Health Care Manag Sci (2012) 15:91–102
DOI 10.1007/s10729-011-9185-4

A Markov decision model for determining optimal outpatient scheduling

Jonathan Patrick

Received: 7 March 2011 / Accepted: 27 October 2011 / Published online: 17 November 2011

Abstract Managing an efficient outpatient clinic can often be complicated by significant no-show rates and escalating appointment lead times. One method that has been proposed for avoiding the wasted capacity due to no-shows is called open or advanced access. The essence of open access is "do today's demand today". We develop a Markov Decision Process (MDP) model that demonstrates that a short booking window does significantly better than open access. We analyze a number of scenarios that explore the trade-off between patient-related measures (lead times) and physician- or system-related measures (revenue, overtime and idle time). Through simulation, we demonstrate that, over a wide variety of potential scenarios and clinics, the MDP policy does as well or better than open access in terms of minimizing costs (or maximizing profits) as well as providing more consistent throughput.

Keywords Clinic scheduling · Open acess · Markov decision processes · Dynamic programming · Simulation

1 An introduction to outpatient scheduling

In recent years, a form of scheduling called advanced access or open access (OA) has been touted as the preferred booking policy for outpatient clinics. The mantra is "do today's demand today". The implicit assumption is that, due to a combination of the cost of no shows and the value the clinic places on same-day access, the benefit associated with being able to smooth demand over a period of time is insufficient to warrant delaying appointments. This begs the question as to how prevalent no-shows need to be and/or how strongly the clinic needs to value same-day access in order to outweigh the benefit of being able to smooth out demand. It is such trade-offs that this research seeks to explore by developing an MDP model to determine the optimal booking policy for given wait-time-dependent no-show rates and by exploring potential cost structures for various clinics.

The detriments of a strict OA policy have begun to appear in the literature as some papers have suggested that a short booking window is more appropriate and that there are certain conditions that need to be satisfied in order for OA to succeed. This work falls in line with that literature and proposes a different outpatient scheduling policy that does in fact use a short booking window combined with overbooking to mitigate the impact of no-shows.

We assume that we are dealing with a clinic that has a fixed capacity to see C clients per day but with additional overtime available if necessary. Clients call for appointments before the start of a service period and can either be booked into that period or else booked into the next available appointment slot in a future service period. Appointments are subject to a no-show probability that increases the further in advance the appointment is booked. Decisions regarding how many requests to serve today and how many to book in advance have to be made prior to knowing if any of the previously booked appointments will fail to show up.

The rest of the paper proceeds as follows. In Section 2, a review of the literature on open access

J. Patrick (✉)
Telfer School of Management, University of Ottawa,
55 Laurier Avenue, Ottawa, ON K1N 6N5, Canada
e-mail: patrick@telfer.uottawa.ca

and clinic scheduling is presented and how our approach differs from what has been done previously is demonstrated. In Section 3, the MDP model for the clinic scheduling problem is described as well as some of the assumptions inherent in the model and two complications that seek to overcome some of those assumptions. In Section 4, the results of testing the MDP policy against OA in a simulation are presented and in Section 5, the form of the MDP policy is described. Finally, Section 6 presents the conclusions from this research.

2 Literature review

Murray and Tantau [10] first proposed the idea of OA as a solution to the high levels of no-shows often present in outpatient clinics. They provide a case study of a successful implementation of open access in the US. Kopach et al. [6] provide a simulation study to determine the impact of various clinic characteristics on the successful implementation of OA. Their simulation allows for a multi-doctor clinic and patient specific no-show probabilities. They allow for limited overbooking in some of the scenarios analyzed in the simulation. Robinson and Chen [12] demonstrate that OA most often outperforms a traditional booking system (with a fixed number of patients per day) but is itself outperformed by a "same-or-next-day" scheduling policy thus demonstrating that there is some advantage to flexibility in scheduling. They use a weighted sum of physician idle time, direct patient waiting time and overtime as the basis for their performance measure.

More generally, Cayirli and Veral [2] as well as Gupta and Denton [4] provide overviews of appointment scheduling research. Kim and Giachetti [5] present a stochastic model that seeks to address the question of how many patients to book in advance given a known distribution for walk-in demand and no-shows. The no-show probability is dependent solely on the advance booking policy (ABP). LaGanga and Lawrence [7] also demonstrate the advantage of overbooking in a stochastic model that computes the expected benefit from overbooking. Neither paper however provides a dynamic scheduling policy where decisions can be dependent on the state of the booking slate. More recently, Muthuraman and Lawley [11] model a clinic with a fixed number of "slots" and where unfinished demand "overflows" into the next slot. They provide a myopic scheduling policy that maximizes revenue but that does not take into account future demand. Patients are added to the booking slate until the expected profit ceases to increase at which point demand is rejected. Zeng et al. [13] build on the model of Muthuraman and Lawley by allowing the no-show probabilities to vary between patients. They propose two sequential scheduling algorithms—one that is myopic in a similar fashion to Muthuraman and another that attempts to include future demand in the decision process through a forecasting model.

Finally, Liu et al. [9] provide a dynamic programming approach that takes into account future demand and allows for state dependent cancelation and no-show rates. Their model tracks the number of booked appointments that are i days out and that were booked j days in advance. This creates an intractable MDP due to the size of the state space. They therefore resort to a one-step policy iteration to improve the policy and demonstrate (in line with the work of Robinson and Chen) that a two day booking window outperforms same day booking and that through their one-step policy iteration they can improve on any initial policy.

This research adds to the above literature by providing a dynamic program that can be solved to optimality, allows for wait time dependent no-show rates and provides clear evidence that a short booking window with overbooking can provide greater benefit to a clinic than a strict adherence to OA. The cost structure in the model is sufficiently flexible to provide a dynamic scheduling policy for a variety of clinics and the simulation results demonstrate the distinct advantage of the scheduling policy derived from the MDP for all the clinic types tested.

3 MDP model for clinic scheduling

Scheduling decisions are assumed to be made before each service period but after today's demand has arrived. Ideally, the probability of a given client keeping his/her appointment would depend on how far in advance the client has been scheduled. However, the MDP model would quickly become intractable if the actual waiting time of each client was incorporated into the state space. Instead, the assumption that clients who are booked in advance are given the first available slot on a first-come-first-served basis allows the size of the queue to act as a proxy for wait time.

The "advanced booking policy" (ABP) is defined as the number of clients who can be booked in advance into each future day. The schedule is not constrained to book only C clients per day as the potential for no-shows may in fact make overbooking desirable and the possibility of same day bookings may make underbooking desirable.

The state is represented by a vector, $\mathbf{s} = (w, x, y)$, with w representing the current ABP, x representing the number of previously booked appointments and y representing new demand. The ABP is restricted to the set $\{1, ..., M\}$ with $M > C$ in order to allow for limited overbooking. Thus, M represents the maximum number of patients who can be booked in advance into each future day. The total number of advanced bookings are also restricted so that $x \in \{0, ..., N\}$ for some finite $N > 0$. Thus, N is the maximum size of the queue. $x \wedge y$ and $x \vee y$ are used to represent the minimum and maximum of x and y respectively.

Each decision epoch, two decisions are required of the manager. S/he must decide whether to change the current ABP as well as how much of the new demand to service today. Let $\mathbf{a} = (a, b)$ represent the combined action where a represents the new ABP and b represents how much new demand to serve today. In this version of the model, any change in the ABP takes place by the next day.

Actions are constrained in four ways. First, bookings cannot exceed demand so $b \leq y$. Second, the number of patients in the advance booking slate cannot exceed the imposed limit so $b \geq [x - x \wedge w + y - N]^+$ (number of today's demand treated today must be at least equal to the current number of bookings − minus today's bookings + demand − limit on the number of advanced bookings). Thirdly, the new ABP is no more than 1 patient per day different from the previous ABP and does not deviate outside the imposed limits so $1 \vee (w - 1) \leq a \leq (w + 1) \wedge M$. The restriction to only a one patient per day change in the ABP is simply to limit the problem to a reasonable size and due to the doubtful benefit of making radical changes. Finally, all actions are positive and integer so $(a, b) \in Z^+ \times Z^+$.

The transition to the next state can be described as:

$$\mathbf{s} = (w, x, y) \rightarrow \mathbf{s}' = (a, x - x \wedge w + y - b, D) \qquad (1)$$

where D is a random variable representing new demand. The new booking slate consists of the previous booking slate minus today's slate $(x \wedge w)$ plus any of yesterday's demand that was not served immediately $(y - b)$.

There are a number of potential rewards/costs associated with booking patients to a clinic. There may be some *revenue* associated with each patient serviced, f^R, a cost to servicing a patient through *overtime*, f^{OT}, a cost associated with *idle time*, f^{IT}, and a cost associated with patient appointment *lead times*, f^{LT}—that is the number of days between the request for service and the date of service. Service times are assumed to be deterministic so that overtime and idle time costs are simply a function of the number of patients who show up to their appointment and capacity. Finally, in order to avoid changes in the ABP that don't result in a major benefit, a cost for *switching* the ABP, f^S, can be imposed. The expected reward function can be written as:

$$\begin{aligned} r(\mathbf{s}, \mathbf{a}) = {} & f^R(p_s(w, x)(w \wedge x) + p_{sd} b) \\ & - f^{OT} \sum_{(i,j) \in W \times B} (i + j - C)^+ Pr(AB = i) Pr(SD = j) \\ & - f^{IT} \sum_{(i,j) \in W \times B} (C - i - j)^+ Pr(AB = i) Pr(SD = j) \\ & - f^{LT} x + f^S (w - a)^+ \end{aligned} \qquad (2)$$

where $p_s(w, x)$ is the probability that an advance booking shows up (henceforth called the show probability) given there are x clients in the queue and w is the current ABP, p_{sd} is the show probability for a same-day booking, W is the set of possible values for $w \wedge x$ and B is the set of possible values for b. AB is a random variable representing the number of clients booked in advance for an appointment today who show for their appointment and SD is a random variable representing the number of clients given a same day appointment for today who show for their appointment. Obviously, it is possible to "shut off" any element of the reward function to better reflect the objectives of the actual clinic. The costs/rewards can be viewed as a trade-off between physician- or system-related measures (revenue, overtime and idle time) and patient-related measures (lead time).

The choice for the probability of a client arriving for his/her appointment is clearly crucial. Kopach et al. [6] provide a logistic regression (with an $R^2 = 0.8071$) for an outpatient clinic that includes age, session (whether the client is booked into the morning or afternoon), weather and insurance type as predictors of the show probability. They recognize that appointment lead time is also crucial but had no data to substantiate the actual impact. They therefore adjust the show probability by an exponential function based on lead times determined by "expert" opinion. Gallucci et al. [3] demonstrate, for a community mental health center, that appointment lead time does in fact have a significant impact, with that impact stabilizing for a lead time greater than 7 days. Of 5901 patients in the data set, 31% failed to show up for their appointment. The show probability for same-day appointments was 88% and dropped to 77% for next day appointments. Appointments booked 7 days out had a show probability of 58% while a 13 day delay only dropped the show probability to 56%. Finally, Lee et al. [8] use a logistic regression

model to demonstrate that age, race, appointment lead time, previously failed appointments, provision of a cell phone number and distance from the hospital were all significant factors in predicting show probabilities. However, their appointment lead time analysis only compares lead times less than 7 days to lead times greater than 21 days.

While it would be ideal to incorporate all the factors mentioned above, such detail would make the model intractable. Thus, this research focuses only on appointment lead time as the factor that is most clearly impacted by the scheduling policy. To that end, the following show probability is used:

$$p_s(w, x) = \max\left(1 - \frac{\beta_1 + \beta_2 * \log(LT + 1)}{100}, \beta_3\right) \quad (3)$$

with LT representing appointment lead time. The log function easily models the Gallucci results (with $\beta_1 = 12$, $\beta_2 = 36.54$, $\beta_3 \cong .5$) that impose a diminishing impact of an extra day's wait the further out a client is booked. β_3 represents a lower bound on the show probability of any client. The addition of one to LT inside the log function insures that the show probability for a same day booking ($LT = 0$) is $p_{sd} = 1 - \beta_1/100$. For advance bookings,

$$LT = \max(1, \lfloor x/w \rfloor) \quad (4)$$

where $\lfloor x/w \rfloor$ is the largest integer smaller than x/w. The maximum reflects the fact that even if the number of bookings is less than the ABP, $\lfloor x/w \rfloor < 1$, any client booked in advance still waits a day. Equation 4 allows the ABP to impact on the relationship between appointment lead times and queue size (if you allow more clients to be booked per day then the same x reflects less of a wait) and therefore better reflects an equal show probability for patients booked an equal number of days out.

The above description provides the five elements (decision epochs, state space, action space, costs and transition probabilities) for a Markov Decision Process (MDP) formulation. An infinite horizon setting was chosen as clinics do not have a closing date and thus any end to the scheduling horizon is necessarily arbitrary. This forces the model to ignore any "day-of-the-week" effect but avoids any arbitrary terminal rewards.

The discounted version of the MDP is used to reflect the fact that upfront costs are always more expensive than future costs. However, a discount factor of $\gamma = 0.99$ is used so as not to unduly bias the model against the future. The standard optimality equations for a discounted infinite horizon MDP of this initial model can therefore be written as:

$$v(\mathbf{s}) = \max_{\mathbf{a} \in A(\mathbf{s})} \left\{ r(\mathbf{s}, \mathbf{a}) + \gamma \sum_{k \in \mathbf{D}} p_d(k) \times v\left(a, x - x \wedge w + y - b, k\right) \right\} \quad (5)$$

where $p_d(k)$ is the probability that $D = k$.

3.1 Assumptions to the model

As with any modeling exercise, there are a number of assumptions that help keep the model tractable. First, client service times are assumed to be deterministic and we do not allow for advance notice of cancelations. Second, as mentioned in Section 3, the model assumes that all demand for a given day has arrived before any scheduling decision needs to be made. In reality scheduling decisions have to be made as each demand request arrives. However, batch arrivals allows decisions to be made in discrete time and the resulting policy, as will be demonstrated in Section 5, is easily translatable into the more realistic setting where the scheduling decision occurs at the time of each demand request. Third, the utilization of the size of the queue as a proxy for wait time when calculating the probability that a client shows up for an appointment means that the no-show rate of a client will depend on the size of the queue at the time of service when in reality it depends on the size of the queue at the time of booking. While this may mean that an individual's no-show rate may not be accurate, it will still penalize a system that books well in advance to the same degree as a more realistic (but intractable) model. It is also an issue only if the length of the queue varies dramatically whereas the policy derived from the MDP model shows no such variation. Fourth, for the sake of simplicity, the issue of patient availability is ignored as each client is assumed to take the offered slot. To do otherwise would require the model to track actual appointment lead times which would again lead to an intractable model. Finally, in the model described in Section 3, changes in the ABP occur before the next decision epoch. This may mean that some appointments need to be postponed (if the ABP is lowered) or that some clients may need to have their appointment day advanced (if the ABP is increased). A more complex model that includes a lag time in the enactment of any change in the ABP is developed in Section 3.3 in order to avoid this problem. However, a small cost for any change to the ABP forestalls this issue as it tends to result in the optimal policy settling on a single ABP. Such a cost is imposed

in the simulation results presented in Section 4 in order to insure a simpler policy that is easily implementable. In all simulation runs in which the model was run with and without a cost for changing the ABP the difference in the objective between the more complex policy and the simpler one was not statistically significant. The next two subsections provide two complications to the model that allow for more realism.

3.2 Complication 1: two stream demand

To add an additional element of realism, the model can easily be adapted to allow for a stochastic stream of clients who must be booked in advance. The transitions would then become:

$$\mathbf{s} = (w, x, y) \rightarrow \mathbf{s}' = (a, x - x \wedge w + y - b + A, D) \tag{6}$$

where A is a random variable representing the demand requiring advanced booking. The optimality equations are now:

$$v(\mathbf{s}) = \max_{\mathbf{a} \in A(\mathbf{s})} \left\{ r(\mathbf{s}, \mathbf{a}) + \gamma \sum_{i \in \mathbf{D},\, j \in \mathbf{A}} p_d(i)\, p_a(j) \times v\left(a, x - x \wedge w + y - b + j, i\right) \right\} \quad \forall s \in S \tag{7}$$

where $\mathbf{A}$ is the set of possible advanced booking requests and $p_a(j)$ is the probability of getting j advanced booking requests in a day. Since advanced booking requests (for a clinic running OA) generally arise as follow-up visits from today's appointments, the number of such visits is assumed to be unknown until after today's booking decisions are made.

An additional issue that this complication to the model raises is that it is now impossible to insure that the booking slate does not exceed N clients simply by restricting the action set. To avoid this issue, a client's request for an appointment is rejected at a high cost should accepting it cause the booking slate to exceed N. Making this cost sufficiently high insures that such extreme measures are never required.

3.3 Complication 2: lag time in ABP changes

As mentioned earlier, it is somewhat unrealistic to expect a change in the ABP to be implemented immediately. However, the model can be adjusted to impose a lag time on any change in the ABP. Let the new state, $\mathbf{s} = (w, x, y, z)$, where w, x and y are as before and z represents the lag time remaining until the next triggered ABP change. The lag can be set to $\lfloor x/w \rfloor \vee 1$ to insure that the lag is equal to the number of days that are already fully booked and thus no previously booked appointments need to be canceled and no new bookings jump the queue. If $z < 0$ then the decision is to reduce the ABP by one and if $z > 0$ then the decision is to increase the ABP by one. More drastic changes are not considered.

The same actions are available to the manager with the same restrictions. The evolution of the system depends on the current lag time and can be described by the following set of transitions:

$$\mathbf{s} = (w, x, y, z) \rightarrow \mathbf{s}' = \begin{cases} (w, x - x \wedge w + y - b\,, D, l + 1), & \text{If } l < -1; \\ (w, x - x \wedge w + y - b\,, D, l - 1), & \text{If } l > 1; \\ (a, x - x \wedge w + y - b\,, D, 0), & \text{If } l = -1,\ \lfloor x/w \rfloor \leq 1 \text{ or } l = 1; \\ (w, x - x \wedge w + y - b\,, D, -\lfloor x/w \rfloor - 1), & \text{If } l = 0,\ \lfloor x/w \rfloor > 1 \text{ and } a < w; \\ (w, x - x \wedge w + y - b\,, D, \lfloor x/w \rfloor - 1), & \text{If } l = 0,\ \lfloor x/w \rfloor > 1 \text{ and } a > w; \\ (w, x - x \wedge w + y - b\,, D, 0), & \text{If } l = 0 \text{ and } a = w. \end{cases} \tag{8}$$

Here, D is again a random variable representing new demand. Notice that once a decision to change the ABP is made, no new change can be made until the day when that decision is implemented. If tomorrow's slate is not yet full ($\lfloor x/w \rfloor \leq 1$) then the ABP change is implemented by tomorrow as in the original model.

Finally, the optimality equations are

$$v(\mathbf{s}) = \max_{\mathbf{a} \in A(\mathbf{s})} \left\{ r(\mathbf{s}, \mathbf{a}) + \gamma \sum_{k \in \mathbf{D}} p_d(k) v\left(\mathbf{s}'\right) \right\} \tag{9}$$

where $\mathbf{D}$ represents the set of all possible demand, p_d is the probability distribution for new demand and $\mathbf{s}'$ is determined by the transitions given in Eq. 8 with $D = k$.

4 Simulation results

In this section, the simulation results are presented that compare the MDP policy to OA in a variety of

scenarios that are chosen to explore the trade-off between the system-related costs/rewards (revenue, overtime and idle time) and the patient-related ones (lead time). We consider three clinic types for the system-related costs/rewards. The first clinic ignores idle time and sets revenue to twice the cost of overtime: $f^R = 20$, $f^{OT} = 10$ and $f^{IT} = 0$. The second clinic adds an additional cost for idle time: $f^R = 20$, $f^{OT} = 10$ and $f^{IT} = 5$. The third clinic does not receive remuneration for each patient seen but seeks to maximize the utilization of the available capacity to meet all demand: $f^R = 0$, $f^{OT} = 10$ and $f^{IT} = 5$. In all cases, the cost of changing the ABP is set equal to the cost of one overtime slot. The relative values of the cost parameters is somewhat arbitrary. We have followed common practice of valuing idle time (when it is considered important) at half the cost of overtime and have chosen to set revenue at twice the value of overtime. The rationale for these choices is to represent both clinics that are privately run or where the physician is paid on a fee-for-service basis (so that revenue is important) as well as clinics that are publicly funded and therefore where revenue is not a factor.

For each of these clinics, lead time costs are set at 0, 1 and 5 leading to a total of 9 clinic types differentiated by the relative weight they place on each cost/reward. If the cost of a day's lead time is set equal to the overtime cost then the optimal policy would clearly be OA. Thus, lead time costs of 1 and 5 seem reasonable options for demonstrating the impact of increasing the value of providing same-day service. For each clinic type, we consider 6 scenarios. The base case for each clinic assumes a Poisson arrival rate and sets average demand, λ, equal to capacity, $C = 10$. Such a small clinic size is chosen in order to restrict the amount of computation time for running numerous scenarios. Larger clinic sizes can obviously be solved. The Gallucci show-rate described in Section 3 is used to determine show rates. Admittedly, a mental health clinic may not be the most standard clinic upon which to base the no-show rates. However, the intent is to demonstrate the benefit of a short booking window in spite of the presence of no-shows thus using a show-rate that is perhaps overly pessimistic is reasonable. Any clinic seeking to implement the MDP policy would have to determine their own show-rate first.

In addition to the base case, five other scenarios are run for each clinic type. The average demand rate is varied above and below the capacity, a stream of demand is introduced that must be booked in advance (arriving with rate μ), the show rate is given a steeper rate of decline and finally same day bookings are given a show probability of one. The actual parameter values for each of these scenarios are summarized in Table 1. No results are provided for the lag time version of the model as the cost (equal to one overtime slot) associated with switching the ABP results in a policy that does not vary the ABP thus making the lag time irrelevant. Running the model with no switching cost leads to a much more complicated policy but provides a statistically insignificant improvement over the policy with a fixed ABP.

For each scenario, 50 replications, each 5000 days in length, are run both for the MDP policy and OA. The two policies are compared on the basis of throughput, overtime, idle time, appointment lead times and profit/cost. Statistics are collected after the first 500 days.

One technical challenge was the calculation of the expected number of clients scheduled today who show for their appointment when the show rate is dependent on the actual lead time (as in the simulation). The challenge is that each client may have been booked a different number of days in advance and thus may have different show probabilities. This problem is avoided in the MDP model by using queue size as a proxy for wait times but in the simulation actual wait times for each client are used. Each day, the probability distribution of the number of clients who show up for their appointment based on the scheduled clients for that day needs to be calculated. This is accomplished based on the algorithm outlined in [1]. Once that probability is

Table 1 Scenarios analyzed

Base case ($\lambda = C$)	$\lambda = 10, \mu = 0, \beta_1 = 12, \beta_2 = 36.54, \beta_3 = 50$
Reduced demand ($\lambda = C - 2$)	$\lambda = 8, \mu = 0, \beta_1 = 12, \beta_2 = 36.54, \beta_3 = 50$
Increased demand ($\lambda = C + 2$)	$\lambda = 12, \mu = 0, \beta_1 = 12, \beta_2 = 36.54, \beta_3 = 50$
Advanced bookings	$\lambda = 7, \mu = 3, \beta_1 = 12, \beta_2 = 36.54, \beta_3 = 50$
Show rate with steep decline	$\lambda = 10, \mu = 0,$ $\beta_1 = 12$ (same day rate unchanged), $\beta_2 = 80$ (steeper decline), $\beta_3 = 20$ (levels out at lower rate)
Show rate with same day = 100%	Same as base case except show probability is 1 for same day appointments

determined, it is an easy step to calculate the expected rewards/costs each day and thus to obtain the optimal action based on the argmax of Eq. 5.

Table 2 provides a comparison of OA and the MDP policy for each of the scenarios for the clinics with system-related parameters equal to, $f^R = 20$, $f^{OT} = 10$, $f^{IT} = 0$ and appointment lead time costs of 0,1 and 5. Throughput is given as a percentage of demand, while overtime and idle time are given as a percentage of capacity. The results of the simulation demonstrate that the two policies are essentially equivalent in terms of average daily profit with the MDP policy slightly outperforming OA in most scenarios. It is not surprising that OA performs relatively well for a clinic where revenue is present since any no-show is detrimental in that it reduces revenue even if overtime is required. Thus, a policy that minimizes the number of no-shows (which is the aim of OA) will clearly have an advantage. The surprise perhaps is that one can do as well by reducing overtime and idle time through a more sophisticated scheduling policy even if that policy implies that there will be less throughput. The crucial additional advantage of the MDP policy is that the day-to-day workload is significantly less variable (see Fig. 1) and the peak load is significantly reduced (from 20 to 15 in the base case with zero lead time costs for instance).

Table 3 provides the comparison of OA and the MDP policy for each of the scenarios for the clinics with system-related parameters equal to $f^R = 20$, $F^{OT} = 10$, $f^{IT} = 5$ and appointment lead time costs of 0, 1 and 5. These clinics demonstrate relatively similar results to those clinics that ignored idle time as revenue is still the primary driver. Thus the two policies are essentially equivalent in terms of average daily profit with the MDP policy outperforming OA by slightly higher margins than for the previous clinics. Again, there is a significant reduction in the peak workload and a smoothing of the variation in daily throughput (see Fig. 2).

Finally, Table 4 provides the comparison of OA and the MDP policy for each of the scenarios for the clinics with system-related parameters equal to $f^R = 0$, $F^{OT} = 10$, $f^{IT} = 5$ and appointment lead time costs of 0, 1 and 5. For such clinics, the MDP policy

Table 2 Simulation results for a clinic with $f^R = 20$, $f^{OT} = 10$, $f^{IT} = 0$

Scenario	Policy	Lead time cost	Average daily			Max lead time (days)	Profit	Increase in profit for MDP
			TH	OT	IT			
Increased demand	OA		88.0	15.7	10.2	0	195.40	
	MDP	0	86.7	11.4	7.4	1	196.69	0.7%*
		1	87.5	13.7	8.7	1	195.96	0.1%*
		5	88.0	15.7	10.2	0	195.40	0.0%
Show rate with same day = 100%	OA		100.0	12.5	12.5	0	187.50	
	MDP	0	98.2	8.3	10.1	1	188.06	0.3%*
		1	99.1	10.1	11.1	1	187.58	0.0%
		5	100.0	12.5	12.5	0	187.5	0.0%
Base case	OA		88.0	6.9	18.9	0	169.08	
	MDP	0	86.8	3.1	16.3	1	170.51	0.8%*
		1	88.0	6.9	18.9	0	169.08	0.0%
		5	88.0	6.9	18.9	0	169.08	0.0%
Show rate with steep decline	OA		88.0	6.9	18.9	0	169.08	
	MDP	0	87.2	4.9	17.8	1	169.39	0.2%*
		1	87.8	6.4	18.6	1	169.11	0.0%
		5	88.0	6.9	18.9	0	169.08	0.0%
Advanced bookings	OA		84.6	5.6	21.0	1	160.66	
	MDP	0	83.7	2.6	18.9	2	164.83	2.6%*
		1	84.0	3.4	19.4	2	161.16	0.3%*
		5	84.6	5.6	21.0	1	160.66	0.0%
Decreased demand	OA		88.0	2.1	31.7	0	138.73	
	MDP	0	87.5	0.5	30.5	1	139.45	0.5%*
		1	87.6	0.7	30.7	1	139.12	0.3%*
		5	88.0	1.9	31.5	1	138.79	0.0%

Throughput (TH) is given as a percentage of demand and overtime (OT) and idle time (IT) as a percentage of regular hour capacity
Starred lines represent scenarios where the MDP policy improved significantly ($\alpha = 0.05$) upon the OA policy in terms of cost based on a matched pairs t-test

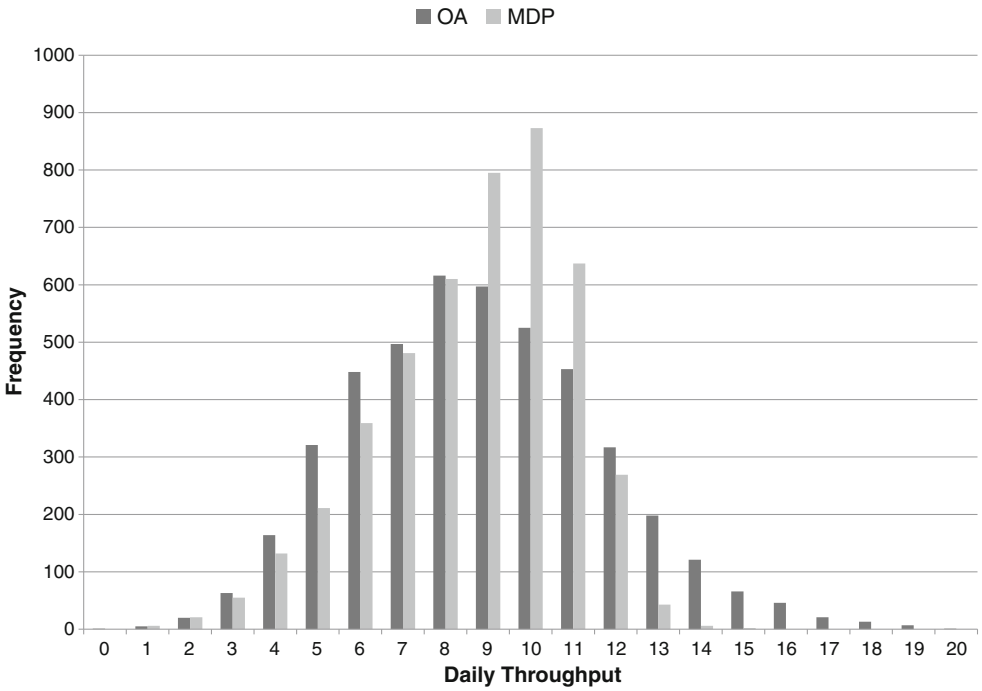

Fig. 1 Daily throughput for a clinic with $f^R = 20$, $f^{OT} = 10$, $f^{IT} = 0$, $f^{LT} = 0$

Table 3 Simulation results for a clinic with $f^R = 20$, $f^{OT} = 10$, $f^{IT} = 5$

Scenario	Policy	Lead time cost	Average daily			Max lead time (days)	Profit	Increase in profit for MDP
			TH	OT	IT			
Increased demand	OA		88.0	15.7	10.2	0	190.33	
	MDP	0	86.2	10.3	6.8	1	193.21	1.5%*
		1	87.1	12.3	7.8	1	191.70	0.7%*
		5	88.0	15.7	10.2	0	190.33	0.0%
Show rate with same day = 100%	OA		100.0	12.5	12.5	0	181.25	
	MDP	0	97.0	6.0	9.0	1	183.44	1.2%*
		1	98.1	8.1	10.0	1	182.38	0.6%*
		5	100.0	12.5	12.5	0	181.25	0.0%
Base case	OA		88.0	6.9	18.9	0	159.63	
	MDP	0	86.6	2.6	16.0	2	162.49	1.8%*
		1	86.9	3.5	16.5	1	161.29	1.0%*
		5	87.8	6.2	18.3	1	159.61	0.0%
Show rate with steep decline	OA		88.0	6.9	18.9	0	159.63	
	MDP	0	86.6	4.0	17.4	1	160.46	0.5%*
		1	87.0	4.7	17.7	1	159.63	0.3%*
		5	88.0	6.9	18.9	0	159.63	0.0%
Advanced bookings	OA		84.6	5.6	21.0	0	153.03	
	MDP	0	83.3	1.9	18.6	3	155.45	1.6%*
		1	83.8	2.8	19.0	2	154.69	1.1%*
		5	84.4	4.8	20.4	1	153.05	0.0%
Decreased demand	OA		88.0	2.1	31.7	0	122.90	
	MDP	0	87.5	0.5	30.5	1	124.21	1.1%*
		1	87.6	0.6	30.5	1	123.95	0.9%*
		5	87.8	1.3	31.0	1	123.16	0.2%*

Throughput (TH) is given as a percentage of demand and overtime (OT) and idle time (IT) as a percentage of regular hour capacity
Starred lines represent scenarios where the MDP policy improved significantly ($\alpha = 0.05$) upon the OA policy in terms of cost based on a matched pairs t-test

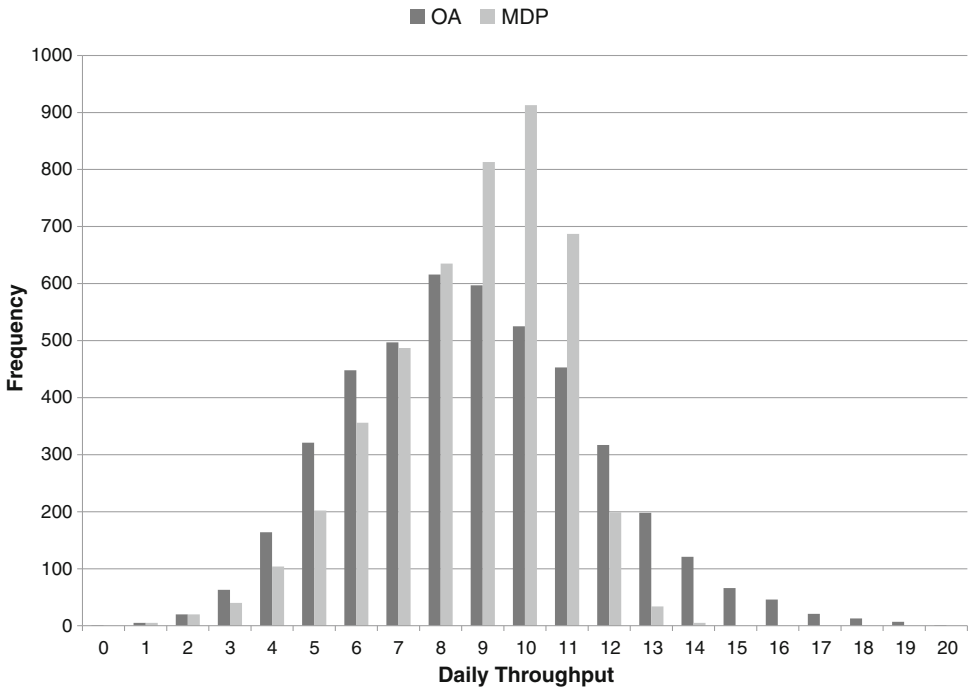

Fig. 2 Daily throughput for a clinic with $f^R = 20$, $f^{OT} = 10$, $f^{IT} = 5$, $f^{LT} = 0$

Table 4 Simulation results for a clinic with $f^R = 0$, $f^{OT} = 10$, $f^{IT} = 5$

Scenario	Lead time cost	Policy	Average daily			Max lead time (days)	Cost	Decrease in cost for MDP
			TH	OT	IT			
Show rate with same day = 100%	OA		100.0	12.5	12.5	0	18.75	
	MDP	0	91.6	1.0	9.4	3	5.67	69.8%*
		1	93.4	1.8	8.4	2	8.92	52.4%
		5	97.1	6.1	9.0	1	16.93	9.7%*
Increased demand	OA		88.0	15.7	10.1	0	20.81	
	MDP	0	78.0	2.9	9.2	4	7.50	64.0%*
		1	83.9	6.9	6.2	2	14.29	31.3%*
		5	87.8	14.7	9.4	2	20.67	0.7%*
Base case	OA		88.0	6.9	18.9	0	16.34	
	MDP	0	84.1	0.7	16.5	3	8.92	45.4%*
		1	85.9	1.7	15.8	2	11.45	29.9%*
		5	87.4	4.5	17.1	1	15.64	4.2%*
Show rate with steep decline	OA		88.0	6.9	18.9	0	16.34	
	MDP	0	82.7	0.8	18.1	2	9.81	40.0%*
		1	84.3	1.5	17.2	2	11.60	29.0%*
		5	86.6	4.0	17.4	1	15.51	5.1%*
Advanced bookings	OA		84.6	5.6	21.0	0	16.16	
	MDP	0	81.5	0.6	19.1	3	10.11	37.4%*
		1	82.9	1.4	18.5	2	12.03	25.5%*
		5	84.2	3.7	19.6	1	13.89	14.0%*
Decreased demand	OA		88.0	2.1	31.7	0	17.89	
	MDP	0	86.9	0.0	30.5	2	15.28	14.6%*
		1	87.2	0.2	30.4	1	15.93	11.0%*
		5	87.6	0.8	30.7	1	17.32	3.2%*

Throughput (TH) is given as a percentage of demand and overtime (OT) and idle time (IT) as a percentage of regular hour capacity
Starred lines represent scenarios where the MDP policy improved significantly ($\alpha = 0.05$) upon the OA policy in terms of cost based on a matched pairs t-test

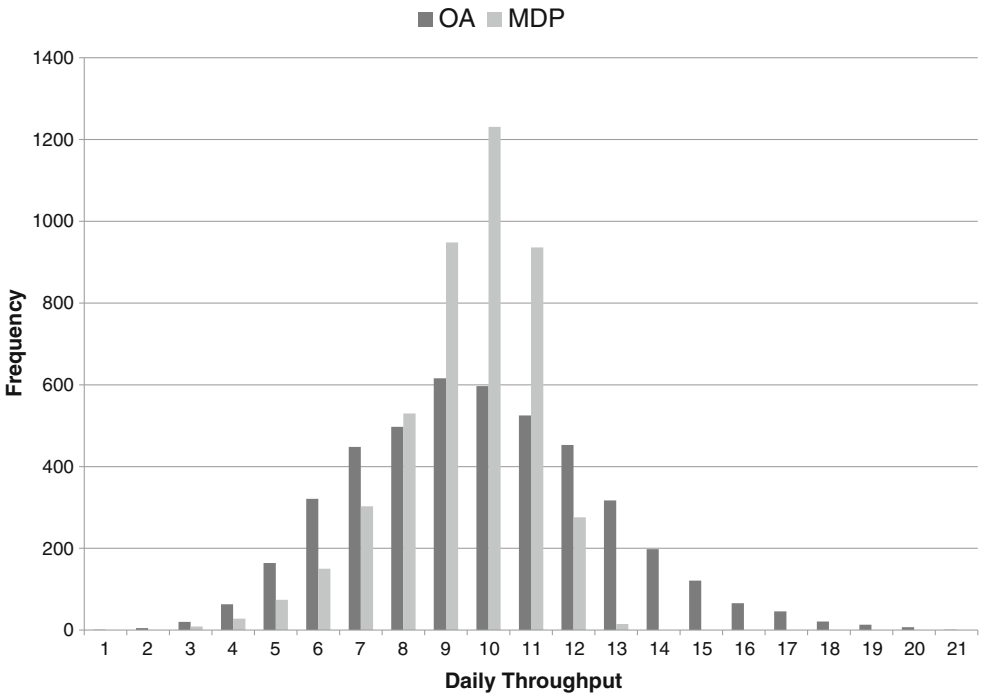

Fig. 3 Daily throughput for a clinic with $f^R = 0$, $f^{OT} = 10$, $f^{IT} = 5$, $f^{LT} = 0$

significantly reduces average daily cost compared to OA in all instances. Figure 3 again demonstrates the significant reduction in variation in daily throughput and in the peak workload achieved by the MDP policy as opposed to OA.

The longest appointment lead times in any of the scenarios examined was four days demonstrating that while the MDP does increase patient lead times it does not do so excessively. It is also worth noting that, for the three clinic types differentiated by the system-related parameter values, the scenarios where the MDP policy performed the best were the scenario with same day show probability equal to 100% (a scenario that one might have thought would benefit OA the most) and the scenario where demand exceeded capacity. This might seem counter-intuitive as one would expect a scenario with higher demand to lead to higher levels of overtime. However, the presence of a significant no-show rate means that the *effective* demand can, to some extent, be controlled by the scheduling policy that is implemented as a larger booking window means higher no-show rates. Finally, for all clinic types, the MDP policy approaches an OA policy the more capacity exceeds demand or, not surprisingly, as the lead time cost increases.

In the simulation and in the MDP, a linear function for overtime costs was used. A more realistic piecewise linear function that incorporates increasing overtime costs with a larger overtime load would increase the benefit of the MDP policy over OA. Additionally, the maximum number of arrivals in a day was restricted to twice the average demand in ordered to keep the state space relatively small. If higher daily demand levels were permitted an even greater improvement in performance would have been achieved by the MDP policy compared to OA.

5 The form of the MDP policy

The MDP policy without a cost associated with changing the ABP is not easily described for any of the clinics as it depends on a number of factors—the current ABP, queue size and demand as well as today's slate of patients and today's expected throughput. However, once a cost is imposed for changing the ABP, the policy becomes quite simple and depends on two factors—the size of the queue (those already booked plus those waiting to be booked) and today's current slate.

For all of the clinics and scenarios analyzed, the MDP policy is defined by a series of thresholds. The policy will act initially like OA booking any new request into today's slate. However, once today's slate reaches a certain threshold (the value being dependent on the scenario and particularly on the lead time cost), it will begin to defer any further requests to the next available appointment slot. The policy will revert back to booking a new request into today's slate if

Fig. 4 Example of the form of the optimal policy from the MDP

	Number of Clients Booked For Today											
DEMAND	0	1	2	3	4	5	6	7	8	9	10	11
0	1	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0	0
2	0	0	1	0	0	0	0	0	0	0	0	0
3	0	0	0	1	0	0	0	0	0	0	0	0
4	0	0	0	0	1	0	0	0	0	0	0	0
5	0	0	0	0	0	1	0	0	0	0	0	0
6	0	0	0	0	0	0	1	0	0	0	0	0
7	0	0	0	0	0	0	0	1	0	0	0	0
8	0	0	0	0	0	0	0	0	1	0	0	0
9	0	0	0	0	0	0	0	0	0	1	0	0
10	0	0	0	0	0	0	0	0	0	0	1	0
11	0	0	0	0	0	0	0	0	0	0	1	0
12	0	0	0	0	0	0	0	0	0	0	1	0
13	0	0	0	0	0	0	0	0	0	0	1	0
14	0	0	0	0	0	0	0	0	0	0	1	0
15	0	0	0	0	0	0	0	0	0	0	1	0
16	0	0	0	0	0	0	0	0	0	0	1	0
17	0	0	0	0	0	0	0	0	0	0	1	0
18	0	0	0	0	0	0	0	0	0	0	1	0
19	0	0	0	0	0	0	0	0	0	0	0	1
20	0	0	0	0	0	0	0	0	0	0	0	1

the queue size reaches its own threshold. Depending on the scenario, the policy may then simply revert to OA and book any additional demand into today or it may book only one additional client today and then defer again until another threshold triggers a second overbook. If the lead time cost is set to zero then the first threshold that triggers the initial decision to begin booking requests into future days is equal to capacity. As the lead time cost is increased, that first threshold is increased as well.

Figure 4 gives a demonstration of the policy for the base case for a clinic with cost parameters equal to $f^R = 0$, $f^{OT} = 10$, $f^{IT} = 5$, $f^{LT} = 0$ and for a day with no advance bookings at the outset. The MDP policy acts like OA until capacity is reached at which point it begins to delay demand to the next day. However, if the number of requests exceeds 18 then it books an additional client into today's schedule before once again delaying any further demand to the next available appointment slot. If there are advance bookings already on the slate then the policy looks very similar to that in Fig. 4 but with the threshold shifted down by the number already in the queue. If the lead time cost is increased to one then the policy acts like OA until today's bookings reach eleven (capacity plus one) and then defers demand to the next available appointment slot. If the lead time cost is further increased to five then the initial decision to stop booking any new requests into today is delayed until today's bookings reach 12 before a series of thresholds on the queue size triggers additional overbooks.

Recall that the MDP formulation assumes that all demand for the day was collected before any scheduling decision is made. However, the form of the policy makes it a simple matter to translate the MDP policy into the more realistic setting where demand is scheduled as it arrives. As a new request for an appointment arrives, the booking clerk checks the MDP policy to determine whether that request should be booked today or delayed based on the look-up tables from the MDP as if this were the last request of the day. Thus, in the example described in the previous paragraph, if the 18th request of the day arrives, it will be booked into the next available appointment slot but if a 19th request arrives, it will be booked into today's slate even though today's slate is already full. The negative aspect of the MDP policy is that now a client who called later may in fact receive an earlier appointment.

6 Conclusion

The argument behind OA is two-fold. Either (1) the presence of no-shows has such a debilitating effect on clinic performance that it is worthwhile to minimize such occurrences as much as possible by "doing today's demand, today" or else (2) the advantage of providing same-day access is worth the cost. What this paper

demonstrates is that the impact of no-shows can be easily mitigated without resorting to OA. A policy that instead uses a short booking window to smooth out demand can reduce both idle time and overtime substantially and thus improve resource utilization and reduce costs. The robustness with which the MDP policy maintains its advantage over OA across all the scenarios demonstrates that this superiority is not a function of the particular parameters used. Even for those clinics where revenue is dominant and thus higher throughput is most advantageous, the MDP demonstrates distinct advantages over OA in that the variation in the day-to-day workload and the peak work load are significantly reduced without sacrificing profit.

If a high enough cost is associated with appointment lead times then clearly OA will eventually become optimal. However, the cost (in resource efficiency) incurred by the clinic for insuring same-day access as opposed to using a short booking window can be quite high. This is not to say that there are not clinics where OA can be readily implemented but that the trade-offs need to be clearly understood. The popularity of OA within the medical community suggests that perhaps they have not. This research contributes to the growing evidence provided by the operations research community that, depending on a clinic's priorities, open access may in fact be sub-optimal. The benefit of this research is to provide an alternative scheduling policy that can significantly improve resource utilization while only marginally increasing client appointment lead times.

Acknowledgements This work was funded in part by a research grant from National Science and Engineering Research Council (NSERC). I would like to thank Martin Puterman from the University of British Columbia his help in revising this paper.

References

1. Butler K, Stephens M (1993) Distribution of a sum of binomial random variables. Technical report No 457 for the Office of Naval Research. Available at http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA266969&Location=U2&doc=GetTRDoc%.pdf
2. Cayirli T, Veral E (2003) Outpatient scheduling in health care: a review of literature. Prod Oper Manag 12:519–549
3. Gallucci G, Swartz W, Hackerman F (2005) Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. Psychiatr Serv 56:344–346
4. Gupta D, Wang L (2008) Revenue management for a primary-care clinic in the presence of patient choice. Oper Res 56:576–592
5. Kim S, Giachetti R (2006) A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. IEEE Trans Syst Man Cybern Part A Syst Humans 36:1211–1219
6. Kopach R et al (2007) Effects of clinical characteristics on successfull open access scheduling. Health Care Manage Sci 10:111–124
7. LaGanga L, Lawrence S (2007) Clinic overbooking to improve patient access and increase provider productivity. Decis Sci 38:251–276
8. Lee V, Earnest A, Chen M, Krishnan B (2005) Predictors of failed attendances in a multi-specialty outpatient centre using electronic databases. BMC Health Serv Res 5:1–8
9. Liu N, Ziya S, Kulkarni V (2009) Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. Manuf Serv Oper Manag 12:347–364
10. Murray M, Tantau C (1999) Redefining open access to primary care. Manag Care Q 7:45–51
11. Muthuraman K, Lawley M (2008) A stochastic overbooking model for outpatient clnical scheduling with no-shows. IIE Trans 40:820–837
12. Robinson L, Chen R (2009) Traditional and open-access appointment scheduling policies: The effects of patient no-shows. Manuf Serv Oper Manag 12:330–346
13. Zeng B, Turkcan A, Lin J, Lawley M (2009) Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. Ann Oper Res 178:121–144