European Journal of Operational Research 280 (2020) 254–265

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

Dynamic multi-priority, multi-class patient scheduling with stochastic service times

Antoine Sauré[a,*], Mehmet A. Begen[b], Jonathan Patrick[a]

[a] Telfer School of Management, University of Ottawa, 55 Laurier Avenue East, Ottawa, ON K1N 6N5, Canada
[b] Ivey School of Business, Western University, 1255 Western Road, London, ON N6G ON1, Canada

ARTICLE INFO

Article history:
Received 16 August 2017
Accepted 19 June 2019
Available online 21 June 2019

Keywords:
OR in health services
Patient scheduling
Markov decision processes
Approximate dynamic programming
Linear programming

ABSTRACT

Efficient patient scheduling has significant operational, clinical and economical benefits on health care systems by not only increasing the timely access of patients to care but also reducing costs. However, patient scheduling is complex due to, among other aspects, the existence of multiple priority levels, the presence of multiple service requirements, and its stochastic nature. Patient appointment (allocation) scheduling refers to the assignment of specific appointment start times to a set of patients scheduled for a particular day while advance patient scheduling refers to the assignment of future appointment days to patients. These two problems have generally been addressed separately despite each being highly dependent on the form of the other. This paper develops a framework that incorporates stochastic service times into the advance scheduling problem as a first step towards bridging these two problems. In this way, we not only take into account the waiting time until the day of service but also the idle time/overtime of medical resources on the day of service. We first extend the current literature by providing theoretical and numerical results for the case with multi-class, multi-priority patients and deterministic service times. We then adapt the model to incorporate stochastic service times and perform a comprehensive numerical analysis on a number of scenarios, including a practical application. Results suggest that the advance scheduling policies based on deterministic service times cannot be easily improved upon by incorporating stochastic service times, a finding that has important implications for practice and future research on the combined problem.

1. Introduction

The large body of literature associated with patient scheduling can broadly be divided into two streams: appointment (or allocation) scheduling and advance scheduling. Advance scheduling refers to the allocation of future service capacity to demand as it arrives. It is most often done on a daily basis. Appointment scheduling, on the other hand, refers to the assignment of specific appointment times and resources to patients but only once all patients for a given service day have been identified. While there is a significant stream of literature regarding both types of scheduling problem, little work has been done that attempts to combine the two despite the potential of their being highly dependent on each other. A typical advance scheduling model will use resource utilization or overtime as a relevant performance metric. However, in the presence of stochastic service times, utilization and overtime are clearly dependent on the form of the appointment schedule. Conversely, the key input to an appointment scheduling problem is the number of patients to be served, which is precisely the output of an advance scheduling model. This interdependency of the two scheduling problems provides the impetus for research that looks to determine the advantage of solving them as a single problem.

Advance scheduling problems typically assume that patients can be classified into multiple types according to their capacity requirements and urgency, that there is (at least) one resource that has a fixed regular-hour capacity, that there exists the possibility of using overtime or an alternative source of surge capacity, and that service durations are deterministic. The aim is to identify effective ways of allocating available service capacity to incoming appointment requests while either maximizing the service level (the number of patients booked within medically acceptable wait times) in a cost-effective manner or else maximizing revenue or throughput. Application areas include the scheduling of diagnostic tests such as MRIs (Schütz & Kolisch, 2012) or CT scans (Patrick, Puterman, & Queyranne, 2008) as well as radiation therapy treatments (Sauré, Patrick, Tyldesley, & Puterman, 2012). Papers in the area of advance

* Corresponding author.
E-mail addresses: asaure@uottawa.ca (A. Sauré), mbegen@ivey.uwo.ca (M.A. Begen), patrick@telfer.uottawa.ca (J. Patrick).

https://doi.org/10.1016/j.ejor.2019.06.040
0377-2217/

scheduling mostly use dynamic programming, or approximate dynamic programming, due to the sequential nature of the scheduling decisions.

Appointment (or allocation) scheduling problems typically consider a single resource with high idle time and overtime costs. The objective is to determine appointment times (sequence) and appointment durations such that some combination of the costs associated with patient within-day waits, resource idle time and overtime is minimized. Appointment times are needed before the service day. The challenge comes from uncertain service times. There are many applications for appointment scheduling with two significant applications being surgery and physician appointment scheduling (Begen & Queyranne, 2011). Stochastic programming, queuing theory, simulation and simulation-optimization are commonly used methodologies to solve this type of problem.

In advance scheduling problems, the assumption of deterministic service times is largely made for convenience and in the general hope that, over time, average service times will work fairly well as an approximation. This assumption allows overtime or idle time to be easily calculated as the number of appointments booked on a given day times the appointment length minus the regular-hour capacity. In the absence of this assumption, the performance metrics of the advance schedule (namely overtime and idle time) are dependent on the appointment schedule being used. In this paper, we show that advance scheduling policies based on deterministic service times cannot be easily improved upon by incorporating stochastic service times, a finding that has important implications for practice and future research on the combined scheduling problem.

To that end, we begin with the advance scheduling model provided in Patrick et al. (2008). Patients request service on a daily basis and are categorized based on urgency into multiple priority classes, each class with its own wait time target (medically acceptable wait time). The aim of the scheduler is to ensure that as many patients as possible are served within their wait time targets and with as little overtime as possible. The model penalizes scheduling a patient past his/her wait time target but does not prohibit it. Patrick et al. (2008) model this problem as a Markov Decision Process (MDP). However, the size of the state space and the corresponding action sets prohibit the determination of a solution via traditional methods. Instead, they resort to the linear programming approach to Approximate Dynamic Programming (ADP). In this approach, the value function in the MDP model is approximated by a function that is affine in the state variables and then the optimal approximation parameter values are determined using mathematical programming. The authors were able to determine the form of the optimal affine approximation under mild conditions on the availability of sufficient regular-hour and overtime capacity. This allowed for the characterization of the approximate optimal policy without having to solve an optimization model every time scheduling decisions were required. The result was an easy-to-implement heuristic policy that performs well in practice.

One of the limiting assumptions of Patrick et al. (2008) is that service times are homogeneous. The first contribution of this paper is to remove that assumption and allow for patient classification both by priority and by resource consumption. Initially, we maintain the assumption that resource consumption is deterministic. In this setting, we are able to show that the optimal affine approximation continues to have a predictable (though different) form and thus the approximate optimal policy can again be characterized and implemented without recourse to an optimization model. The proof of this theoretical result is provided in the online supplement.

The addition of stochastic service times significantly complicates the advance scheduling problem due to the fact that the computation of overtime now depends on the appointment schedule that defines the appointment time for each patient. For the purposes of this paper, we focus on the case where the only performance metric of interest for the appointment schedule is the overtime or idle time that results at the end of the service day. This assumption is reasonable in settings such as surgical scheduling where patient wait times on the day of service are not considered a major factor and where the possibility of resource idle time between appointments is removed by requiring patients to arrive well before their scheduled surgery times. The value of this simplifying assumption is that the calculation of the performance metrics does not depend on the sequencing of patients but only on their service times. Thus, the MDP model continues to evaluate the scheduling policy based on a combination of meeting the wait time target for each patient and minimizing the overtime and idle time at the end of the day of service. However, the calculation of the overtime or idle time is now based on a method described in Begen and Queyranne (2011) that determines the optimal appointment schedule as a function of known stochastic service time distributions.

We consider service times with discrete probability distributions. Thus, the expected cost on the service day can be computed efficiently by using recursive equations for a given number of patients. Expected cost computations are fast (i.e., polynomial in the number of patients and in the largest service time) and can be done easily on an as-needed basis. Furthermore, these computations are flexible. We can either work with discrete probability distributions, if they are available, or samples of service durations. In the case of samples, we can work with correlated durations. Expected service-day costs are computed and incorporated into the MDP model.

In this paper, we describe a model setting that is meant to be a first step towards bridging the advance and appointment scheduling problems. We obtain theoretical and numerical results for the advance scheduling problem with deterministic service times in the case of multi-priority, multi-class patients. Next, we incorporate stochastic service times into our model. We conduct a comprehensive analysis to compare the performance of the model with stochastic service durations to that of the model with deterministic service durations as well as against the performance of benchmark policies used in practice. In addition, we provide an extensive literature review of the advance and appointment scheduling problems.

The rest of the paper is organized as follows. In Section 2, we provide an exhaustive overview of the relevant literature. In Section 3, we describe an enhanced version of the MDP model presented in Patrick et al. (2008) that can be used to incorporate patient classes differentiated by both priority and resource consumption as well as stochastic service times. In Section 4, we first characterize the model with deterministic service times and then the model with stochastic service times. In particular, we state the theoretical results for the deterministic model with multiple service classes and show how stochastic service durations are incorporated. Then, in Section 5, we present extensive numerical results designed to determine the impact of incorporating stochastic service times and compare the performance of the resulting scheduling policies to that of benchmark policies used in practice. We also present results from a practical application based on data provided by a medium-size clinic in Canada. Finally, in Section 6, we provide a discussion and review our main conclusions.

2. Related literature

Patient scheduling is complex due to, among other aspects, the existence of multiple priority levels, the presence of multiple service requirements, and its stochastic nature. As much as it is complex, it is also critical to have good and effective schedules to

ensure that patients receive timely access to medical services in a cost-efficient manner. This is especially important now that health care costs and demand for medical services are on the rise and almost all countries are under constant pressure to improve health care efficiency while reducing costs.

Two important challenges in patient scheduling are the presence of random patient arrivals and the existence of random resource requirements. To address these challenges, researchers have mostly focused on one source of uncertainty at a time. Advance scheduling models deal with random patient arrivals whereas appointment (or allocation) scheduling models consider random service durations. Next, we provide an extensive literature survey on advance and appointment scheduling.

Appointment scheduling has been studied extensively over the last 50 years, starting with the well-known paper by Bailey (1952). This paper recommends booking two patients at the beginning of the day and patients equally spaced thereafter in order to avoid any idle time for doctors. Application areas of appointment scheduling include surgery scheduling (Begen & Queyranne, 2011; Denton, Viapiano, & Vogl, 2007) and physician appointment scheduling (Kaandoorp & Koole, 2007; Robinson & Chen, 2003). The main objective in appointment scheduling is to determine how much time to reserve for each appointment (Begen & Queyranne, 2011; Denton & Gupta, 2003; Mancilla & Storer, 2012; Robinson & Chen, 2003; Wang, 1993) and/or the number of appointments to book at each pre-determined time interval (Bosch, Vanden, Dietz, & Simeoni, 1999; Cayirli, Kum, & Quek, 2012; Chakraborty, Muthuraman, & Lawley, 2010; Kaandoorp & Koole, 2007; Muthuraman & Lawley, 2008; Zeng, Turkcan, Lin, & Lawley, 2009) in order to minimize a weighted combination of the expected overtime and idle time of the resource (e.g., doctor's time or operating room hours) and the expected waiting time for patients. Almost all papers in the literature focus on the optimization of expected costs or rewards though there are some studies that consider other objectives. For example, Mittal, Schulz, and Stiller (2014) consider the worst-case scenario of the realized service times.

In appointment scheduling, the total number of patients is assumed to be known in advance and planned appointment times are needed before any patient is served. The main challenge comes from uncertain appointment durations. Most papers in the literature use continuous probability distributions to describe service durations. However, there are some important advantages to considering discrete distributions instead (Begen & Queyranne, 2011). Discrete probability distributions enable efficient expected value computations and polynomial time algorithms for optimizing schedules. In addition, they can be easily used to incorporate no-shows and, to a certain extent, emergencies (Begen & Queyranne, 2011). There are also a few papers in the literature that discuss settings with partial, limited or no information on the service time distributions (Begen, Levi, & Queyranne, 2012; Ge, Wan, & Zhang, 2013; Kong, Lee, Teo, & Zheng, 2013; Mak, Rong, & Zhang, 2014).

Although most studies assume that a fixed sequence of patients is given, and only determine an optimal schedule, the optimal sequencing of patients at the same time and in addition to determining the optimal appointment times is also considered in the literature. The problem, however, increases in complexity as the number of possible sequences grows quickly with the number of patients. Unsurprisingly, this problem is NP-hard (Mancilla & Storer, 2009). Some results exist on the optimal sequencing of two patients. For example, it has been demonstrated that scheduling patients in increasing order of service time variance is optimal (Denton et al., 2007; Gupta, 2007; Weiss, 1990). For a higher number of patients, the problem seems to be open and it is very unlikely that a universal optimal sequencing rule could be found unless some restrictions on the service time distributions and costs are imposed. For example, a recent paper by Guda, Dawande, Janakiraman, and Jung (2016) demonstrated that the shortest-variance-first rule is optimal for the single-machine earliness/tardiness problem if the earliness and tardiness cost parameters are the same for all the jobs and there is a dilation ordering of the processing times. The authors discuss the application of this rule to the appointment scheduling problem with sequencing in the case where idling is not allowed. Although an optimal sequencing rule for the general problem is not yet available, Denton et al. (2007) show that ordering patients by increasing service time variance works well. In addition, Chen and Robinson (2014) describe heuristic policies that work well in the presence of two classes of patients. Mak et al. (2014) show, using inventory approximations, that sequencing jobs based on increasing (standard deviation)/(overtime cost)$^{\gamma}$, where $\gamma \in \{1, 0.5\}$, performs better than simply ordering based on increasing variance.

While it is common to assume that both medical resources (e.g., doctors) and patients are punctual, there is substantial evidence suggesting that this assumption is often unwarranted. Furthermore, doctors' schedules can be interrupted by other, sometimes more urgent, tasks. Motivated by these considerations, there are a few papers that study the effect of unpunctuality and interruptions on the optimal appointment schedule (Klassen & Yoogalingam, 2009; 2014; Luo, Kulkarni, & Ziya, 2012; Samorani & LaGanga, 2015). There are also papers that consider the potential impact of patient no-shows, overbooking, and cancellations. Some of these papers resemble appointment scheduling, others come closer to advance scheduling.

The main techniques used for appointment scheduling are stochastic programming (Begen & Queyranne, 2011; Denton & Gupta, 2003; Denton et al., 2007; Mancilla & Storer, 2012; Robinson & Chen, 2003), the newsvendor approach (Weiss, 1990), queuing theory (Kaandoorp & Koole, 2007; Klassen & Rholeder, 1996; Wang, 1993), simulation (Ma, Sauré, Puterman, Taylor, & Tyldesley, 2016; Santibáñez, Begen, & Atkins, 2007; White, Froehle, & Klassen, 2011), and simulation-optimization (Klassen & Yoogalingam, 2008; 2009).

Advance scheduling differs from appointment scheduling in that arrivals are random while service durations are generally assumed to be deterministic (i.e., fixed). Advance scheduling thus deals with patient waiting times until the day of service (e.g., days until an MRI appointment or surgery), whereas appointment scheduling deals with costs incurred on the day of service (e.g., waiting time penalties, overtime and idle time costs). In advance scheduling one needs a booking (or capacity allocation) policy to be used continuously, whereas in appointment scheduling one needs an appointment schedule before any patient is served.

Advance scheduling can be viewed as a stochastic capacity allocation problem in which the trade-off between capacity utilization and waiting times, or between revenue and service levels, is modelled for different types of patients (Dobson, Hasija, & Pinker, 2011; Erdelyi & Topaloglu, 2009; Patrick et al., 2008; Sauré et al., 2012; Truong, 2015). Alternatively, it can be considered a revenue management problem in which the main decision is to determine whether or not to accept an incoming service request (Schütz & Kolisch, 2012; 2013). While most studies in advance scheduling focus on a single resource (Dobson et al., 2011; Patrick et al., 2008), there are others that consider multiple resources (Astaraky & Patrick, 2015; Gocgun & Ghate, 2012; Truong, 2015). Other than in a few cases (Feldman, Liu, Topaloglu, & Ziya, 2014; Gupta & Denton, 2008; Wang & Gupta, 2011), patient preferences have not been considered.

No-shows, in both appointment and advance scheduling, have received a significant amount of attention from researchers in the last few years (Liu, 2016; Tang, Yan, & Cao, 2014; Tsai & Teng, 2014). High no-show rates can significantly paralyze a service and cause the double negative effect of low resource utilization and high waiting times. To overcome these negative outcomes, re-

searchers have considered overbooking (Huang & Zuniga, 2012; La-Ganga & Lawrence, 2012; Zacharias & Pindeo, 2014), studied the relationship between panel size and no-show rates (Green & Savin, 2008; Liu, 2016), and examined the impact of open-access policies (Patrick, 2012; Robinson & Chen, 2010). Open-access policies are also known as "same-day" policies. Patients call on the day of their appointments or only a few days before. There are also studies that consider cancellations (Liu, Ziya, & Kulkarni, 2009; Schütz & Kolisch, 2012) and daily no-show estimates (Samorani & LaGanga, 2015).

Papers in the area of advance scheduling mostly use dynamic programming or approximate dynamic programming due to the stochastic nature of the appointment request arrivals and the sequential nature of the decision process (Patrick et al., 2008; Sauré, Patrick, & Puterman, 2015; Sauré et al., 2012; Schütz & Kolisch, 2012; 2013; Truong, 2015). Some of the objectives considered in these studies are: maximizing the number of patients booked within their medically acceptable wait times (Patrick et al., 2008; Sauré et al., 2015; Sauré et al., 2012), maximizing revenue (Gupta & Denton, 2008; Schütz & Kolisch, 2013), improving resource utilization (Santibáñez, Chow, French, Puterman, & Tyldesley, 2009), satisfying specific appointment date windows (Gocgun & Puterman, 2014), taking patient preferences into account (Feldman et al., 2014; Gupta & Denton, 2008; Wang & Gupta, 2011), and reducing wait times (Green, Savin, & Wang, 2006). Application areas of advance scheduling include the scheduling of diagnostic tests such as MRI or CT scans (Green et al., 2006; Patrick et al., 2008; Schütz & Kolisch, 2012), radiation therapy treatment scheduling (Sauré et al., 2012), primary care clinics (Dobson et al., 2011; Green & Savin, 2008; Liu, 2016), and surgical scheduling (Astaraky & Patrick, 2015).

To the best of our knowledge, most related to our work are the papers of Muthuraman and Lawley (2008), Zeng et al. (2009), Chakraborty et al. (2010), and Erdogan and Denton (2013). In Muthuraman and Lawley (2008), the authors consider an advance scheduling problem with no-shows and determine the number of appointments per day required to maximize revenue. They assume appointments slots of fixed length and exponentially distributed service times. Zeng et al. (2009) and Chakraborty et al. (2010) extend Muthuraman and Lawley (2008)'s work to the case of heterogeneous no-show probabilities and general service time distributions, respectively. To some degree, these three papers combine advance scheduling with appointment scheduling. However, the authors assume that appointment durations are fixed, of equal length, and determined exogenously. Except for no-show rates, patients are homogeneous (i.e., same service time distributions and priority) and the admission decisions are made at the time of the booking requests. In Erdogan and Denton (2013), the authors use stochastic programming to formulate a dynamic appointment scheduling problem with uncertain demand and homogeneous patients. The maximum number of patients that can be scheduled is known and each appointment request is probabilistic given the state of the previous one. The authors only consider single-period models and provide a conceptual multi-stage stochastic programming formulation for the case of probabilistic appointment requests. They provide some structural properties and numerical results for the case of deterministic patient arrivals with no-shows. None of these four papers attempts to characterize the optimal solution. Except for these papers, we are not aware of any other studies that combine both random patient arrivals and random service times in the context of appointment scheduling.

In this paper, we incorporate stochastic service times into the advance scheduling problem as a first step towards bridging advance and appointment scheduling problems. We consider random appointment requests coming from multiple types of patients as well as random appointment durations. Patients are classified on the basis of resource consumption and priority (given by urgency level or maximum recommended wait time). We take into account the wait until the day of service as well as the medical resource (e.g., operating room or surgeon) idle time and overtime at the end of the day of service. Patients' waiting within the day of service is not considered.

3. An MDP formulation for the patient scheduling problem

In this section, we formulate a discounted infinite-horizon MDP model for the problem under study. We assume that demand for service has been broken down into I priority classes and J service classes. Service classes are differentiated by the service time probability distribution or by the mean service time in the deterministic case. For notational simplicity, we let $[A] = \{1, \ldots, A\}$. We denote a vector by bolding it, such as $\mathbf{s}$.

3.1. Decision epochs and the booking horizon

We consider a system that has a capacity of C^R regular time units and C^{OT} overtime units per day. At a specific point of time every day, referred to as the decision epoch, the scheduler observes the number of booked appointments from each service class on each future day over an N-day booking horizon and the number of cases in each priority class-service class pairing waiting to be scheduled. The booking horizon is defined by the maximum number of days in advance that the scheduler is allowed to schedule patients.

We assume that the number of patients scheduled into a given service day is known prior to the start of the day. This, for example, reflects a hospital that chooses to reserve capacity for emergency surgeries rather than impinge on the elective surgical schedule. Our model is complicated by the fact that the booking horizon is not static but rolling. Thus, day n at the current decision epoch becomes day $n-1$ at the subsequent decision epoch. Since no patient is scheduled more than N days in advance, at the beginning of each decision epoch, the N^{th} day has no appointments booked.

3.2. The state space

As mentioned above, we assume that demand for service is broken down into multiple priority classes, based on urgency, and multiple service classes, based on the probability distribution of the length of service. Our state takes the form $\mathbf{s} = (\mathbf{x}, \mathbf{y})$, where x_{jn} is the number of patients from service class j already booked on day n and y_{ij} is the number of priority i patients from service class j waiting to be booked. The state space needs to capture the number of patients of each service class booked into each day of the booking horizon as the service class mix will play a key role in determining overtime and idle time costs and the optimal appointment schedule, particularly in the scenario with stochastic service times. In contrast, prioritization does not need to be tracked once a patient has been scheduled to a specific day as the value of any late booking penalty is determined at the time of booking.

3.3. The potential action sets

The scheduler's task is to decide at each decision epoch on which day to schedule each of the patients waiting to be booked. Thus, a vector of possible actions can be written as $\mathbf{a}$, where a_{ijn} is the number of priority i patients from service class j to book on day n. To accommodate the potential for overtime, bookings are allowed to exceed the daily regular-hour capacity up to a limit C^{OT} of overtime units. To be valid, the number of bookings cannot exceed

the number of patients waiting,

$$\sum_{n=1}^{N} a_{ijn} \le y_{ij} \qquad \forall (i,j) \in [I] \times [J], \tag{1}$$

and the overtime capacity cannot be exceeded,

$$\sum_{j=1}^{J} \mu_j \left(x_{jn} + \sum_{i=1}^{I} a_{ijn} \right) \le C^R + C^{OT} \qquad \forall n \in [N], \tag{2}$$

where μ_j is the number of time units required by a patient of service class j. For the stochastic case, μ_j can represent either the average service time or else some percentile of the service time distribution depending on how conservative the scheduler wants to be. Finally, all actions are constrained to be positive and integer,

$$\mathbf{a} \in \mathbb{Z}_{[I]\times[J]\times[N]}. \tag{3}$$

We define the action set $A(\mathbf{s})$, for any given state $\mathbf{s}$, as the set of actions $\mathbf{a}$ satisfying Eqs. (1)–(3).

3.4. Transition probabilities

Once decisions are made, the only stochastic element in the transition to the next state of the system is due to the number of new appointment requests in each priority class-service class pairing. Demand that is not booked today re-appears in tomorrow's demand. If the number of new patient arrivals is represented by vector $\mathbf{y}'$, then

$$x_{jn} \leftarrow x_{j,n+1} + \sum_{i=1}^{I} a_{ij,n+1} \qquad \forall (j,n) \in [J] \times [N],$$

$$y_{ij} \leftarrow y'_{ij} + y_{ij} - \sum_{n=1}^{N} a_{ijn} \qquad \forall (i,j) \in [I] \times [J],$$

where $x_{j,N+1} = a_{ij,N+1} = 0$. We assume demand from each patient priority class-service class pairing is independent and that each day's demand is independent as well. Expected demand values are denoted by λ_{ij}.

3.5. Immediate costs

The cost associated with a given state-action pair derives from three sources: penalties associated with booking patients beyond their priority-specific wait time targets, a cost associated with the day of service (overtime or idle time cost), and penalties associated with delaying the booking decisions for some of the waiting demand. We write the immediate cost function as

$$c(\mathbf{s},\mathbf{a}) = \sum_{(i,n)\in[I]\times[N]} f^{WT}(i,n) \sum_{j\in[J]} a_{ijn} + f^{AS}(\mathbf{s},\mathbf{a}) + \sum_{(i,j)\in[I]\times[J]} f^{D}(i) \left(y_{ij} - \sum_{n\in[N]} a_{ijn} \right), \tag{4}$$

where $f^{WT}(i,n)$ is the penalty associated with booking a priority i patient on day n and $f^{AS}(\mathbf{s},\mathbf{a})$ is the cost associated with the appointment schedule. The latter could consist of a combination of within-day wait times of patients as well as idle time and overtime at the end of the day. This cost will be discussed further in Section 4 as we look at both the deterministic and the stochastic versions of the model separately. $f^{D}(i)$ is the penalty associated with delaying the booking of a priority i patient one day. We represent the wait time target for priority i by $T(i)$. The choice of $f^{WT}(i,n)$, though arbitrary, should consider certain characteristics. It is clearly reasonable to assume that it should be decreasing in i and zero if $n \le T(i)$. Furthermore, it would seem advisable to ensure that the penalty associated with delaying the booking of a patient k days and then booking him/her within the corresponding wait time target should be equal to the penalty associated with booking the patient k days late initially. Thus, a natural form for the wait time penalty is

$$f^{WT}(i,n) = \begin{cases} \sum_{k=1}^{n-T(i)} \gamma^{k-1} f^{D}(i), & \text{for all } n > T(i); \\ 0, & \text{otherwise.} \end{cases}$$

where γ is the daily discount factor.

The immediate cost function $c(\mathbf{s},\mathbf{a})$ explicitly balances the cost to the patients in wait times and the cost to the system in having to resort to overtime or having idle time. The scheduler's role is to maintain reasonable patient wait times in a cost-effective manner.

3.6. The structure of the Bellman equation

The value function v of the MDP specifies the minimum expected discounted cost over the infinite horizon for each state and satisfies the following optimality equations:

$$v(\mathbf{s}) = \min_{\mathbf{a}\in A(\mathbf{s})} \left\{ c(\mathbf{s},\mathbf{a}) + \gamma \sum_{\mathbf{y}'\in D} p(\mathbf{y}') v\left(x_{12} + \sum_{i=1}^{I} a_{i12}, \ldots, x_{JN} + \sum_{i=1}^{I} a_{ijN}, 0; y'_{11} + y_{11} - \sum_{n=1}^{N} a_{11n}, \ldots, y'_{IJ} + y_{IJ} - \sum_{n=1}^{N} a_{IJn} \right) \right\} \quad \forall \mathbf{s} \in S, \tag{5}$$

where D is the set of all possible incoming demand streams and $p(\mathbf{y}')$ is the probability of a vector $\mathbf{y}'$ of new demand.

4. Deterministic and stochastic models

The formulation presented in the previous section provides a comprehensive model that incorporates multiple patient classes with a flexible function $f^{AS}(\mathbf{s},\mathbf{a})$ for within-day costs. In this section, we show how $f^{AS}(\mathbf{s},\mathbf{a})$ can be computed for the deterministic case and the stochastic case and describe an ADP version of the model.

4.1. Deterministic model

In the deterministic case, the within-day costs are reduced to overtime and idle time costs and can be computed as

$$f^{AS}(\mathbf{s},\mathbf{a}) = h \left[\sum_{j=1}^{J} \mu_j \left(x_{j1} + \sum_{i=1}^{I} a_{ij1} \right) - C^R \right]^+ + u \left[C^R - \sum_{j=1}^{J} \mu_j \left(x_{j1} + \sum_{i=1}^{I} a_{ij1} \right) \right]^+, \tag{6}$$

where h is the overtime cost per time unit, u is the idle time cost per time unit, and $[z]^+ = \max\{0, z\}$.

Even in the deterministic case, the size of the state space and the size of the corresponding action sets require that we approximately solve our MDP model via ADP. To that end, we assume that the value function in our formulation can be adequately represented by an affine approximation architecture in the form:

$$V(\mathbf{s}) = V_0 + \sum_{\substack{(j,n)\in \\ [J]\times[N]}} V_{jn} x_{jn} + \sum_{\substack{(i,j)\in \\ [I]\times[J]}} W_{ij} y_{ij} \quad \forall \mathbf{s} \in S \qquad \mathbf{V}, \mathbf{W} \ge 0, V_0 \in \mathbb{R}. \tag{7}$$

There are two main options in seeking to solve an MDP model using ADP. Simulation-based ADP iteratively produces simulated

runs of the scheduling process to approximate the value function at a subset of initial states while using a form of least squares regression or a recursive update function to converge on a good approximation (e.g., see Sauré et al., 2015). The other option is to transform the MDP model into an equivalent linear program and then substitute into it the value function approximation of choice. This is the approach taken here and in Patrick et al. (2008).

Once the approximation given in Eq. (7) is substituted into the linear programming formulation of the discounted MDP, we get the following approximate linear program (ALP):

$$\min_{\substack{\mathbf{V},\mathbf{W}\geq 0,\\ V_0\in\mathbb{R}}} \quad \sum_{\mathbf{s}\in S}\alpha(\mathbf{s})\left(V_0+\sum_{\substack{(j,n)\in\\ [J]\times[N]}}V_{jn}x_{jn}+\sum_{\substack{(i,j)\in\\ [I]\times[J]}}W_{ij}y_{ij}\right) \tag{8}$$

subject to

$$(1-\gamma)V_0+\sum_{\substack{(j,n)\in\\ [J]\times[N]}}\left(x_{jn}-\gamma x_{j,n+1}-\gamma\sum_{i\in[I]}a_{ij,n+1}\right)V_{jn}$$
$$+\sum_{\substack{(i,j)\in\\ [I]\times[J]}}\left((1-\gamma)y_{ij}+\gamma\sum_{n\in[N]}a_{ijn}-\gamma E[Y_{ij}]\right)W_{ij}\geq c(\mathbf{s},\mathbf{a})$$
$$\forall(\mathbf{s},\mathbf{a})\in S\times A(\mathbf{s}),$$

where $\boldsymbol{\alpha}$ is any positive-valued vector. In traditional MDP theory, the choice of $\boldsymbol{\alpha}$ is arbitrary as any $\boldsymbol{\alpha}>0$ will lead to the same solution. This turns out not to be the case when using ADP as the choice of $\boldsymbol{\alpha}$ plays a key role in the solution. The value of $\boldsymbol{\alpha}$ in the latter case is best interpreted as the probability distribution over the initial state of the system.

While the ALP remains intractable due to the existence of a constraint for every state-action pair, its dual can be solved via column generation. The dual of the ALP can be written as:

$$\max_{\mathbf{X}\geq 0} \quad \sum_{\substack{(\mathbf{s},\mathbf{a})\in\\ S\times A(\mathbf{s})}}c(\mathbf{s},\mathbf{a})X(\mathbf{s},\mathbf{a}) \tag{9}$$

subject to

$$(1-\gamma)\sum_{\substack{(\mathbf{s},\mathbf{a})\in\\ S\times A(\mathbf{s})}}X(\mathbf{s},\mathbf{a})=1,$$

$$\sum_{\substack{(\mathbf{s},\mathbf{a})\in\\ S\times A(\mathbf{s})}}\left(x_{jn}-\gamma x_{j,n+1}-\gamma\sum_{i=1}^{I}a_{ij,n+1}\right)X(\mathbf{s},\mathbf{a})\geq E_{\alpha}[X_{jn}]$$
$$\forall(j,n)\in[J]\times[N],$$

$$\sum_{\substack{\mathbf{s},\mathbf{a}\in\\ S\times A(\mathbf{s})}}\left((1-\gamma)y_{ij}+\gamma\sum_{n\in[N]}a_{ijn}-\gamma E[Y_{ij}]\right)X(\mathbf{s},\mathbf{a})\geq E_{\alpha}[Y_{ij}]$$
$$\forall(i,j)\in[I]\times[J].$$

The dual variable $X(\mathbf{s},\mathbf{a})$ can be interpreted as the frequency of taking action $\mathbf{a}$ when in state $\mathbf{s}$. Following our earlier comment regarding $\boldsymbol{\alpha}$, one can interpret $E_{\boldsymbol{\alpha}}[X_{jn}]$ as the expected number of patients of type j booked into day n and $E_{\boldsymbol{\alpha}}[Y_{ij}]$ as the expected number of new arrivals of priority class i and service class j both associated with the initial state of the system.

Once the optimal value function approximation has been determined then the approximate optimal decision policy is derived by determining the arg min of the optimality equation given in Eq. (5), with the optimal approximation inserted in the place of the value function. If the regular-hour capacity on day 1 is full, that is $\sum_{j\in[J]}\mu_j(x_{j1}+\sum_{i\in[I]}a_{ij1})-C^R>0$, then the approximate optimal action $\mathbf{a}^*$ is given by

$$\min_{\mathbf{a}\in A(\mathbf{s})}\left[\mathcal{C}+\sum_{\substack{(i,j)\in\\ [I]\times[J]}}\left(h\mu_j-f^D(i)-\gamma W^*_{ij}\right)a_{ij1}\right.$$
$$\left.+\sum_{\substack{(i,j)\in\\ [I]\times[J]}}\sum_{n=2}^{N}\left(f^{WT}(i,n)+\gamma V^*_{j,n-1}-f^D(i)-\gamma W^*_{ij}\right)a_{ijn}\right]. \tag{10}$$

Otherwise, if there is regular-hour capacity available on day 1, then $\mathbf{a}^*$ is given by

$$\min_{\mathbf{a}\in A(\mathbf{s})}\left[\mathcal{C}+\sum_{\substack{(i,j)\in\\ [I]\times[J]}}\left(-u\mu_j-f^D(i)-\gamma W^*_{ij}\right)a_{ij1}\right.$$
$$\left.+\sum_{\substack{(i,j)\in\\ [I]\times[J]}}\sum_{n=2}^{N}\left(f^{WT}(i,n)+\gamma V^*_{j,n-1}-f^D(i)-\gamma W^*_{ij}\right)a_{ijn}\right], \tag{11}$$

where V^*_{jn} and W^*_{ij} represent the optimal values of the approximation parameters and $\mathcal{C}$ is a constant independent of the action taken. The coefficients in Eqs. (10) and (11) have a nice intuitive explanation. In the case where day 1's capacity is full, the bookings on day 1 trade off the cost associated with overtime for the cost of delaying a booking and having an additional patient in the wait list tomorrow. In the case where there is excess capacity on day 1, there is only a benefit to making use of that capacity as it reduces the idle time cost. Thus, the coefficients for bookings on day 1 are negative. The bookings on day $n>1$, in both cases, trade off a potential wait time penalty plus the cost of having less available capacity on day $n-1$ tomorrow for the cost of delaying a booking and having an additional patient in the wait list tomorrow.

Somewhat surprisingly but as a natural extension to Patrick et al. (2008), the form of the optimal affine value function approximation can be proven under certain conditions on the cost values and the available capacity as outlined in Theorem 1. The value of the wait time target $T(i)$ is assumed to increase with i as a high priority patient is, by definition, a patient who must be served sooner. We define the indicator function $I(\cdot)$ as

$$I(X>x)=\begin{cases}1, & X>x;\\ 0, & \text{otherwise.}\end{cases}$$

to ease the presentation.

Theorem 1. *Assuming that $T(i)$ is non-decreasing in i, that the wait time penalties are non-decreasing in n and non-increasing in i, and that the following conditions are satisfied:*

$$f^D(i)>\left(\gamma^{T(i)-1}-\gamma^{T(i)}\right)\mu_j h \quad \forall(i,j)\in[I]\times[J] \tag{12}$$

$$\sum_{j\in J}\mu_j\left[\sum_{i=1}^{I}\frac{\gamma^{T(i)-1}I(T(i)>1)\lambda_{ij}}{1-\gamma}+\sum_{m=1}^{N}\gamma^{m-1}E_{\alpha}[X_{jm}]\right]>\frac{C^R}{1-\gamma} \tag{13}$$

$$\sum_{j\in J}\mu_j\left[\sum_{i=1}^{I}\frac{\gamma^{T(i)-n}I(T(i)>n)\lambda_{ij}}{1-\gamma}+\sum_{m=n}^{N}\gamma^{m-n}E_{\alpha}[X_{jm}]\right]$$
$$<\frac{C^R+C^{OT}}{1-\gamma} \quad \forall n\in[N] \tag{14}$$

Then, the optimal affine value function approximation for the discounted MDP will be given by

$$V_{jn}^{*} = \begin{cases} \mu_j h, & \text{n=1;} \\ \gamma V_{j,n-1}^{*}, & 2 \le n \le N-1; \\ 0, & n = N. \end{cases} \tag{15}$$

$$W_{ij}^{*} = \begin{cases} V_{jT(i)}^{*}, & \lambda_{ij} > 0; \\ 0, & \lambda_{ij} = 0. \end{cases} \tag{16}$$

$$V_0^{*} = \frac{1}{1-\gamma}\left(\sum_{\substack{(i,j)\in \\ [I]\times[J]}} \gamma^{T(i)} \mu_j h E[Y_{ij}] + hC^R\right). \tag{17}$$

Conditions (12) to (14) have a nice intuitive appeal. Condition (12) ensures that the cost of delaying a booking decision is greater than the difference between serving a patient through overtime $T(i)$ days from now versus $T(i)-1$ days from now. If this condition is not satisfied then the optimal action is simply not to book patients unless there is regular capacity available. In other words, the cost of overtime is prohibitive. If $\boldsymbol{\alpha}$ is viewed as a probability distribution on the initial state of the system then the left-hand side of Condition (13) can be seen as the present value of the expected demand over the infinite horizon plus the initial bookings (all in time units). This is required to be greater than the present value of the regular-hour capacity over the infinite horizon. Satisfying this condition ensures that there is in fact a congestion problem. If it is not satisfied then a first-come, first-served booking policy would work equally well as regular-hour capacity would be sufficient. Finally, Condition (14) ensures that, for any given day n, the present value of the expected future demand with wait time targets greater than n over the infinite horizon plus the initial bookings on days that are greater than n (all in time units) are less than the present value of the combined overtime and regular-hour capacity. In other words, it guarantees that there is sufficient overtime capacity to deal with the fluctuations in demand. This condition is particularly appealing as it provides a capacity requirement to ensure that the resulting booking policy will work well in practice. The proof of Theorem 1 is provided in the online supplement.

4.2. Stochastic model

As mentioned earlier in this paper, we only consider the end-of-day overtime and idle time costs to determine the within-day cost. Thus, in the stochastic case, we can compute $f^{AS}(\mathbf{s}, \mathbf{a})$ as

$$f^{AS}(\mathbf{s}, \mathbf{a}) = h\left[D(\mathbf{s}, \mathbf{a}) - C^R\right]^+ + u\left[C^R - D(\mathbf{s}, \mathbf{a})\right]^+, \tag{18}$$

where $D(\mathbf{s}, \mathbf{a})$ is the sum of the durations of all the appointments booked on day 1 after taking action $\mathbf{a}$ in state $\mathbf{s}$. That is $D(\mathbf{s}, \mathbf{a}) = \sum_{k=1}^{K(\mathbf{s},\mathbf{a})} d_k$, where d_k is the patient-specific duration of appointment k and $K(\mathbf{s}, \mathbf{a}) = \sum_{j=1}^{J}\left(x_{j1} + \sum_{i=1}^{I} a_{ij1}\right)$ is the total number of appointments booked on day 1. The value of d_k depends on the service class of the kth patient booked on day 1, which is known at the moment of computing $f^{AS}(\mathbf{s}, \mathbf{a})$. We assume d_k to be bounded from above and below so that $\underline{d}_k \le d_k \le \bar{d}_k \ \forall k$. In our setting, it does not actually matter the order in which patients receive care. Thus, we can assume that patients are served in increasing service class order.

We use discrete probability distributions to model appointment durations and the algorithm in Begen and Queyranne (2011) to compute the probability distribution of $D(\mathbf{s}, \mathbf{a})$ and the expected value of $f^{AS}(\mathbf{s}, \mathbf{a})$, $\forall(\mathbf{s}, \mathbf{a}) \in S \times A(\mathbf{s})$, when the service time distributions are known and independent. When the probability distributions are unknown and only samples of $D(\mathbf{s}, \mathbf{a})$ are available, then we can use the approach in Begen et al. (2012) instead. In this case, we do not require any independence assumption as we can work with correlated appointment durations as well. For sake of completeness, we will briefly describe these two algorithms below. The algorithm in Begen and Queyranne (2011) was the one used in our numerical study. It is also important to note that, in the case of known probability distributions, there are other methods for computing the probability distribution of $D(\mathbf{s}, \mathbf{a})$ (e.g., see Drew, Evans, Glen, and Leemis, 2008).

To compute the expected value of $f^{AS}(\mathbf{s}, \mathbf{a})$ when the duration of appointment k follows a known discrete probability distribution $\Pr\{d_k = m\}$, $\underline{d}_k \le m \le \bar{d}_k$, we first need to determine the probability distribution of $D(\mathbf{s}, \mathbf{a})$. If we let $D_{ab} = \sum_{k=a}^{b} d_k$, then this is equivalent to computing the probability distribution of $D_{1K(\mathbf{s},\mathbf{a})}$. To that end, we start by computing the probability distribution of $D_{11} = d_1$, then we find the probability distribution of $D_{12} = d_1 + d_2$, $D_{1k} = D_{1(k-1)} + d_k$, $\ldots$, $D_{1K(\mathbf{s},\mathbf{a})} = D_{1(K(\mathbf{s},\mathbf{a})-1)} + d_{K(\mathbf{s},\mathbf{a})}$, recursively. For example,

$$\begin{aligned} \Pr\{D_{1b} = q\} &= \Pr\{D_{1(b-1)} + d_b = q\} \\ &= \sum_{m=\underline{d}_b}^{\bar{d}_b} \Pr\{D_{1(b-1)} = q - m, d_b = m\} \\ &= \sum_{m=\underline{d}_b}^{\bar{d}_b} \Pr\{D_{1(b-1)} = q - m | d_b = m\} \Pr\{d_b = m\} \\ &= \sum_{m=\underline{d}_b}^{\bar{d}_b} \Pr\{D_{1(b-1)} = q - m\} \Pr\{d_b = m\} \quad \text{(independence)} \end{aligned}$$

We must repeat these steps for all possible values of b and q as indicated in the pseudo-code presented in the online supplement. These computations are polynomial in $K(\mathbf{s}, \mathbf{a})$ and $\bar{d}_{\max}$, where $\bar{d}_{\max} = \max\{\bar{d}_1, \bar{d}_2, \ldots, \bar{d}_{K(\mathbf{s},\mathbf{a})}\}$.

Once we have the probability distribution of $D(\mathbf{s}, \mathbf{a})$, that is $\Pr\{D_{1K(\mathbf{s},\mathbf{a})} = q\}$ $\forall q$, we can compute the expectation of $f^{AS}(\mathbf{s}, \mathbf{a})$ by using Eq. (18) as follows:

$$\sum_{q \le C^R} u(C^R - q) \Pr\{D_{1K(\mathbf{s},\mathbf{a})} = q\} + \sum_{q > C^R} h(q - C^R) \Pr\{D_{1K(\mathbf{s},\mathbf{a})} = q\} \tag{19}$$

As mentioned above, we can also use a sampling approach to compute the expected value of $f^{AS}(\mathbf{s}, \mathbf{a})$ when the service time probability distributions are unknown. Let us assume that we have access to $\mathcal{N}$ samples $\{D_t\}_{t=1}^{\mathcal{N}}$ of $D(\mathbf{s}, \mathbf{a})$, then the expected within-day cost can be computed in polynomial time as

$$\frac{1}{\mathcal{N}}\left[\sum_{t=1}^{\mathcal{N}} h(D_t - C^R)^+ + u(C^R - D_t)^+\right]. \tag{20}$$

After the expected within-day cost is computed for all possible patient-service class combinations on day 1, we can use the same ADP approach as for the deterministic case.

The implementation of the column generation algorithm and the optimization model employed to identify the approximate optimal actions for both versions of the model, the deterministic case and the stochastic case, was performed in GAMS 24.2 with CPLEX 12.6 as the solver. The algorithm used to compute the expected within-day costs was implemented in Java.

5. Numerical results

To explore the impact of incorporating stochastic service times into the MDP model described in Section 3, we performed an extensive numerical analysis under a number of scenarios. Scenarios are categorized into four problem settings. These settings were designed to reflect common observed practices in health systems as

Table 1
Arrival rate for each priority class-service class combination for Problem Setting 1 (Base Case).

Priority Class	Service Class			Total
	1	2	3	
1	1.0	1.0	1.0	3.0
2	1.0	1.0	–	2.0
3	–	–	1.0	1.0
Total	2.0	2.0	2.0	6.0

well as situations where stochastic service times might reasonably be expected to have a higher impact. Next, we compare the performance of the appointment scheduling policies suggested by the proposed approach, that are greedy with respect to the final affine value function approximations, to that of the two other policies described below:

- First Available Slot (FAS): Patients are booked as soon as possible, in increasing priority class and service class order, according to their expected service times and the available regular-hour capacity. This policy resorts to overtime only when there is no available regular-hour capacity within the booking horizon. Overtime is then booked starting with day 1 and working up to day N.
- Myopic (M): Patients are booked as soon as possible, in increasing priority class and service class order, according to their expected service times and the available regular-hour capacity. Unlike the previous policy, this policy resorts to overtime for patients of type i only when there is no available regular-hour capacity within the first n_i days of the booking horizon, where $n_i = \max\left\{n : f^{WT}(i,n) < h\right\}$. Overtime is then booked starting with day 1 and working up to day N.

These two policies are the most common representations of actual scheduling practices reported in the literature and observed in practice. The approximate optimal policies obtained from the deterministic and the stochastic versions of our model are denoted by AOPD and AOPS, respectively. The AOPD policy can be viewed as a revised version of the policy described in Patrick et al. (2008) since our model extends the MDP model in Patrick et al. (2008) to include multiple service classes.

Although the MDP model described in Section 3 is formulated in terms of the expected (DC) over the infinite horizon, the different patient scheduling policies are also compared in terms of mean daily average cost (AC), mean daily average capacity utilization (ACU), mean average time to first available appointment slot (ATFAS), mean average wait times (AWTs) and mean service levels (SLs). The ACU is computed as the average number of slots booked on day one of the booking horizon. The ATFAS is computed for each service class as the average time to the first day for which the available regular-hour capacity exceeds the corresponding expected service time. Capacity utilization and time to first available appointment slot values are computed on a daily basis. The SL is computed for each priority class as the percentage of patients booked within the corresponding wait time target. It is important to note that capacity utilization values that are different from the regular-hour capacity are associated with either overtime or idle time cost. Thus, ACU values close to the regular-hour capacity are preferred.

To compare the performance of the different policies for the first three problem settings we ran 100 replications of a simulation of the scheduling process with a total length of 2500 days with statistics collected after a warm-up period of 1000 days. We used a computer with a 3.00 Gigahertz Quad Core CPU and 16 Gigabytes of RAM for all the numerical experiments in this paper. The simulation model was implemented using GAMS Java Application Program Interface (API) with common patient arrivals and service durations.

5.1. Problem setting 1: base case

We first consider a system with a regular-hour capacity of 18 appointment slots per day. The overtime capacity is set arbitrarily large at 9 appointment slots per day to ensure that it is not a limitation. The system divides demand into three priority classes, with wait time targets of 4, 8 and 12 days, and three service classes, with expected service times of 2, 3 and 4 appointment slots. Demand from each priority class-service class combination is assumed to be Poisson with the means given in Table 1. The total expected demand is equal to the regular-hour capacity. We consider service time probability distributions that are geometric, negative binomial, Poisson, and discrete uniform. The overtime cost is 100 per appointment slot, the idle time cost is 50 per appointment slot, the postponement penalties are 20, 10 and 5 per patient for each priority class, and the discount factor is 0.99. We assume that no patient is scheduled more than 12 days in advance.

The simulation results are summarized in Table 2 for initial states generated using the FAS policy and negative binomial service time distributions. The results for initial states generated using the M policy, included in the online supplement, are similar, suggesting that the simulation outcomes could be independent of the warm-up policy. Table 3 shows, for each policy, the percent deviation from the lowest mean discounted cost across policies for the different discrete service time distributions.

Although the AOPS policy tended to outperform the AOPD policy in the simulation, the difference in the mean discounted cost

Table 2
Summary of the simulation results for Problem Setting 1 (Base Case). The bold font indicates the policy (policies) that provides (provide) the best mean performance for each metric in a statistical sense ($\alpha = 0.05$).

Metric	Priority/Service Class	Policy			
		AOPD	AOPS	FAS	M
DC [$]	–	34054 ± 667	**33802 ± 699**	58312 ± 2482	46793 ± 1422
AC [$]	–	314.46 ± 1.83	**312.13 ± 1.76**	589.87 ± 13.82	439.34 ± 7.82
ACU [slots]	–	18.05 ± 0.05	18.05 ± 0.05	**18.00 ± 0.05**	18.03 ± 0.05
ATFAS [days]	1	**1.26 ± 0.01**	1.28 ± 0.01	2.27 ± 0.02	2.19 ± 0.02
	2	**1.30 ± 0.01**	1.33 ± 0.01	2.82 ± 0.03	2.66 ± 0.03
	3	**1.35 ± 0.01**	1.38 ± 0.01	3.60 ± 0.05	3.29 ± 0.04
AWT [days]	1	**2.02 ± 0.02**	2.18 ± 0.02	8.72 ± 0.19	6.27 ± 0.11
	2	**5.51 ± 0.04**	**5.50 ± 0.04**	8.85 ± 0.18	7.02 ± 0.13
	3	9.49 ± 0.05	9.64 ± 0.05	8.68 ± 0.15	**7.45 ± 0.13**
SL [%]	1	**99.89 ± 0.03**	99.84 ± 0.04	9.64 ± 1.73	23.32 ± 1.84
	2	**100.00 ± 0.00**	**100.00 ± 0.00**	36.22 ± 2.77	67.55 ± 1.85
	3	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**

Table 3
Percent deviation from the lowest mean discounted cost (DC*) for Problem Setting 1 (Base Case) assuming different discrete service time distributions. The bold font isssndicates the policy that provides the best value for each probability distribution.

	Probability Distributions			
	Geometric	Neg. Binomial	Poisson	Uniform
AOPD	0.8%	0.7%	0.7%	1.0%
AOPS	**0.0%**	**0.0%**	**0.0%**	**0.0%**
FAS	60.3%	72.5%	79.3%	76.3%
M	32.9%	38.4%	41.2%	39.9%
DC*	42,809 ± 767	33,802 ± 699	29,878 ± 674	30,989 ± 708

Table 4
Arrival rate for each priority class-service class combination for Problem Setting 2.

Priority Class	Service Class		Total
	1	2	
1	1.0	1.0	2.0
2	–	1.0	1.0
3	1.0	–	1.0
Total	2.0	2.0	4.0

was minor (around 1%) even for service time distributions with high variance. On the other hand, both ADP policies outperform the FAS and M policies by significant margins. In particular, for the higher priority patients, the service levels and average wait times provided by the ADP policies are dramatically better as shown in Table 2. The ability of the booking policy obtained from the deterministic model to almost match the performance of the one coming from the stochastic version is perhaps surprising but also encouraging as it implies that advance schedules based on average service times are in fact reasonably effective. In the following sections, we present a number of scenarios to determine if similar results for the deterministic case can be seen in other settings.

5.2. Problem setting 2: increased difference between service classes

In the second setting, we increase the difference in the expected service time and in the standard deviation of the service time between classes. We now consider a system with regular-hour capacity of 14 appointment slots per day. The overtime capacity is 7 appointment slots per day. The system divides demand into only two service classes, with expected service times of 2 and 5 appointment slots and standard deviations of the service time of 1.01 and 3.03 appointment slots, respectively. Thus, one class has a relatively short average service time without much variance while the other has a significantly higher average service time and variance. Both service classes are assumed to have negative binomial service time distributions. Demand from each priority class-service class combination is assumed to be Poisson with means given in Table 4. All the other parameters remain the same as in the Base Case scenario.

The intent of choosing this problem setting was to pick a scenario that should more likely lead to differences between the performance of the policies coming from the two versions of our model and yet even here the two policies appear to do equally well with the stochastic version only slightly outperforming the deterministic one (difference of around 3%). Both continue to outperform the two comparator policies though by smaller margins perhaps due to the more limited flexibility in a setting with only two service classes. The simulation results are summarized in Table 5.

5.3. Problem setting 3: varying the ratio between idle time and overtime costs

In the third setting, we set the average service times to be the same across service classes but keep a significant difference in their standard deviations. We now consider a system with regular-hour capacity of 16 appointment slots per day. The overtime capacity is 8 appointment slots per day. The system divides demand into two service classes, with expected service times of 4 appointment slots each and standard deviations of 0.71 and 2.72 appointment slots, respectively. All the other parameters, including the patient arrival rates, remain the same as in the previous setting.

The simulation results are summarized in Table 6. It is worth noting that even though the differences between the comparator and the ADP policies are reduced in the last two settings, the ADP policies continue to provide much higher service levels for the high priority patients. Should the value placed on meeting wait time targets be increased (currently set quite low in comparison to the idle time and overtime costs), the difference in performance between the AOPD and AOPS policies and the comparators would undoubtedly increase accordingly.

Table 7 shows the percent difference between the mean discounted cost associated with the AOPD policy and that associated with the AOPS policy, for Problem Setting 3, assuming different idle time and overtime cost values. Here we see the greatest divergence in performance between the two ADP policies with the greatest discrepancy observed when the idle time and overtime costs are equal. Nonetheless, even in this setting, the differences are less than 3%.

Table 5
Summary of the simulation results for Problem Setting 2. The bold font indicates the policy (policies) that provides (provide) the best mean performance for each metric in a statistical sense ($\alpha = 0.05$).

Metric	Patient Class	Policy			
		AOPD	AOPS	FAS	M
DC [$]	–	34547 ± 694	**34283 ± 709**	39287 ± 1561	35654 ± 1084
AC [$]	–	331.73 ± 1.82	**327.88 ± 1.80**	387.00 ± 5.50	342.07 ± 3.27
ACU [slots]	–	**14.00 ± 0.04**	**14.00 ± 0.04**	13.98 ± 0.04	13.99 ± 0.04
ATFAS [days]	1	**1.34 ± 0.01**	1.44 ± 0.01	1.98 ± 0.02	1.90 ± 0.02
	2	**1.56 ± 0.01**	1.71 ± 0.02	3.38 ± 0.06	3.06 ± 0.04
AWT [days]	1	**2.22** ± 0.02	2.45 ± 0.02	5.77 ± 0.12	4.44 ± 0.07
	2	**5.77** ± 0.03	6.26 ± 0.03	7.54 ± 0.12	6.66 ± 0.09
	3	8.80 ± 0.05	8.89 ± 0.05	4.49 ± 0.12	**3.82** ± 0.08
SL [%]	1	**99.72 ± 0.06**	99.35 ± 0.08	41.16 ± 1.54	53.45 ± 1.22
	2	99.94 ± 0.01	**99.99 ± 0.01**	55.59 ± 1.75	70.97 ± 1.29
	3	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**

Table 6
Summary of the simulation results for Problem Setting 3. The bold font indicates the policy (policies) that provides (provide) the best mean performance for each metric in a statistical sense ($\alpha = 0.05$).

Metric	Patient Class	Policy			
		AOPD	AOPS	FAS	M
DC [$]	–	32256 ± 825	**31407 ± 828**	36751 ± 2139	32715 ± 1394
AC [$]	–	302.82 ± 1.74	**295.07 ± 1.73**	361.25 ± 8.42	310.21 ± 4.46
ACU [slots]	–	**16.01 ± 0.04**	**16.01 ± 0.04**	**15.99 ± 0.04**	**16.00 ± 0.04**
ATFAS [days]	1 & 2	**1.50 ± 0.01**	**1.50 ± 0.01**	3.67 ± 0.08	3.34 ± 0.07
AWT [days]	1	**2.29 ± 0.02**	2.51 ± 0.02	6.37 ± 0.21	4.92 ± 0.12
	2	5.83 ± 0.03	5.68 ± 0.04	6.60 ± 0.20	**5.53 ± 0.13**
	3	9.95 ± 0.05	8.81 ± 0.06	6.67 ± 0.19	**5.77 ± 0.14**
SL [%]	1	**99.83 ± 0.06**	99.76 ± 0.06	33.92 ± 2.71	45.24 ± 2.20
	2	**99.99 ± 0.01**	**99.99 ± 0.02**	66.18 ± 2.45	82.75 ± 1.37
	3	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**

Table 7
Percent difference between the mean discounted cost (DC) associated with the AOPD policy and that associated with the AOPS policy for Problem Setting 3 assuming different idle time and overtime cost values.

Overtime Cost/Idle Time Cost				
$100/$0	$100/$25	$100/$50	$100/$75	$100/$100
1.03%	1.88%	2.70%	2.57%	2.94%

Table 8
Arrival rate for each priority class-service class combination for a practical application.

Priority Class	Service Class			Total
	1	2	3	
1	0.0	1.0	1.0	2.0
2	1.0	1.0	1.0	3.0
3	1.0	2.0	1.0	4.0
4	6.0	8.0	2.0	16.0
Total	8.0	12.0	5.0	25.0

5.4. Problem setting 4: a practical application

Finally, we consider a practical application based on data provided by a medium-size clinic in Canada. The clinic divides demand into four priority classes with wait time targets of 4, 8, 12 and 24 days and three service classes with expected service times of 4, 6 and 8 appointment slots. Each appointment slot is 5 minutes in length. Demand from each priority class-service class combination is assumed to be Poisson with arrival rates given in Table 8. Service times follow the empirical discrete probability distributions shown in Fig. 1. The regular-hour capacity is set at 144 appointment slots, which is equal to the average daily demand and equivalent to two identical medical resources operating six hours a day. The overtime capacity is 24 appointment slots or one extra hour per medical resource. The overtime cost is 100 per appointment slot, the idle time cost is 50 per appointment slot, the postponements penalties are 25, 20, 15 and 10 per patient, and the discount factor is 0.99. We assume that no patient is scheduled more than 24 days in advance.

The corresponding simulation results are summarized in Table 9. To speed up the performance evaluation process for this problem setting, each policy was simulated for 1000 days with statistics collected for each of 30 simulation runs after a warm-up period of 250 days. The total time needed to simulate each policy was a couple of seconds for the FAS and M policies, close to 30 hours for the AOPD policy, and about 12 hours for the AOPS policy. The AOPS again shows a slight improvement in the mean discounted cost over the AOPD (around 3.5%) and both continue to outperform the comparator policies in terms of this metric as well as others.

The total time required to identify the approximate optimal policy was around 7 hours for the deterministic case and 17 hours for

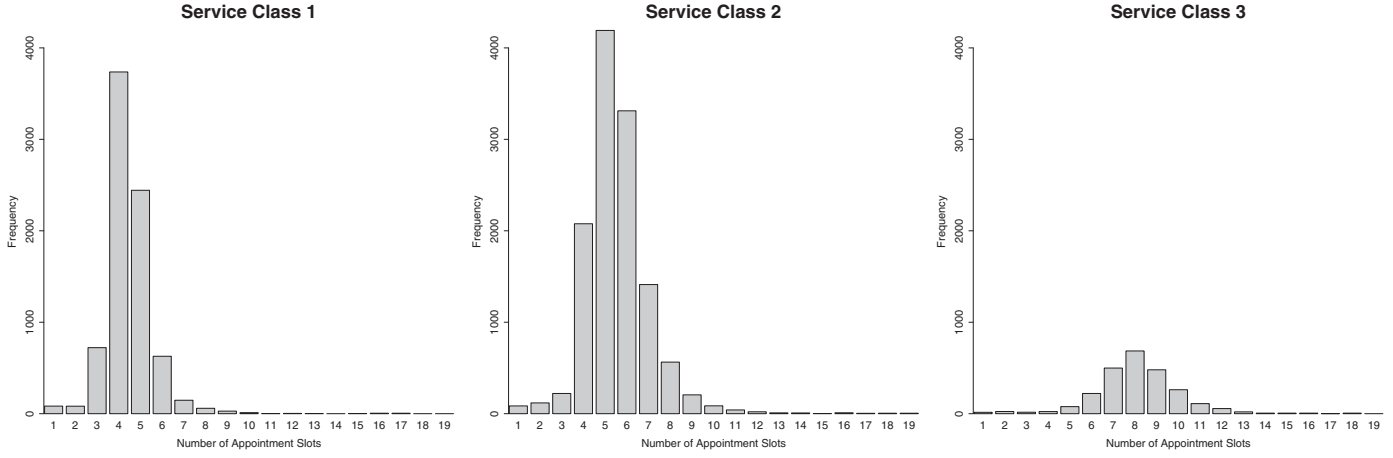

Fig. 1. Service time histograms for a practical setting based on data provided by a medium-size clinic in Canada. The clinic divides demand into three service classes with mean services times of 20, 30 and 40 minutes, respectively. Each appointment slot is 5 minutes in length.

Table 9
Summary of the simulation results for Problem Setting 4 (A Practical Application). The bold font indicates the policy (policies) that provides (provide) the best mean performance for each metric in a statistical sense (α = 0.05).

Metric	Patient Class	Policy			
		AOPD	AOPS	FAS	M
DC [$]	–	65424 ± 2944	**63210 ± 2982**	140154 ± 15821	105573 ± 8156
AC [$]	–	536.56 ± 9.26	**512.74 ± 9.17**	1618.69 ± 183.20	700.96 ± 29.14
ACU [slots]	–	142.84 ± 0.32	142.76 ± 0.32	142.04 ± 0.15	**144.01 ± 0.34**
ATFAS [days]	1	**1.13 ± 0.02**	1.14 ± 0.02	2.26 ± 0.04	2.24 ± 0.07
	2	**1.15 ± 0.02**	1.16 ± 0.03	2.88 ± 0.06	2.73 ± 0.10
	3	**1.16 ± 0.03**	1.17 ± 0.03	3.89 ± 0.12	3.37 ± 0.13
AWT [days]	1	**1.30 ± 0.04**	1.39 ± 0.05	15.11 ± 1.26	5.76 ± 0.30
	2	**2.31 ± 0.15**	2.45 ± 0.16	15.17 ± 1.25	6.93 ± 0.39
	3	**4.95 ± 0.31**	5.05 ± 0.31	15.36 ± 1.25	7.18 ± 0.39
	4	18.93 ± 0.48	18.74 ± 0.52	15.70 ± 1.25	**7.55 ± 0.39**
SL [%]	1	**99.41 ± 0.17**	99.36 ± 0.18	1.35 ± 1.93	24.79 ± 6.22
	2	**99.81 ± 0.11**	**99.82 ± 0.11**	4.85 ± 5.98	84.41 ± 2.34
	3	**99.96 ± 0.04**	**99.98 ± 0.03**	21.43 ± 10.91	98.33 ± 0.67
	4	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**

the stochastic case. These execution times are satisfactory, considering that the coefficients defining the approximate optimal policies only need to be redetermined when there is a significant change in the problem parameters. It is also important to note that solving the integer programming model used to identify the approximate optimal actions, which is done on a daily basis, takes less than one or two seconds.

6. Discussion and conclusion

Through this research we have begun to bridge the advance and the appointment scheduling problems – two problems that have been treated separately in past research despite their significant interdependencies – by incorporating stochastic service times into the advance scheduling problem and, in this way, taking into account not only the waiting time until the day of service but also the idle time and overtime of the medical resource on the day of service.

This paper provides three contributions to the literature. First, we describe a model that schedules patients with varying capacity requirements and urgency levels to a single resource and that can be adapted for either deterministic or stochastic service times. Second, we provide solutions and insights for multiple problem settings, including a practical application, assuming deterministic and stochastic appointment durations with extensive numerical analyses. For the deterministic case, we present analytical results and prove the form of the optimal affine approximation under reasonable conditions. The conditions provide limits on the available capacity in order for the resulting policies to work effectively. For the stochastic version, we provide numerical results that suggest that the deterministic version of the model works reasonably well (within 4% difference in our numerical experiments) even in cases where the service time distributions demonstrate significant variability and when service classes vary significantly in both the mean service time and the standard deviation of the service time. The latter provides good evidence that in fact the advance scheduling policies based on deterministic service times cannot be easily improved upon by incorporating stochastic service times. This is, in our opinion, a very important finding for those studying advance scheduling as it allows us to avoid significant complications to the models we build and to continue to combine the advance scheduling and the appointment scheduling problems without worrying that we are losing significantly. Third, but not least, we provide an extensive literature review of advance and appointment scheduling.

It is, however, possible that the success of the policies obtained from the deterministic model in (almost) matching the performance of those obtained using the stochastic version may be attributable to either the use of an affine approximation architecture (possibly curtailing the advantage of incorporating stochastic service times) or the fact that we concentrated our efforts on a setting where within-day waiting/idle time was of little consequence. Thus, two further avenues of research are the implementation of a non-linear approximation architecture similar to Sauré et al. (2015) as well as the incorporation of the sequencing of patients into the MDP model in order to address scenarios where the within-day waiting/idle time is relevant.

Acknowledgments

This work was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) [Grants RGPIN-2015-03911 and RGPIN-2018-05225].

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2019.06.040

References

Astaraky, D., & Patrick, J. (2015). A simulation based approximate dynamic programming appraoch to multi-class, multi-resource surgical scheduling. *European Journal of Operational Research, 245*, 309–319.
Bailey, N. (1952). A study of queues and appointment systems in hospital out-patient departments with special reference to waiting times. *Journal of the Royal Statistical Society B, 14*, 185–199.
Begen, M., Levi, R., & Queyranne, M. (2012). Technical note - a sampling-based approach to appointment scheduling. *Operations Research, 60*, 675–681.
Begen, M., & Queyranne, M. (2011). Appointment scheduling with discrete random durations. *Mathematics Operations Research, 36*, 240–257.
Bosch, P., Vanden, P., Dietz, D., & Simeoni, J. (1999). Scheduling customer arrivals to a stochastic service system. *Naval Research Logistics, 46*, 217–229.
Cayirli, T., Kum, K., & Quek, S. (2012). A universal appointment rule in the presence of no-shows and walk-ins. *Production and Operations Management, 21*, 682–697.
Chakraborty, S., Muthuraman, K., & Lawley, M. (2010). Sequential clinical scheduling with patient no-shows and general service time distributions. *IIE Transactions, 42*, 354–366.
Chen, R., & Robinson, L. (2014). Sequencing and scheduling appointments with potential call-in patients. *Production and Operations Management, 23*, 1522–1538.
Denton, B., & Gupta, D. (2003). A sequential bounding approach for optimal appointment scheduling. *IIE Transactions, 35*, 1003–1016.

Denton, B., Viapiano, J., & Vogl, A. (2007). Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Science, 10*, 13–24.

Dobson, G., Hasija, S., & Pinker, E. (2011). Reserving capacity for urgent patients in primary care. *Production and Operations Management, 20*, 456–473.

Drew, J., Evans, D., Glen, A., & Leemis, L. (2008). *Computational probability: algorithms and applications in the mathematical sciences*. Springer Science & Business Media.

Erdelyi, A., & Topaloglu, H. (2009). Computing protection level policies for dynamic capacity allocation problems by using stochastic approximation methods. *IIE Transactions, 41*, 498–510.

Erdogan, S., & Denton, B. (2013). Dynamic appointment scheduling of a stochastic server with uncertain demand. *INFORMS Journal Computer, 25*, 116–132.

Feldman, J., Liu, N., Topaloglu, H., & Ziya, S. (2014). Appointment scheduling under patient preference and no-show behaviour. *Operations Research, 62*, 794–811.

Ge, D., Wan, Z., & Zhang, J. (2013). A note on appointment scheduling with piecewise linear costs functions. *Mathematics Operations Research, 39*, 1244–1251.

Gocgun, Y., & Ghate, A. (2012). Lagrangian relaxation and constraint generation for allocation and advanced scheduling. *Computer Operations Research, 39*, 2323–2336.

Gocgun, Y., & Puterman, M. (2014). Dynamic scheduling with due dates and time windows: an application to chemotherapy patient appointment booking. *Health Care Management Science, 17*, 60–76.

Green, L., & Savin, S. (2008). Reducing delays for medical appointments: A queuning approach. *Operations Research, 56*, 1526–1538.

Green, L., Savin, S., & Wang, B. (2006). Managing patient service in a diagnostic medical facility. *Operations Research, 54*, 11–25.

Guda, H., Dawande, M., Janakiraman, G., & Jung, K. (2016). Optimal policy for a stochastic scheduling problem with applications to surgical scheduling. *Production and Operations Management, 25*, 1194–1202.

Gupta, D. (2007). Surgical suites operations management. *Production and Operations Management, 16*, 689–700.

Gupta, D., & Denton, B. (2008). Appointment scheduling in health care: challenges and opportunities. *IIE Transactions, 40*, 800–819.

Huang, Y., & Zuniga, P. (2012). Dynamic overbooking scheduling system to improve patient access. *Journal of the Operational Research Society, 63*, 810–820.

Kaandoorp, G., & Koole, G. (2007). Optimal outpatient appointment scheduling. *Health Care Management Science, 10*, 217–229.

Klassen, K., & Rholeder, T. (1996). Scheduling outpatient appointments in a dynamic environment. *Journal Operations Management, 14*, 83–101.

Klassen, K., & Yoogalingam, R. (2008). An assessment of the interruption level of doctors in outpatient appointment scheduling. *Operations Management Research, 1*, 95–102.

Klassen, K., & Yoogalingam, R. (2009). Improving performance in outpatient appointment services with a simulation optimization approach. *Production and Operations Management, 18*, 447–458.

Klassen, K., & Yoogalingam, R. (2014). Strategies for appointment policy design with patient unpunctuality. *Decision Science, 45*, 881–911.

Kong, Q., Lee, C., Teo, C., & Zheng, Z. (2013). Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operational Research, 61*, 711–726.

LaGanga, L., & Lawrence, S. (2012). Appointment overbooking in healht care clinics to improve patient service and clinic performance. *Production and Operations Management, 21*, 874–888.

Liu, N. (2016). Optimal choice for appointment scheduling window under patient no-show behaviour. *M&SOM-Manufacturing Services Operations, 25*, 128–142.

Liu, N., Ziya, S., & Kulkarni, V. (2009). Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *M&SOM-Manufacturing Services Operations, 12*, 347–364.

Luo, J., Kulkarni, V., & Ziya, S. (2012). Appointment scheduling under patient no-shows and service interruptions. *M&SOM-Manufacturing Services Operations, 14*, 670–684.

Ma, X., Sauré, A., Puterman, M., Taylor, M., & Tyldesley, S. (2016). Capacity planning and appointment scheduling for new patient oncology consults. *Health Care Management Science, 19*, 347–361.

Mak, H., Rong, Y., & Zhang, J. (2014). Appointment scheduling with limited distributional information. *Management Science, 61*, 316–334.

Mancilla, C., & Storer, R. (2009). *Stochastic sequencing and scheduling of an operating room* Ph.D Thesis.

Mancilla, C., & Storer, R. (2012). A sample average approximation approach to stochastic appointment sequencing and scheduling. *IIE Transactions, 35*, 655–670.

Mittal, S., Schulz, A. S., & Stiller, S. (2014). Robust appointment scheduling. In K. Jansen, J. D. P. Rolim, N. R. Devanur, & C. Moore (Eds.), *Approximation, randomization, and combinatorial optimization. algorithms and techniques: 28* (pp. 356–370). Leibniz International Proceedings in Informatics (LIPIcs).

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

Patrick, J. (2012). A markov decision model for optimal outpatient scheduing. *Health Care Management Science, 15*, 91–102.

Patrick, J., Puterman, M., & Queyranne, M. (2008). Dynamic multi-priority patient scheduling for a diagnostic resource. *Operations Research, 56*, 1507–1525.

Robinson, L., & Chen, R. (2003). Scheduling doctors' appointments: optimal and empirically based heuristic policies. *IIE Transactions, 35*, 295–307.

Robinson, L., & Chen, R. (2010). Traditional and open-access appointment scheduling policies: The effects of patient no-shows. *M&SOM-Manufacturing Services Operations, 12*, 330–346.

Samorani, M., & LaGanga, L. (2015). Outpatient appointment scheduling given individual day-dependent no-show prediction. *European Journal of Operational Research, 240*, 245–257.

Santibáñez, P., Begen, M., & Atkins, D. (2007). Surgical block scheduling in a system of hospitals: An application to resource and wait list management in a British Columbia health authority. *Health Care Management Science, 10*, 269–282.

Santibáñez, P., Chow, V., French, J., Puterman, M., & Tyldesley, S. (2009). Reducing patient wait times and improving resource utilization at British Columbia Cancer Agency's ambulatory care unit through simulation. *Health Care Management Science, 12*, 392–407.

Sauré, A., Patrick, J., & Puterman, M. (2015). Simulation-based approximate policy iteration with generalized logistic functions. *INFORMS Journal Computer, 27*(3), 579–595.

Sauré, A., Patrick, J., Tyldesley, S., & Puterman, M. (2012). Dynamic multi-appointment patient scheduling for radiation therapy. *European Journal of Operational Research, 223*, 573–584.

Schütz, H., & Kolisch, R. (2012). Approximate dynamic programming for capacity allocation in the service industry. *European Journal of Operational Research, 218*, 239–250.

Schütz, H., & Kolisch, R. (2013). Capacity allocation for demand of different customer-product-combinations with cancellations, no-shows, and overbooking when there is sequential delivery of service. *Annals of Operations Research, 206*, 401–423.

Tang, J., Yan, C., & Cao, P. (2014). Appointment scheduling algorithm considering routine and urgent patients. *Expert Systems with Applications, 41*, 4529–4541.

Truong, V. (2015). Optimal advance scheduling. *Management Science, 61*, 1584–1597.

Tsai, P., & Teng, G. (2014). A stochastic appointment scheduling system on multiple resources with dynamice call-in sequence and patient no-shows for an outpatient clinic. *European Journal of Operational Research, 239*, 427–436.

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

Wang, W., & Gupta, D. (2011). Adaptive appointment systems with patient preferences. *M&SOM-Manufacturing Services Operations, 13*, 373–389.

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

White, D., Froehle, C., & Klassen, K. (2011). The effect of integrated scheduling and capacity policies on clinical efficiency. *Production and Operations Management, 20*, 442–455.

Zacharias, C., & Pindeo, M. (2014). Appointment scheduing with no-shows and overbooking. *Production and Operations Management, 23*, 788–801.

Zeng, B., Turkcan, A., Lin, J., & Lawley, M. (2009). Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. *Annals of Operations Research, 178*, 121–144.