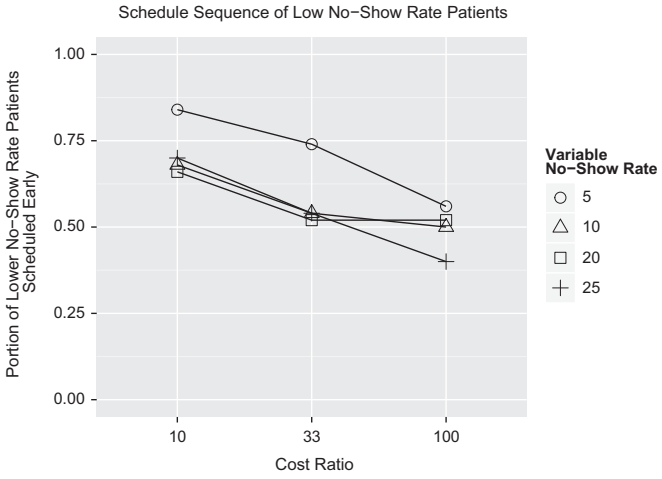

Fig. 4. The sequences are presented for a two patient classes where no-show probabilities are 0.15 for one class and range from 0.05 to 0.25 for the other class, with 5 patients in each class. Being scheduled early is defined by being sequenced in the first half of the schedule.

7. Case study: optimal overbooking for a gastroenterology practice

A detailed discrete event simulation model of the GI Advanced Practice at Mayo Clinic in Rochester, MN, was developed and used to compare booking, sequencing, and scheduling decisions based on our single server booking model and the actual sequences and

schedules employed in practice. The simulation model was developed using Arena 12.0 [23]. The patient flow process in the model includes registration, intake, procedure, and recovery. Resources include registration staff, intake beds, intake nurses, procedure rooms, endoscopists, endoscopes, procedure support staff, recovery beds, and recovery nurses. The suite opens at 7 A.M. and overtime was measured according to procedures ending past 12 P.M. or 4 P.M., depending on the procedure type.

Fig. 6 illustrates the overall process, which is similar to other OPCs described by Berg et al. [4], Gul et al. [18], and Huschka et al. [21]. Intake consists of registering at the front desk, changing into a procedure gown, and meeting with a nurse who collects patient information and explains the procedure. When the patient finishes intake, and a procedure room for their specific procedure is available, the patient is taken to the procedure room where the endoscopist joins them and the procedure begins. Following the procedure the patient is taken to recovery where they will stay until they are ready to be discharged.

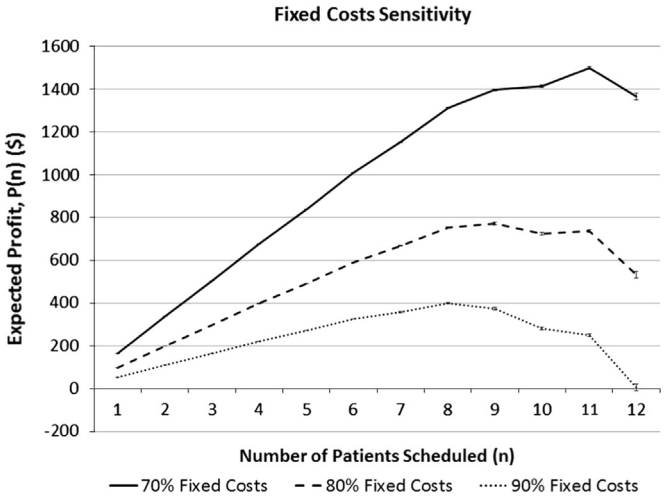

Fig. 5. The sensitivity of the expected profit for booking n patients is compared for varying fixed cost estimates. As the portion of revenue consumed by fixed costs increases, the optimal number of patient to schedule decreases. 95% confidence intervals are included.

As in most procedure centers, procedures require specific resources. For example, an EUS–EGD requires use of ultrasound imaging equipment only available in EUS procedure rooms. Furthermore, each patient is associated with a certain provider. Thus, the process can be approximated as several single servers, where the server is defined by the combination of procedure room and provider. Shared resources exist, such as intake and recovery, but these are generally much less costly and typically not the bottleneck in the system. Therefore, the proposed booking, sequencing, and scheduling method are applied to each procedure room separately. The appointment times generated for each procedure are then used in the simulation model as the times at which the corresponding patients arrive at registration.

The five procedures described in Section 6.1 are each allocated to one of three types of procedure rooms shown in Fig. 6. Colonoscopies and EGD's are performed in the Complex procedure room, EUS Colonoscopies and EUS–EGD's are performed in the EUS procedure rooms, and ERCP's are performed in the ERCP procedure room. Procedure room, endoscopist, and case mix information for the suite are summarized in Table 6.

Test instances were generated based on historical appointment schedules from five different days. Each historical appointment schedule was simulated to compare (a) actual sequences and schedules used in practice, (b) the corresponding two-stage stochastic program solutions assuming the fixed number of patients booked in practice (referred to as the SP solution below), and (c) the optimal booking decision. The actual and SP solutions both assume a fixed number of patients according to the instance. The optimal booking decision refers to the optimal number of

Table 6
Patients are assigned to a procedure room and provider according to the type of procedure they will be receiving. Each procedure type has a specified procedure room and provider. The procedure case mix is based on historical data.

Procedure type	Rooms	Endoscopists	Procedure case mix
ERCP	1	1	100% ERCP
EUS	2	2	88% EUS–EGD 12% EUS Colonoscopy
Complex	1	1	55% Colonoscopy 45% EGD

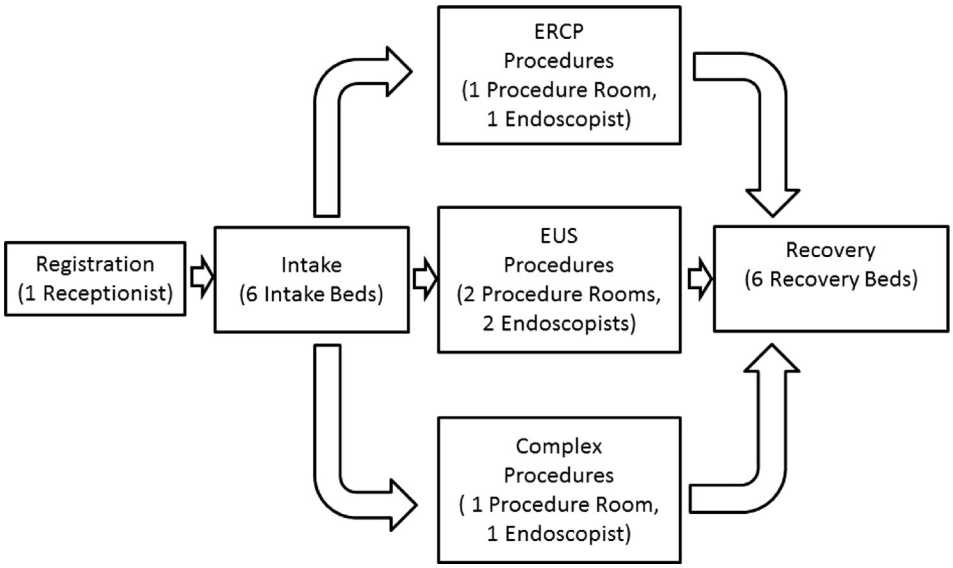

Fig. 6. Each patient goes through the intake, procedure, and recovery processes in the endoscopy suite. Intake and recovery resources are shared between patients receiving different procedures, but the procedure rooms and providers act as independent single servers.

patients scheduled, the optimal sequence, and the optimal time allowances. The optimal number of patients to schedule was defined by enumerating the feasible solutions to (1) and using the n that resulted in the maximum profit. It is important to remember that in this enumeration process, a stochastic program is solved for the cost component of each n. The stochastic program associated with each procedure room was solved to optimality. For larger problems, the CPU budget limit that was used in the computational experiments was eliminated. Each appointment schedule was simulated for 1000 replications using historical data described in Section 6.1 along with historical time stamp data for the registration, intake, and recovery processes. Procedure room turn around times were modeled with a triangular distribution of (10, 15, 20 min) based on a subjective estimate from the endoscopy suite director.

Expected patient waiting time and overtime with 95% half-widths are reported in Table 7. The comparison of the actual and SP results are both with respect to the number of patients actually booked for each instance. From Table 7, the SP solution resulted in higher expected patient waiting time (1–10 min) and lower expected overtime (7–74 min).

Table 7 compares the actual sequence and schedule of patients to the SP solution, assuming no change in the number of patients booked for each instance. The optimal booking decision, which includes the optimal number of patients to book in addition to the corresponding optimal sequence and time allowances, is also included in Fig. 7. The optimal booking decision increases expected profits by a range of 2.49%–63.11% from the actual booking, sequencing, and time allowance schedules. As seen in Fig. 7, the optimal booking decision results in higher expected profit in all of the instances except the SP solution for Historical Schedule 1. In Historical Schedule 1 the profit for the SP solution is actually slightly higher than the optimal booking decision. This is a result of the single server model approximation. Overall, as Fig. 7 illustrates, the optimal booking decision generally performs very well.

The performance of the optimal booking decision in Fig. 7 is based on the parameter estimates used in the model such as overtime cost estimates and the portion of reimbursements that are consumed by fixed costs. Table 8 includes results where model parameter estimates are varied for Historical Schedule 4. Specifically, overtime estimates were varied $\pm 50\%$ resulting in c^l/c^w ratios of 49.5 and 16.5, respectively. Further, the fixed costs estimates were varied between 70% and 90%. Table 8 illustrates that again, in general, the optimal booking decisions provide significant increases in the expected profit. However, while these are extreme estimates of overtime, it can be seen that the historical schedule resulted in higher expected profits when fixed cost estimates were high and overtime costs were low, as well as when fixed cost estimates were low and overtime costs were high. While not presented for brevity, 95% confidence interval half widths ranged between 28.5 and 98.09.

While the objective of our model is to maximize expected profit, an improvement in patient access can also be seen. The optimal booking decision resulted in more patients being booked, $n=36$, than in all of the instances in Table 7, an increase ranging from 24% to 80% (when compared to Historical Schedules 5 and 1 scheduling 29 and 20 patients, respectively). That is, the increase in expected profit also resulted in improved patient access through more patients being booked.

8. Conclusions

In this article we extended the model in [12] to formulate a model for optimal booking, sequencing, and scheduling of a single

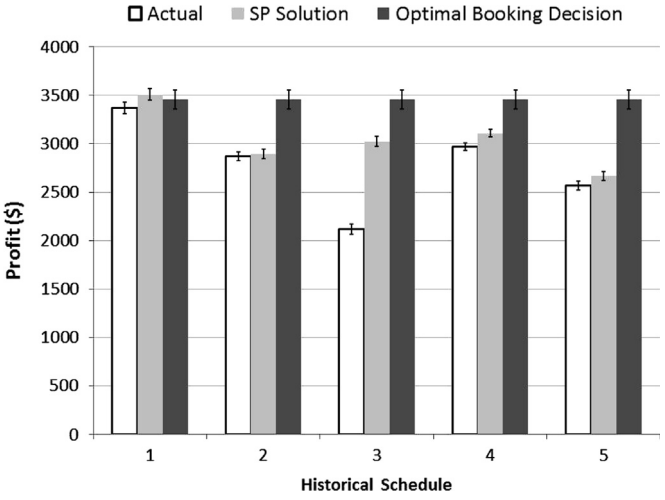

Fig. 7. The expected profits for the actual sequences and schedules and the corresponding SP solutions are presented for a fixed number of patients for each instance. The expected profit for the optimal booking, sequencing, and scheduling decisions is presented for comparison with each historical schedule. 95% confidence intervals are included.

Table 8
Sensitivity analysis was performed on overtime and fixed cost estimates for Historical Schedule 4.

c^l/c^w	Fixed cost %	P (Schedule 4)	P (optimal booking)	% Improvement
16.5	70	3185.58	3411.49	7.09
	80	1699.11	1897.81	11.69
	90	775.10	536.22	−30.82
33	70	2969.29	3455.25	16.37
	80	1583.85	1654.34	4.45
	90	614.68	836.62	36.11
49.5	70	2753.00	2513.97	−8.68
	80	1266.53	1577.54	24.56
	90	342.52	813.76	137.58

Table 7
Actual sequences and schedules and the corresponding two-stage stochastic program solutions (SP) were simulated assuming the fixed number of patients booked in practice. Average patient waiting time and average overtime are compared for five instances with 95% half-widths in parentheses.

Instance	Number of patients			Average patient waiting (min)		Average overtime (min)	
	ERCP	EUS	Complex	Actual	SP solution	Actual	SP solution
1	5	19	5	71.97 (2.13)	73.40 (2.55)	54.25 (4.01)	42.20 (4.00)
2	5	13	4	30.93 (0.77)	40.43 (1.07)	40.51 (3.57)	33.61 (3.57)
3	5	14	4	38.57 (0.97)	40.05 (0.98)	108.87 (3.98)	34.94 (3.41)
4	4	13	6	23.89 (0.87)	41.66 (1.05)	34.97 (3.05)	13.94 (2.33)
5	5	11	4	33.61 (0.73)	41.48 (0.98)	39.08 (3.49)	27.33 (3.22)

stochastic server. We evaluated three alternative solution methods for solving the underlying stochastic mixed-integer program. We provided analytic insights based on special cases including sufficient conditions for the optimal sequence and double booking in the presence of attendance uncertainty. Computational experiments showed that realistic problem instances are very challenging to solve. However, the alternative methods presented provided tight optimality gaps for problems likely to be encountered in practice. Heuristics motivated by theoretical results and the computational challenges observed were evaluated and shown to perform well.

Our analysis revealed that optimal sequencing decisions are quite sensitive to both the procedure duration variance and no-show probability. Our theoretical results for special instances, and heuristic analysis, provide supporting evidence that it is optimal to sequence patients with higher procedure duration variance, and higher no-show probability, later in the sequence during a given day. For example, in the OPC that we studied, our results demonstrate that is optimal to schedule the EGD and EUS–Colonoscopy procedures at the beginning of the day and schedule the Colonoscopy and EUS–EGD procedures later. Our numerical results for larger instances are also consistent with these findings, providing evidence that these findings may generalize to larger problems. Further, our sensitivity analysis illustrated the importance of solving the stochastic program as optimal sequences varied among different problem instances, particularly when the ratio of overtime costs to patient waiting costs is low.

Our heuristic analysis illustrated that there are certain cases when the easy-to-implement heuristics should be used. When overtime costs are significantly greater than patient waiting time costs, the heuristics resulted in solutions with very small optimality gaps. However, the optimality gaps are higher for the heuristic solutions when there is less of a difference between patient waiting time costs and overtime costs, implying that exact solution methods should be used in these cases. Further, the results demonstrated that when there is a direct relationship between procedure duration standard deviation and no-show rate, the heuristics provide near optimal solutions. For example, when procedures that have high duration standard deviation also have high no-show rates, this would be a case to use the heuristics. On the other hand, if there is not a clear relationship between procedure duration standard deviation and no-show rate, then using exact solutions may be preferable.

Our findings indicate that as the probability of no-shows increases, it becomes optimal to double book some patients. Double booking is common in practice and our results show that it is also optimal in some cases where overtime or idling costs are high, or no-show probabilities are high, or both. Although the idea of double booking was alluded to in the literature as early as 1966 by [35] in his *two-at-a-time* policy, we are unaware of theoretical insights, such as ours, about the potential optimality of this practice. In general, we observe the amount of double booking increases with no-show probability. The trend of double booking at the beginning of the day was observed in the structure of the optimal solutions to the scenarios analyzed. While this rule has been heuristically demonstrated to perform well, we have provided numerical results that show it is optimal in certain cases. Intuitively this means double booking is most appropriate when there is a high risk of low utilization resulting from a patient no-show, i.e., when a queue of waiting patients has yet to develop.

Our results show that the optimal number of patients to book is sensitive to the fixed costs associated with a particular practice, and decrease as fixed cost estimates increase. Our case study shows that there may be significant benefits to implementing our model in a realistic multi-server OPC. Based on our experience in working with many different types of OPCs at several institutions, this research presents a model and results that can be generalizable to other outpatient procedure settings. The most significant managerial insights can be summarized as follows:

- While the benefits of overbooking depend on an OPC's cost structure, overbooking resulted in a 17% increase in profit in the most likely cost scenario, with an increase as high as 137% in experiments.
- Sequencing patients with high no-show rates or high procedure duration variance later in the day results in lower costs of waiting, idling, and overtime.
- Double booking is optimal, and increasingly so as no-show probabilities become high; and in the optimal schedules, double booking at the beginning of the day is observed.
- The optimal number of patients to book depends on fixed cost estimates and decreases as fixed costs increase.

There are some limitations to our study that present opportunities for future research. First, we assumed that there is an unlimited supply of patient demand for each procedure type. Thus, in some OPCs, there may be additional constraints on the booking decisions. However, case mix decisions are a result of many factors such as patient population demand and appointment volume at referring departments and clinics. Second, based on our numerical experiments we found that the stochastic mixed integer program is extremely difficult to solve. Thus, an important direction for future research is the study of new solution methods. It is a generic model that underlies many types of industrial service systems. We were able to solve instances up to $n=10$ to optimality, which is suitable for most OPCs, which are the focus of this article. However, other types of service systems that pre-schedule patients, such as lab services, that involve larger numbers of customers, may benefit from additional computational advances. Finally, the model considered here focuses on the single day booking, sequencing, and scheduling decisions. While incorporating indirect waiting for patients is beyond the scope of this immediate work, it is an important component to integrate into future models. Similarly, extending this model and its results to a dynamic appointment scheduling setting may extend the relevance to other service settings where static booking templates are less realistic.

References

[1] Bailey NTJ. A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. J R Stat Soc Ser B Methodol 1952;14(2):185–99.
[2] Batun S, Denton BT, Huschka TR, Schaefer AJ. Operating room pooling and parallel surgery processing under uncertainty. INF J Comput 2011;23(2):220.
[3] Begen MA, Queyranne M. Appointment scheduling with discrete random durations. Math Oper Res 2011;36(2):240–57.
[4] Berg B, Denton BT, Nelson H, Balasubramanian H, Rahman A, Bailey A, et al. A discrete event simulation model to evaluate operational performance of a colonoscopy suite. Med Decis Mak 2010;30(3):380–7.
[5] Berger JO. Statistical decision theory and Bayesian analysis. New York: Springer-Verlag; 1993.
[6] Birge JR, Louveaux F. Introduction to stochastic programming. New York: Springer; 1997.
[7] Birge JR, Louveaux FV. A multicut algorithm for two-stage stochastic linear programs. Eur J Oper Res 1988;34(3):384–92.
[8] Bruni ME, Beraldi P, Conforti D. A stochastic programming approach for operating theatre scheduling under uncertainty. IMA J Manag Math. Advanced Access published January 12, 2014. http://dx.doi.org/10.1093/imaman/dpt027.
[9] Cayirli T, Veral E. Outpatient scheduling in health care: a review of literature. Prod Oper Manag 2003;12(4):519–49.
[10] Cayirli T, Yang KK, Quek SA. A Universal Appointment Rule in the Presence of No-Shows and Walk-Ins. Production and Operations Management 2012; 21(4):682–97.
[11] Cullen KA, Hall MJ, Golosinskiy A. Ambulatory surgery in the United States, 2006. National Health Statistics Reports no. 11; 2009.

[12] Denton BT, Viapiano J, Vogl A. Optimization of surgery sequencing and scheduling decisions under uncertainty. Health Care Manag Sci 2007;10 (1):13–24.
[13] Denton BT, Miller AJ, Balasubramanian HJ, Huschka TR. Optimal allocation of surgery blocks to operating rooms under uncertainty. Oper Res 2010;58(4-Part-1): 802–16.
[14] Dexter F, Ledolter J. Bayesian prediction bounds and comparisons of operating room times even for procedures with few or no historic data. Anesthesiology 2005;103(6):1259.
[15] Erdogan SA, Denton B. Dynamic appointment scheduling of a stochastic server with uncertain demand. INFORMS Journal on Computing 2013;25(1):116–32.
[16] Festinger DS, Lamb RJ, Marlowe DB, Kirby KC. From telephone to office: intake attendance as a function of appointment delay. Addict Behav 2002;27 (1):131–7.
[17] Geraghty MK, Johnson E. Revenue management saves national car rental. Interfaces 1997;27(1):107–27.
[18] Gul S, Denton BT, Fowler JW, Huschka T. Bi-criteria scheduling of surgical services for an outpatient procedure center. Prod Oper Manag 2011;20 (3):406–17.
[19] Gupta D, Denton BT. Appointment scheduling in health care: challenges and opportunities. IIE Trans 2008;40(9):800–19.
[20] Hassin R, Mendel S. Scheduling arrivals to queues: a single-server model with no-shows. Manag Sci 2008;54(3):565–72.
[21] Huschka TR, Denton BT, Narr BJ, Thompson AC. Using simulation in the implementation of an outpatient procedure center. In: Winter simulation conference; 2008. p. 1547–52.
[22] Jonas DE, Russell LB, Sandler RS, Chou J, Pignone M. Value of patient time invested in the colonoscopy screening process: time requirements for colonoscopy study. Med Decis Mak 2008;28(1):56–65.
[23] Kelton WD, Sadowski RP, Sturrock DT. Simulation with Arena. fourth edition. Boston: McGraw-Hill; 2007.
[24] Kim S, Giachetti RE. A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. IEEE Trans Syst Man Cybern —Part A 2006;36(6):1211–9.
[25] Klassen KJ, Rohleder TR. Scheduling outpatient appointments in a dynamic environment. J Oper Manag 1996;14(19):83–101.
[26] LaGanga LR, Lawrence SR. Clinic overbooking to improve patient access and increase provider productivity. Decis Sci 2007;38(2):251–76.
[27] Macharia WM, Leon G, Rowe BH, Stephenson BJ, Haynes RB. An overview of interventions to improve compliance with appointment keeping for medical services. J Am Med Assoc 1992;267(13):1813–7.
[28] Mancilla Camilo, H Storer Robert. Stochastic sequencing of surgeries for a single surgeon operating in parallel operating rooms. IIE Trans Healthc Syst Eng 2013;3(2):127–38.
[29] Muthuraman K, Lawley M. A stochastic overbooking model for outpatient clinical scheduling with no-shows. IIE Trans 2008;40(9):820–37.
[30] Ostrowski J, Linderoth J, Rossi F, Smriglio S. Orbital branching. Math Program 2011;126(1):147–78.
[31] Rockafellar RT, Wets RJ-B. Scenarios and policy aggregation in optimization under uncertainty. Math Oper Res 1991;16(1):119–47.
[32] Rohleder TR, Klassen KJ. Using client-variance information to improve dynamic appointment scheduling performance. Omega 2000;28(3):293–302.
[33] Rothstein M. Hotel overbooking as a Markovian sequential decision process. Decis Sci 1974;5(3):389–404 ISSN: 1540-5915.
[34] Rothstein M. OR and the airline overbooking problem. Oper Res 1985;33(2): 237–248.
[35] Soriano A. Comparison of two scheduling systems. Oper Res 1966;14(3): 388–397.
[36] Van Slyke RM, Wets R. L-shaped linear programs with applications to optimal control and stochastic programming. SIAM J Appl Math 1969;17(4):638–63.
[37] Vanden Bosch PM, Dietz DC. Minimizing expected waiting in a medical appointment system. IIE Trans 2000;32(9):841–8.
[38] Vargo JJ, Bramley T, Meyer K, Nightengale B. Practice efficiency and economics: the case for rapid recovery sedation agents for colonoscopy in a screening population. J Clin Gastroenterol 2007;41(6):591–8.
[39] Weiss EN. Models for determining estimated start times and case orderings in hospital operating rooms. IIE Trans 1990;22(2):143–50.
[40] Zeng B, Turkcan A, Lin J, Lawley M. Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. Ann Oper Res 2010;178(1):121–44.