Intra-day dynamic rescheduling under patient no-shows

Aditya Shetty
Simon School of Business
Univeristy of Rochester
aditya.shetty@simon.rochester.edu

Harry Groenevelt
Simon School of Business
Univeristy of Rochester
groenevelt@simon.rochester.edu

Vera Tilson
Simon School of Business
Univeristy of Rochester
vera.tilson@simon.rochester.edu

Abstract

Existing work on appointment scheduling assumes that appointment times cannot be updated once they have been assigned. However, advances in communication technology and the adoption of online (as opposed to in-person) appointments make it possible for appointments to be flexible. In this paper, we describe an intra-day dynamic rescheduling model that adjusts upcoming appointments based on observed no-shows. We formulate the problem as a Markov Decision Process in order to compute the optimal pre-day schedule and the optimal policy to update the schedule for every scenario of no-shows. We also propose an alternative formulation based on the idea of 'atomic' actions that can solve for the optimal policy more efficiently. Based on a numerical study, we estimate that using intra-day dynamic rescheduling can lead to a 5-7% decrease in expected cost when compared to static scheduling.

1. Introduction

The problem of appointment scheduling has been the subject of multiple works in the past, starting with the seminal work of [1]. It involves assigning appointment times to a set of patients to be served by a doctor over the course of a day. Existing work on appointment scheduling has largely focused on the creation of pre-day schedules that can minimize clinic idling and patient waiting costs ([2], [3]). However, advances in communication technology and the adoption of online (as opposed to in-person) appointments have lead to the possibility that appointment times given before the start of the day, can be updated as the day progresses. In this paper, we describe an intra-day dynamic rescheduling model that takes into account the state of the system during the day, to make adjustments to subsequent appointments. This work is intended to be a first step in understanding the benefits of rescheduling. There are further challenges that must be overcome for a rescheduling policy to be implemented in practice. We highlight some of these challenges and describe how they can be addressed by building on our model. Even though our primary inspiration in developing this model is its potential use in outpatient clinics, the idea of intra-day dynamic rescheduling can be applied to appointment scheduling in professional services outside healthcare like lawyer's offices, accountant offices, barber shops etc.

The central challenge in appointment scheduling is to mitigate the effect of uncertainty on patients and providers. There are various sources of uncertainty that a clinic may need to plan for. Some examples include: Non-deterministic service duration, tardiness of patients and doctors, walk-in patients and emergencies. In our model, we restrict our attention to the uncertainty that arises from patient no-shows. A common approach to hedge against the effect of no-shows is to overbook. Overbooking reduces the probability that the server will remain idle, but it comes at the cost of increased waiting time if multiple patients show up in the same time slot. We compute a rescheduling policy that allows for the possibility of postponing appointments. Such a rescheduling policy reduces cost in two ways: (1) it reduces the expected amount of time patients must wait at the clinic, and (2) it allows for a greater level of overbooking without significantly increasing waiting times. However, these advantages must be evaluated against the inconvenience cost of patients' appointment times being changed. Therefore, in addition to the cost of waiting, overtime and idling considered by earlier work, we introduce a cost of rescheduling. We assume rescheduling cost increases if the update is made closer to the appointment time and if the appointment is delayed longer.

We formulate the problem as a Markov Decision Process. The state of the system is defined by the time slot, the number of patients waiting and the current schedule. At the beginning of each time slot, the appointment times of all subsequent patients may be postponed. In the case of static scheduling, scheduling

URI: https://hdl.handle.net/10125/79784
978-0-9981331-5-7

a patient earlier generally reduces expected cost for the provider while it increases expected patient costs, but in the case of dynamic scheduling, there is a secondary effect to consider: an earlier appointment gives the clinic information about a no-show earlier in the day. Due to this "information effect", optimal dynamic schedules do not follow some seemingly intuitive characteristics. For example: (1) it may be optimal to not postpone a patient's appointment even if the patient is guaranteed to have to wait, and (2) it may be optimal to have some slack in the schedule, i.e., no patient is scheduled in a certain slot even though patients are scheduled after that slot.

The optimal rescheduling decision from a given state must not only consider the no-show behavior of future patients, but also their appointment times. Having to account for these upcoming appointment times greatly increases the size of the state space and therefore, the computational complexity of the problem. We propose an alternative, but equivalent formulation based on the idea of 'atomic' schedule updates that can solve for the optimal policy more efficiently. We show that the problem of finding the optimal schedule update at any point reduces to finding a sequence of optimal atomic updates. Solving for the optimal policy using the alternative formulation is an order of magnitude faster than using the conventional formulation. However, the run times for both formulations increase exponentially with the number of periods. To solve larger problem instances may require the use of heuristics. We briefly describe some potential heuristics in section 6.

Using parameter estimates from empirical work in the exiting literature, we conduct a numerical study to estimate the advantage of rescheduling over a static scheduling policy. We find that using intra-day dynamic rescheduling can lead to a 5-7% reduction in expected cost. Further, we find that this reduction in expected cost is largely driven by reducing cost in some of the worst case realizations.

The rest of this paper is structured as follows: Section 2 gives an overview of related work, Section 3 lays out the setup of our model and expresses the optimization problem as a Markov Decision Process (MDP), Section 4 provides a more efficient formulation of the base MDP from Section 3. It also highlights the structural differences between optimal static and dynamic schedules. Section 5 presents a numerical study that quantifies the advantage and robustness of rescheduling. Finally, we conclude our findings in Section 6 and discuss potential future directions for exploration.

2. Literature

Outpatient appointment scheduling was first studied by [1]. They proposed a simple heuristic of scheduling two patients at the beginning of the day with subsequent patients being spaced by their expected service times. Since then, the problem has been studied in a lot of different settings. [4], [5] provide an extensive overview of this literature. In this paper, we restrict our attention to the uncertainty resulting from no-shows. Multiple studies have shown that no-shows have a significant impact on the operating costs of healthcare systems ([6], [7], [8], [9] and [10]).

Various approaches have been studied to reduce no-show rates. Some examples of these approaches are: using open-access scheduling ([11], [12], [13], [14]), controlled release of capacity ([15]) and sending out appointment reminders ([16], [17], [18]). Another stream of work attempts to hedge against the uncertainty of no-shows through overbooking. This could mean allocating less time to each patient ([19]) or scheduling multiple patients in the same time slot ([2], [3]).

The term "dynamic scheduling" has been applied to different types of scheduling problems. One is a form of online scheduling where appointment requests arrive sequentially and decisions on appointments or capacity must be made before subsequent requests are known ([20], [21], [22], [23]). Another form of dynamic scheduling arises in operating room planning where the major source of intra-day stochasticity is the randomness in the demand for inpatient and emergent procedures. [24] studied the problem of dynamically sequencing inpatients, outpatients and emergent patients that are awaiting service at the beginning of each slot. They also suggest heuristics to compute efficient schedules for outpatients. Their model was extended to the case of multiple servers by [25] while [26] considered add-on patients and different objectives. A closely related work is one by [27] that considers the problem of assigning appointment times to customers during the day, after an earlier 'anchor' patient has started service. However, they do not look at the problem of creating a schedule at the start of the day and the resulting inconvenience cost of updates.

Existing work on appointment scheduling under no-shows assumes that schedules are non-adaptive or static, in the sense that once a patient has been assigned an appointment time, that time cannot be updated. The focus of our work is the dynamic *rescheduling* of patients. To the best of our knowledge, this is the first study that models and evaluates an intra-day dynamic version of the appointment scheduling problem that allows for appointment times to be updated throughout

the day as information is revealed.

3. Model

We model a service where N patients are to be served by a single server. The service duration for all patients is the same and deterministic. We define this time duration to be one time slot or period. The patients can be scheduled over T consecutive time slots. Patients are punctual. They show up exactly at their appointment time. However, there is a probability, p, that a patient will be a no-show. The provider learns whether or not a patient is a no-show at the patient's appointment time.

Before the start of the day, each patient must be given an appointment time. We will call these appointment times, the pre-day schedule. As the day progresses, depending on the state of the system, upcoming appointments can be moved to a later time on the same day. An appointment may be moved multiple times during the day. Patients are immediately notified of changes to their appointment time, and will show up based on the appointment time assigned to them in the latest notification they receive.

The schedule must assign an appointment time to each patient, but because patients are homogeneous (equal no-show probability and service duration) and all appointment times must correspond to the start of a time slot, the schedule can be represented as a vector of length T, where the j^{th} element is the number of patients scheduled to arrive at the beginning of time slot j. At the start of each time slot, the state of the system can be completely represented by the current time slot, t, the number of patients awaiting service, q, and the number of patients currently scheduled to arrive at the start of each slot, $\mathbf{x}$ [1]. The minimum expected cost of being in a given state can be written in the form of a Bellman equation having three components: (1) the operating costs that will be incurred in the current slot, (2) the minimum operating cost from the next period onward if the schedule is optimally updated and (3) the cost incurred at the end of the planning horizon, T. The remainder of this section will elaborate on each of these components in order to lead up to the Bellman equation (Formulation (DR)).

Any schedule update made at the beginning of slot t can only affect the cost incurred in slots $t+1$ onward. Therefore, the cost incurred during slot t is only dependent on the current state $(t, q, \mathbf{x})$ and is independent of the updated schedule. If the clinic has no patients to serve in the current slot, it incurs a cost of idling, γ_I. However, if there are no more patients scheduled, the clinic can close and does not incur idling cost. Therefore, the cost of idling is given by: $\gamma_I I(q = 0 \wedge x_{t+1} + ... + x_T > 0)$, where $I(.)$ is an indicator function. On the other hand, multiple patients may be in the system at the start of the current slot. In such cases, patients will be served in the order in which they arrive and will need to wait until all patients with earlier appointment times have been served. Waiting in the current period incurs a cost of γ_W per slot per patient which leads to a total waiting cost of $\gamma_W(q-1)^+$ for the current slot.

When the schedule is updated, the appointment times of patients may be postponed by one or more slots. We model the cost incurred by a patient when a schedule update at time t postpones the patient's appointment from time j to j' as:

$$U_t(j, j') = \gamma_W \left(u_{j-t} + u_{j-t+1} + ... + u_{j'-t-1}\right) \quad (1)$$

where $\mathbf{u}$ is a vector whose k^{th} element is the update cost incurred by a patient when their appointment time is k slots away and is postponed by one slot. Estimating $\mathbf{u}$ is an empirical problem but we conjecture that $1 > u_1 \geq u_2 \geq ... \geq u_{T-1} \geq 0$. This implies that $U_t(j, j')$ satisfies three intuitive properties: (1) $0 \leq U_t(j, j') \leq (j' - j)\gamma_W$, i.e., updating an appointment is inconvenient but the cost of an update never exceeds the cost of waiting. (2) $U_t(j, j')$ is increasing and convex in t, i.e., the longer an update is postponed, the larger the update cost becomes, and this happens at an increasing rate. (3) $U_t(j, j')$ is increasing and concave in j', i.e., the update cost increases with the length of the update, but at a decreasing rate. Further, we define a transformation, $L : \mathbb{Z}_+^T \rightarrow \mathbb{Z}_+^N$, which transforms the schedule vector $\mathbf{x}$, representing the number of patients scheduled in each slot to an equivalent list of appointments, such that $L(\mathbf{x})_i$ is the appointment time of the i^{th} patient. The definition of L is as follows:

$$L(\mathbf{x})_i \equiv \min\left\{j : F_j(\mathbf{x}) \geq i\right\}, \quad (2)$$

where $F_j(\mathbf{x}) = \sum_{k=1}^{j} x_k$ is the number of patients scheduled to arrive at or before slot j. The total cost of updating the schedule from $\mathbf{x}$ to $\mathbf{x}'$ at time t, is

$$R_t(\mathbf{x}, \mathbf{x}') \equiv \sum_{i=1}^{N} U_t(L(\mathbf{x})_i, L(\mathbf{x}')_i); \quad (3)$$

which adds up the costs of all updated appointments. Patients that do not show up will not incur a rescheduling cost. Therefore, the expected cost of updating the schedule is $(1-p)R_t(\mathbf{x}, \mathbf{x}')$.

The clinic must serve all patients that show up. This might require staying open beyond the planned duration

[1] Given q, the schedule up till time t is irrelevant. But for ease of notation, we use the full schedule, $\mathbf{x}$, in the state definition instead of truncating the first t elements

of T periods. An additional cost of γ_O is incurred for each time slot beyond T during which the clinic remains open. We assume that patients cannot be *scheduled* to arrive beyond period T, however if $q > 0$ patients remain in the system at time $T+1$, the clinic must remain open for an additional q periods incurring an overtime cost of $\gamma_O q$. Additionally, in each of the q periods, all remaining patients, except the one being served would have to wait. The total waiting cost over all q periods would be given by the progression: $\gamma_W((q-1)+(q-2)+..0) = \gamma_W \sum_{i=0}^{(q-1)^+} i$.

We can now write down the Bellman equation for the cost incurred starting from any given state as follows:

$$
\begin{aligned}
V_t(q,\mathbf{x}) &= \gamma_W(q-1)^+ + \gamma_I I(q=0 \wedge \sum_{j=t+1}^{T} x_j > 0) \\
&\quad + \min_{\mathbf{x}' \in \mathcal{A}_t(\mathbf{x})} \Big[(1-p) R_t(\mathbf{x},\mathbf{x}') \\
&\quad + \mathop{\mathbb{E}}_{X'_{t+1}} V_{t+1}((q-1)^+ + X'_{t+1}, \mathbf{x}') \Big] \\
V_{T+1}(q,\mathbf{x}) &= \gamma_W \sum_{i=0}^{(q-1)^+} i + \gamma_O q
\end{aligned}
\tag{DR}
$$

where,

$$
\begin{aligned}
\mathcal{A}_t(\mathbf{x}) = \Big\{ \mathbf{x}' | F(\mathbf{x}') \le F(\mathbf{x}) \wedge \sum_{j=1}^{T} x'_j = N \\
\wedge\, x'_j = x_j \forall j \le t \Big\},
\end{aligned}
\tag{4}
$$

is the set of all valid schedule updates at time t when the current schedule is $\mathbf{x}$, and X'_{t+1} is the number of patients that show up in slot $t+1$. We assume that patient no-shows are independent of each other. Therefore, the expectation of $V_{t+1}((q-1)^+ + X'_{t+1}, \mathbf{x}')$ over X'_{t+1} is given by $\sum_{i=0}^{x'_{t+1}} b(i, x'_{t+1}, 1-p) V_{t+1}((q-1)^+ + i, \mathbf{x}')$, where $b(k,n,p)$ is the probability mass function of the binomial distribution for k successes over n trials with success probability p.

Formulation (DR) can be used to find the optimal policy, π^*, such that if the current state is $(t,q,\mathbf{x})$, updating to schedule $\pi^*(t,q,\mathbf{x})$ has the lowest expected cost. To compute the pre-day schedule, i.e., the schedule at the start of the day before any information on patient no-show is revealed, we search over all feasible schedules. The pre-day schedule is the first time that patients are allotted appointment times and therefore, there are no immediate schedule update costs associated with it. However, just like an intra-day schedule update, the initial appointment times must account for rescheduling costs that will be incurred later in the day. The cost incurred by a pre-day schedule, $\mathbf{x}$, is $TC(\mathbf{x}) = \mathbb{E}_{X_1} V_1(X_1, \mathbf{x})$. The optimal pre-day schedule ($\mathbf{a}$) is: $\mathbf{a} = \arg\min \Big\{ TC(\mathbf{x}) : \sum_{j=1}^{T} x_j = N \Big\}$.

Table 1. Summary of model notation

t	Current time slot
q	Number of patients awaiting service at the beginning of slot 't'
$\mathbf{x}$	Number of patients scheduled to arrive *at* the start of each time slot
N	Total number of patients
T	Total number of slots
p	Probability of a patient being a no-show
γ_I	Unit cost of server idling
γ_W	Unit cost of patient waiting (can be normalized to 1)
γ_O	Unit cost of operating in overtime
u_k	Cost of postponing an appointment that is k slots away, by one slot
$b(k,n,p)$	Probability mass function of the binomial distribution
$F(\mathbf{x})$	Number of patients scheduled to arrive *by* the start of each time slot
$L(\mathbf{x})$	Appointment time of the i^{th} patient under schedule $\mathbf{x}$
$U_t(j,j')$	The cost of moving an appointment from slot j to j' when t slots have elapsed
$R_t(\mathbf{x},\mathbf{x}')$	Function for cost of updating the schedule from $\mathbf{x}$ to $\mathbf{x}'$ at time t
$V_t(q,\mathbf{x})$	The minimum expected cost incurred from the beginning of slot t till the end of the day when the current state is $(t,q,\mathbf{x})$
$TC(\mathbf{x})$	Cost incurred when the pre-day schedule is $\mathbf{x}$
$\mathbf{a}$	The optimal pre-day schedule
$\pi*$	The optimal dynamic policy

4. Structural properties and solution methodology

As with most dynamic programs, the formulation specified in (DR) presents computational challenges. The time taken to compute the optimal policy can be quite large even for a standard problem size with $T = 8$ slots (column (DR) in Table 2). Finding structural properties of the optimal policy can help reduce the space of feasible schedules while also providing insight for the creation of heuristics. Unfortunately, as noted in the following observations, the optimal dynamic policy does not exhibit some seemingly intuitive properties that have been shown to hold for static schedules (see [13]).

Observation 1 *The optimal dynamic policy can leave slack in the schedule, i.e.* $\pi^*_j(t,q,x) = 0 \not\Longrightarrow \pi^*_{j'}(t,q,x) = 0, \forall j' > j$.

Observation 2 *Even if the server is guaranteed to be busy during a slot when a patient is scheduled to arrive, the optimal dynamic schedule does not necessarily push that patient to a later time, i.e. the following condition is not necessarily true:* $\pi^*_{t+q-1}(t,q,x) = 0$.

Observation 3 *Even though the dynamic scheduling policy can only move patients to a later time, it is possible that the optimal pre-day dynamic schedule brings in at least one patient later in the day than the optimal static schedule, i.e., the following may be true :* $F(\pi^*(t,q,x)) < F(\pi^*_s(t,q,x))$*, where* π^*_s *is the optimal solution to the static problem.*

These observations seem counter-intuitive if evaluated on the basis of direct idling and waiting costs. However, there is a parallel trade-off from two secondary effects of rescheduling that must be considered. Observation 1 and Observation 2 arise from the information effect, where bringing patients in earlier (even if it increases waiting costs) gives the provider more information on no-shows, therefore allowing better schedule adjustments for future slots. On the other hand, rescheduling can only be effective if there are patients with appointments far enough into the future such that their appointments can be moved without incurring high rescheduling costs. Therefore, in some scenarios where the static schedule places most of the patients in the earlier slots of the day (when no show rate and idling cost are high), the pre-day dynamic schedule places a relatively larger number of patients in later slots.

In the absence of any strong structural properties, the state space of the formulation (DR) cannot be reduced. However, note that solving for the optimal policy involves, for each state, a minimization over each schedule in the feasible set of schedule updates. As seen in Equation 4, the number of elements in the action set grows combinatorially in the number of patients and time slots. We now provide an equivalent formulation of the base dynamic program defined in the previous section, where, for each state, a maximum of T actions need to be evaluated. We label this formulation (DRE). The (DRE) formulation computes the same optimal policy and expected cost as the (DR) formulation (Proposition 1) while significantly reducing the computation time.

$$\begin{aligned}\tilde{V}_t(q,\mathbf{x}) = \min\Bigg\{ & \gamma_W (q-1)^+ \\ & + \gamma_I I(q=0 \wedge \sum_{j=t+1}^{T} x_j > 0) \\ & + \mathop{\mathbb{E}}_{X_{t+1}} \tilde{V}_{t+1}((q-1)^+ + X_{t+1}, \mathbf{x}), \\ & \min_{t<t'<T \wedge x_{t'}>0} \Big[\tilde{V}_t((q, \psi(\mathbf{x},t')) \\ & \qquad + (1-p) u_{t'-t} \Big] \Bigg\} \\ \tilde{V}_{T+1}(q,\mathbf{x}) = \gamma_W & \sum_{i=0}^{(q-1)^+} i + \gamma_O q \end{aligned} \tag{DRE}$$

where,

$$\psi(\mathbf{x},t')_j = \begin{cases} x_j - 1 & \text{if } j = t' \\ x_j + 1 & \text{if } j = t'+1 \\ x_j & \text{otherwise} \end{cases}$$

is an 'atomic update' that moves one patient from slot t' to slot $t'+1$.

Complete proofs of Proposition 1 and Lemma 1 have been omitted for the sake of brevity, but the following is an intuitive explanation to show how the (DR) and (DRE) formulations are equivalent. Given a state $(t,q,\mathbf{x})$, let $\mathbf{x}^* = \pi^*(t,q,\mathbf{x})$ be the optimal updated schedule. Lemma 1 shows that there must exist a sequence of atomic updates that takes the schedule from $\mathbf{x}$ to $\mathbf{x}^*$, and that each of these sequences must have the same rescheduling cost, $R_t(\mathbf{x},\mathbf{x}^*)$. From state $(t,q,\mathbf{x})$, the cost of making an atomic update at slot t' would be $V_t(q,\psi(\mathbf{x},t')) + (1-p)u_{t'-t}$. Proposition 1 shows that if t' is recursively chosen in a greedy way, i.e., $t' = \arg\min_{t<t'<T\wedge x_{t'}>0}[\tilde{V}_t((q,\psi(\mathbf{x},t')) + (1-p)u_{t'-t}]$, then the sequence of atomic updates generated would eventually lead to the schedule becoming $\mathbf{x}^*$.

To further clarify the equivalence between the (DR) and (DRE) formulations, there are a couple of points to be noted: (1) Even though the (DRE) formulation is restricted to making atomic updates to the schedule, it recursively makes multiple atomic updates at the start of the same time slot. Therefore, it still achieves the same schedule updates in each slot as the (DR), and (2) Even though the updated schedule is computed one atomic update at a time, when implemented in a real system, updates to patients will only be sent out once the entire sequence, and therefore $\mathbf{x}^*$ has been computed. In other words, the (DRE) formulation would be identical to the (DR) formulation as seen by the patients and providers because only the final updated schedule will be used to send out notifications while the underlying sequence of atomic updates remains internal to the model.

Lemma 1 *Given the current schedule, $\boldsymbol{x}$, for every feasible schedule update $\boldsymbol{x}'$:*

a There exists a sequence of time slots $\left\{\tilde{t}^k\right\}_{k=1}^{K}$ and corresponding sequence of schedules $\left\{\tilde{\boldsymbol{x}}^k\right\}_{k=1}^{K}$, such that $\tilde{\boldsymbol{x}}^1 = \boldsymbol{x}$, $\tilde{\boldsymbol{x}}^k = \psi(\tilde{\boldsymbol{x}}^{k-1}, \tilde{t}^{k-1})$, and $\tilde{\boldsymbol{x}}^K = \boldsymbol{x}'$.

b For every sequence, $\left\{\tilde{t}^k\right\}_{k=1}^{K}$, that satisfies the above condition, $\sum_{k=1}^{K} u_{\tilde{t}^k - t} = R_t(\boldsymbol{x}, \boldsymbol{x}')$.[2]

Proposition 1 *The optimization problem as formulated in* (DRE) *is equivalent to the formulation in* (DR). *$\tilde{V}_t(q, \boldsymbol{x}) = V_t(q, \boldsymbol{x})$ for all states $(t, q, \boldsymbol{x})$.*

Table 2 shows a comparison of the run time for the (DR) and (DRE) formulations. The experiments are conducted using a 3 GHz Intel i7 processor. Our dynamic program is implemented in Python 3.8. It can be seen that for realistic problem sizes ($T = 8$), the (DRE) is about a 100 times faster than the (DR) formulation, and can solve problem instances up to $T = 11$ time slots under the specified time limit of 300 seconds. However, the runtime of both formulations increases exponentially with problem size.

5. Numerical Study

In this section, we study how each model parameter affects the cost incurred when using dynamic scheduling as opposed to a static schedule. As an intermediate between the two extremes of dynamic and static scheduling, we also define a myopic schedule, where the optimal rescheduling decision in any given state is made under the assumption that there will be no more rescheduling. Clearly, the cost under myopic scheduling may be higher than that under dynamic scheduling, but the computation time for myopic schedules is much lower.

[2] All proofs in this paper are available upon request.

T	N	(DR) Runtime	(DRE) Runtime
5	6	0.05	0.01
5	8	0.31	0.03
5	9	0.78	0.05
6	8	1.11	0.07
6	9	3.33	0.12
8	9	27.94	0.45
8	10	119.56	1.09
9	10	293.69	1.95
9	11	>300	5.02
9	12	>300	12.14
10	11	>300	9.77
10	13	>300	120.80
11	12	>300	127.76
11	13	>300	>300

Table 2. Comparison of time (in seconds) required to solve for the optimal policy and pre-day schedule using the (DR) **and** (DRE)**) formulations. Only instances where the solution was computed within 300 seconds were recorded.**

To calibrate baseline values for our model parameters, we start by normalizing the cost of waiting (γ_W) to one. For the cost of idling and overtime, we use from ([5]) empirical estimates of the relative cost of idling and overtime for outpatient clinics. The probability of no-show varies widely even within the context of outpatient appointments. [5] and [28] report that no-show rates may vary between 0% and 80% across clinics and estimate the average no-show rate to be 38% and 31% respectively. Therefore, we set our baseline no-show rate to an intermediate value of 0.35. To calibrate the cost of rescheduling (**u**), we set it to be exponentially decaying with respect to how far out the appointment time is from the current time ($u_k = e^{-\delta k}$). This reflects the notion that for an appointment that is further out (larger k), the inconvenience cost of moving the appointment decreases at a lower rate. The decay parameter (δ) reflects the level of patient flexibility with a higher value of δ implying that it is cheaper to reschedule patients. Figure 1 shows a plot of the baseline chosen for the rescheduling cost ($u_k = e^{-0.231k}$) along with two alternatives: a linear functional form to capture a constant rate of decrease in the cost of rescheduling and a 'shifted' exponential for scenarios where rescheduling within a certain time of the appointment must not be allowed. For all three choices, the value of δ is chosen by setting the cost of

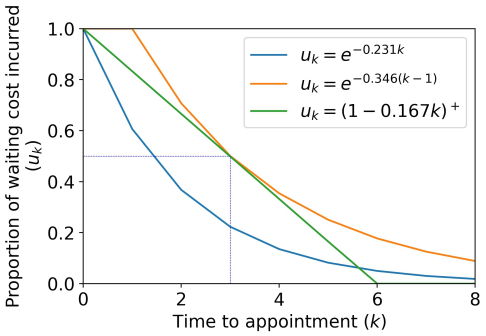

Figure 1. Choices of functional forms used to calibrate the rescheduling cost(u). $u_k = e^{-0.231k}$ was chosen as the baseline.

rescheduling an appointment that is three hours away (u_3) to be 50% of the waiting cost (shown by the dotted blue lines). Table 3 summarizes the baseline values used for all model parameters.

Table 3. Baseline values of parameters used in numerical analysis. (Based on estimates in [1]Rust et al. (1995) and [2]Cayirli et al. (2006)).

Waiting cost(γ_W)	=	1 (Normalized)
Idling cost(γ_I)	=	$10^{[2]}$
Overtime cost(γ_O)	=	$1.5\gamma_I^{[2]}$
Pr. no-show(p)	=	$0.35^{[1][2]}$
Rescheduling cost(u_k)	=	$e^{-\delta k}$
Patient Flexibility(δ)	=	0.231

Figure 2 shows the expected cost of the static, myopic and dynamic scheduling policies with respect to the patient flexibility parameter (δ). Under the static schedule, there is no rescheduling, and the optimal static schedule remains the same irrespective of changes in patient flexibility. However, as rescheduling becomes less expensive, the myopic and dynamic policies perform better with the myopic policy capturing a significant part of the cost reduction. The breakdown of the expected cost incurred by the dynamic policy in Figure 3 shows that along with a decrease in expected cost, higher values of patient flexibility also significantly change the cost composition. When patients are inflexible ($\delta = 0$), the cost breakdown is identical to that from static scheduling. As δ increases, patient waiting costs decrease as appointments that are likely to be delayed are moved to later times. A portion of this waiting cost reduction though is incurred as

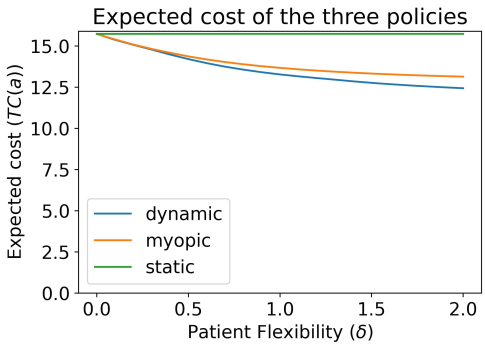

Figure 2. Expected cost of the static, myopic and dynamic policies.

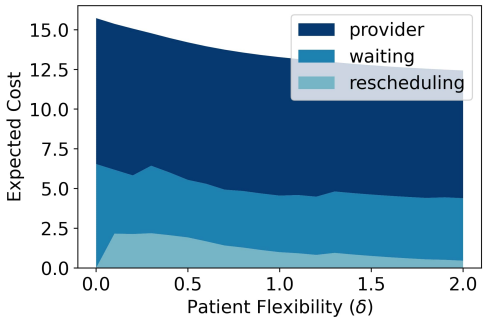

Figure 3. Breakdown of expected cost incurred by dynamic policy. Cost of idling and overtime is combined into 'provider cost' for ease of exposition.

rescheduling cost. As δ increases further, the optimal pre-day schedule changes to bring patients in earlier. Therefore, the 'provider cost' (sum of provider idling and overtime cost) decreases while patient waiting costs go up.

Another perspective to understand what drives the reduction in expected cost when rescheduling, is to look at the cumulative distribution of the cost incurred under each policy (Figure 4). The cost incurred by the dynamic and myopic policies does not stochastically dominate the cost incurred by the static policy. However, it can be seen that the dynamic and myopic policies are better at reducing cost at the tail end of the cost distribution, i.e., they provide the largest gains in some of the worst scenarios. In the example plotted in Figure 4, the costs for the optimal static and dynamic policies are almost equal, upto the 60^{th} percentile. But the 80^{th} percentile of the cost of the optimal static policy is 26.0, as opposed to 23.3 for the optimal dynamic policy (a reduction of 10.4%).

As pointed out earlier, the probability of no-show and the costs of idling and overtime can vary significantly across clinics. Figure 5 plots the

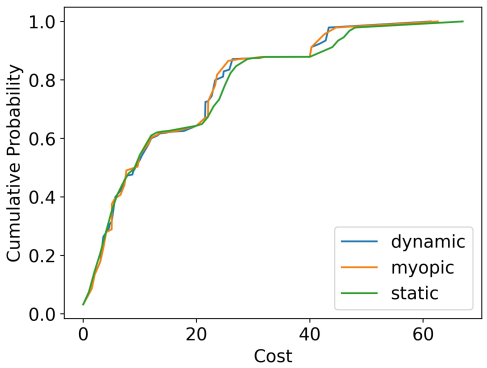

Figure 4. Cumulative distribution of the cost incurred by the static, myopic and dynamic policies. Rescheduling policies are better at cutting cost at the tail end of the cost distribution.

percentage reduction in expected cost that is achieved through rescheduling (as compared to using static scheduling) when the probability of no-show and cost of idling are changed. Baseline values for both parameters are highlighted by the dotted blue line. It can be seen that the reduction in expected cost is robust to small deviations in the parameter values away around their baseline. Note that in contrast to Figure 2, we now plot the percentage reduction in cost instead of the absolute cost because unlike patient flexibility, the probability of no-show and cost of idling also affect the cost of the static policy. This also explains why the plots in Figure 5 are not smooth.

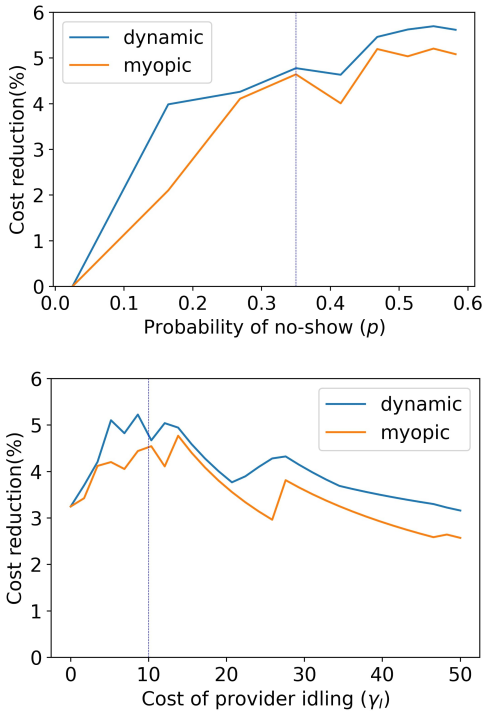

Figure 5. Sensitivity of the cost savings from rescheduling to model parameters (as measured by percentage cost reduction relative to static scheduling). The dotted blue lines represent the baseline values of $p = 0.35$ and $\gamma_I = 10$. The plots show that cost savings from rescheduling are robust to changes in these model parameters.

6. Conclusion

Outpatient appointments have conventionally been treated as a one-time contract between a patient and the provider. In this work, we introduced the idea of intra-day rescheduling where the initial appointment time assigned to patients may be updated based on the state of the system during the day. We find that the ability to update appointment times can help reduce both patient and provider costs.

This work is an important first step in understanding the benefits of rescheduling, albeit in a simplified setting. Implementing a rescheduling policy in the real world will require work that helps account for various factors that we have not considered here. For example: there may be clinical restrictions on which appointments can or cannot be moved. With such a mix of flexible and inflexible appointments, it would be important to find the optimal slots for inflexible appointments in addition to finding the optimal scheduling policy. A relatively easier relaxation is to allow for stochastic (but discrete) service times.

Another direction to explore is to augment the state space to distinguish between patients that have been rescheduled from the ones that have not. This enables the model to: (1) restrict patients from being rescheduled multiple times, and (2) account for the effect of rescheduling on patients' no-show probabilities. Quantifying the effect of rescheduling on no-show probability could be done using estimation techniques similar to those used by earlier work on individualized no-show prediction ([29], [30]).

We model the inconvenience associated with changing appointment times through the cost of rescheduling. Even though an exponentially decreasing function seems to be a reasonable choice for calibrating the cost of rescheduling, further empirical work to estimate rescheduling cost would be very beneficial in improving the quality of the optimal dynamic policy.

As shown by Table 2, there are significant computational challenges in being able to solve for the optimal rescheduling policy as the problem sizes get larger. It may be possible to solve the problem

more efficiently by exploiting structural properties of the optimal policy. The existence of such structural properties remains an open question. Another approach is to explore the use of heuristics that can reduce either the state space or the number of actions. Some examples of promising heuristic are: (1) not allowing slack within the schedule, (2) restricting the number of patients rescheduled at each decision epoch, and (3) rescheduling once every $k > 1$ slots.

References

[1] J. Welch and N. J. Bailey, "Appointment systems in hospital outpatient departments," *The lancet*, vol. 259, no. 6718, pp. 1105–1108, 1952.

[2] L. R. LaGanga and S. R. Lawrence, "Clinic overbooking to improve patient access and increase provider productivity," *Decision Sciences*, vol. 38, no. 2, pp. 251–276, 2007.

[3] S. Kim and R. E. Giachetti, "A stochastic mathematical appointment overbooking model for healthcare providers to improve profits," *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and humans*, vol. 36, no. 6, pp. 1211–1219, 2006.

[4] D. Gupta and B. Denton, "Appointment scheduling in health care: Challenges and opportunities," *IIE transactions*, vol. 40, no. 9, pp. 800–819, 2008.

[5] T. Cayirli and E. Veral, "Outpatient scheduling in health care: a review of literature," *Production and operations management*, vol. 12, no. 4, pp. 519–549, 2003.

[6] T. Cayirli, E. Veral, and H. Rosen, "Designing appointment scheduling systems for ambulatory care services," *Health care management science*, vol. 9, no. 1, pp. 47–58, 2006.

[7] C.-J. Ho and H.-S. Lau, "Evaluating the impact of operating conditions on the performance of appointment scheduling rules in service systems," *European Journal of Operational Research*, vol. 112, no. 3, pp. 542–553, 1999.

[8] C.-J. Ho and H.-S. Lau, "Minimizing total cost in scheduling outpatient appointments," *Management science*, vol. 38, no. 12, pp. 1750–1764, 1992.

[9] C. G. Moore, P. Wilson-Witherspoon, and J. C. Probst, "Time and money: effects of no-shows at a family practice residency clinic," *Family Medicine-Kansas City-*, vol. 33, no. 7, pp. 522–527, 2001.

[10] V. Pesata, G. Pallija, and A. A. Webb, "A descriptive study of missed appointments: families' perceptions of barriers to care," *Journal of Pediatric Health Care*, vol. 13, no. 4, pp. 178–182, 1999.

[11] M. Murray and C. Tantau, "Same-day appointments: exploding the access paradigm," *Family practice management*, vol. 7, no. 8, p. 45, 2000.

[12] M. Murray and C. Tantau, "Redefining open access to primary care.," *Managed care quarterly*, vol. 7, no. 3, pp. 45–55, 1999.

[13] L. W. Robinson and R. R. Chen, "A comparison of traditional and open-access policies for appointment scheduling," *Manufacturing & Service Operations Management*, vol. 12, no. 2, pp. 330–346, 2010.

[14] B. Rajan and A. Seidmann, "Improving open access policy for scheduling outpatient appointments," in *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pp. 4760–4763, IEEE, 2016.

[15] F. Yuan, V. Tilson, J. Szmerekovsky, and R. Spurr, "Simulation model to study provider capacity release schedules under time-varying demand rate for acute appointments, demand for follow-up appointments, and time-dependent no show rate," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[16] P. E. Hasvold and R. Wootton, "Use of telephone and sms reminders to improve attendance at hospital appointments: a systematic review," *Journal of telemedicine and telecare*, vol. 17, no. 7, pp. 358–364, 2011.

[17] A. Parikh, K. Gupta, A. C. Wilson, K. Fields, N. M. Cosgrove, and J. B. Kostis, "The effectiveness of outpatient appointment reminder systems in reducing no-show rates," *The American journal of medicine*, vol. 123, no. 6, pp. 542–548, 2010.

[18] L. Cui, V. Tilson, and G. Dobson, "Reminder systems for reducing no-shows in general practices," in *2012 45th Hawaii International Conference on System Sciences*, pp. 4820–4827, IEEE, 2012.

[19] L. W. Robinson and R. R. Chen, "Scheduling doctors' appointments: optimal and empirically-based heuristic policies," *Iie Transactions*, vol. 35, no. 3, pp. 295–307, 2003.

[20] N. Liu, S. Ziya, and V. G. Kulkarni, "Dynamic scheduling of outpatient appointments under patient no-shows and cancellations," *Manufacturing & Service Operations Management*, vol. 12, no. 2, pp. 347–364, 2010.

[21] K. Muthuraman and M. Lawley, "A stochastic overbooking model for outpatient clinical scheduling with no-shows," *Iie Transactions*, vol. 40, no. 9, pp. 820–837, 2008.

[22] Y. Lu, X. Xie, and Z. Jiang, "Dynamic appointment scheduling with wait-dependent abandonment," *European Journal of Operational Research*, vol. 265, no. 3, pp. 975–984, 2018.

[23] C. Laan, M. van de Vrugt, J. Olsman, and R. J. Boucherie, "Static and dynamic appointment scheduling to improve patient access time," *Health systems*, vol. 7, no. 2, pp. 148–159, 2018.

[24] L. V. Green, S. Savin, and B. Wang, "Managing patient service in a diagnostic medical facility," *Operations Research*, vol. 54, no. 1, pp. 11–25, 2006.

[25] R. Kolisch and S. Sickinger, "Providing radiology health care services to stochastic demand of different customer classes," *OR spectrum*, vol. 30, no. 2, pp. 375–395, 2008.

[26] Y. Gocgun, B. W. Bresnahan, A. Ghate, and M. L. Gunn, "A markov decision process approach to multi-category patient scheduling in a diagnostic facility," *Artificial intelligence in medicine*, vol. 53, no. 2, pp. 73–81, 2011.

[27] G. Cheng, K. Chandrasekher, and J. Walrand, "Static & dynamic appointment scheduling with stochastic gradient descent," in *2019 American Control Conference (ACC)*, pp. 2092–2099, IEEE, 2019.

[28] C. T. Rust, N. H. Gallups, W. S. Clark, D. S. Jones, and W. D. Wilcox, "Patient appointment failures in pediatric resident continuity clinics," *Archives of pediatrics & adolescent medicine*, vol. 149, no. 6, pp. 693–695, 1995.

[29] Y. Li, S. Y. Tang, J. Johnson, and D. A. Lubarsky, "Individualized no-show predictions: Effect on clinic overbooking and appointment reminders," *Production and Operations Management*, vol. 28, no. 8, pp. 2068–2086, 2019.

[30] M. Dashtban and W. Li, "Deep learning for predicting non-attendance in hospital outpatient appointments," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.