An Integrated Approach to Demand and Capacity Planning in Outpatient Clinics

Navid Izady
University of Southampton, n.izady@soton.ac.uk

An outpatient clinic serving two independent demand streams, one representing advance booking requests and the other same-day requests, is considered. Advance requests book their appointments through an electronic booking system for a future day, and same-day requests are served on the day they arise. Taking an integrated approach to demand and capacity planning, a policy formulation compatible with electronic booking systems is proposed that incorporates major operational levers suggested in the literature. It combines a static slot publication policy, which specifies the pattern under which slots are released to the booking system, with an allocation policy that dynamically adjusts the daily workload of advance patients. The optimal policies are found numerically by developing a novel queueing model that efficiently evaluates major performance metrics. The application of the model with real data, obtained from one clinic with carve-out delivery and another with advanced access, demonstrates substantial savings.

Key words: Health care; Queues: Applications; Demand and Capacity Planning

History:

1. Introduction

Outpatient medical facilities must typically serve both patients who require a same-day visit as well as those who book an appointment in advance. This applies not only to clinics that operate a “carve-out” mode of delivery, where a few slots in each day are reserved for patients with urgent medical needs and the rest are available for advance booking, but also to clinics that have implemented “advanced access”. The primary objective in advanced access is to offer every patient a same-day appointment regardless of urgency (Murray and Berwick 2003). However certain patient groups, such as commuters who often experience difficulty in making and attending same-day appointments, tend to decline this offer in favour of advance booking (Pope et al. 2008). As such with both delivery modes clinics must decide how much of their daily capacity should be allocated to "advance booking" patients, and consequently how much be left open in anticipation of "same-day" demand. Clinics may also decide to serve more pre-booked patients on particular days than originally scheduled for those days in order to eliminate excessive appointment backlogs following a temporary surge (decline) in demand (supply).

In addition to the capacity planning decisions mentioned above, clinics may exercise some control over the demand streams. In some clinics this is achieved through assigning each physician a panel of patients, the size of which specifies the total demand for that physician. Appointment scheduling window, i.e. the length of time in advance a patient can schedule an appointment with a provider, is another operational lever clinics may deploy to regulate the demand for advance appointments. In this paper, we develop an integrated approach that guides clinics in making such capacity and demand planning decisions. We assume advance booking patients schedule their appointments through an electronic booking system (EBS), and propose a novel policy formulation that utilizes the existing functionalities of such systems. We find the optimal values of the policies proposed numerically through developing a new queueing model that captures the major complexities observed in outpatient environments.

A prime example of an EBS implemented on a large scale is that of the Choose and Book (CaB) system used in the UK National Health Service. It enables routine patients referred to specialty clinics to book their first outpatient appointment online. Under this system, the providers store the information about their time slots, including timings, the clinicians providing them, and whether they are publishable on CaB or not, in their computerized systems linked to the CaB. The free and publishable time slots on each day are released to the CaB system a given number of days in advance as specified by the appointment scheduling window (or the "polling range" as referred to in the CaB context) selected by the provider. These slots will then be available to routine patients seeking appointments: once a routine referral is deemed necessary for a patient, first the referring clinician jointly with the patient chooses where the patient must be referred to depending on patient's condition and preferences, and second the patient books the most convenient slot from a menu of available slots displayed by the CaB for her chosen provider. The slots not released to the CaB will be offered to urgent referrals and walk-in patients on the day they arise. There is a wide range of other EBS's used in primary and specialty clinics in different countries, e.g. *ZorgDomein* for referral to specialty clinics in Netherlands (Dixon et al. 2010) and *ZocDoc* for primary care practices in the US (Zocdoc 2015). The majority of these systems function on a similar basis to the CaB.

When the EBS shows no available slot for the first choice provider of an advance booking patient, the patient might switch to a different provider, or insist on being served by that particular provider. The latter may be due to reasons such as geographical proximity or reputation of the provider. Similar to Jiang et al. (2012), we call these two categories of advance booking patients "flexible" and "dedicated", respectively. Some EBS's have built-in features enabling dedicated patients to enforce an appointment with their chosen provider when the EBS shows no available slot. In the CaB, for instance, this is provided through the "Defer to Provider" option. Alternatively, in some EBS's patients are advised to phone the clinic directly when they cannot find an appointment online. Flexible patients on the other hand forgo the difficulties associated with securing an appointment through these alternative routes and seek care elsewhere.

To enable the clinics to effectively manage the demand for and supply of appointments through an EBS, we develop an integrated approach that combines a static *slot publication* policy with a dynamic *allocation* policy. We define the slot publication policy as a two-dimensional policy where the first dimension specifies the number of slots in each period of time that are made publishable to the EBS by the clinic, and the second dimension determines the number of periods in advance that such slots are released to the EBS. The slot publication policy is in fact a combination of two major operational levers investigated separately in the literature. The first dimension is similar to the optimal urgent reservation level studied in Qu et al. (2007), Robinson and Chen (2010) and Dobson et al. (2011) for capacity allocation, and the second dimension is the appointment scheduling window proposed in Liu (2015) as a mechanism for demand management.

The slot publication policy, or its first dimension to be accurate, is a *threshold* policy as it allocates a fixed number of slots to advance requests in each time period. However, studies on allocation and advance scheduling suggest that a threshold policy may not be optimal. For example, Truong (2015, p. 3) prove that with a specific cost structure, "the optimal policy does not schedule the same number of regular [i.e. advance] patients for each day but dynamically increases this daily regular workload as the total number of regular patients in the system increases." Our discussions with clinic managers suggest that this is consistent with what happens in practice; more advance patients are seen in the clinic when there is a large appointment backlog. To reflect this, one could make the first dimension of the slot publication policy dynamic, varying it by the numbers in the backlog. But the resulting impact would appear only after the scheduling window which might stretch up to several weeks. Therefore, we define a second policy called the dynamic allocation policy. It specifies the additional number of pre-booked patients the clinic serves in each time period, which varies depending on the size of the appointment backlog in the beginning of that period.

The combination of slot publication and dynamic allocation policies gives clinics a framework to plan their capacity and demand optimally depending on the optimality criteria they choose. Here we define the optimal joint policy as the one that minimizes the average cost associated with providing overtime slots whilst ensuring that advance patients' access-to-care requirement and provider's service level requirement are met. The overtime cost is incurred when the number of same-day requests on a day exceeds the number of slots available for them. The access requirement we consider for advance booking requests is to be served on average in less than a pre-specified amount of time. The service level requirement is to ensure that the average number of advance flexible patients forced to switch provider as a result of no slots being available online falls below a given threshold. To find the optimal values for the two policies, we enumerate over a range of possible values using a new queueing model for performance evaluation. As will be illustrated through two real examples, the efficiency of the queueing model enables us to explore a wide range of policies in a relatively short time. Drawing on empirical results obtained from our experiments, we also propose a search process that reduces the number of policy combinations that must be evaluated.

Our planning approach is integrated as (i) it considers capacity allocation and demand management decisions in a single model, and (ii) it combines a static publication policy, which is essential for compatibility with EBS's, with a dynamic allocation policy that could substantially improve the efficiency. Our approach captures the intrinsic complexities observed in many outpatient environments including: arbitrary distributions for same-day and advance booking requests; patient no-show and subsequent rescheduling, and their potential dependence on appointment delays; and clinic slot cancellations caused by provider vacations, illnesses, or absences to attend professional meetings. A restrictive aspect of our model is that a first-come first-serve (FCFS) discipline is assumed in the analysis. However we also express a milder condition under which our analysis remains valid even without FCFS.

We also believe that our optimization model is more relevant to practical applications than the vast majority of models developed in the literature. This is because in these models cost values are typically defined for every day an advance patient is made to wait, e.g. Gerchak et al. (1996) and Truong (2015), and/or for every flexible patient who is forced to switch provider, e.g. Patrick et al. (2008) and Jiang et al. (2012). Although this makes the analysis simpler by consolidating all performance metrics in a single cost function, validating and implementing the resulting models would be difficult as it is incredibly hard to find realistic estimates for the two cost elements. We take a different approach by setting maximum thresholds for the average access time, i.e. the time between a patient request for an appointment and the actual appointment, of advance booking patients as well as the average number of flexible patients turned away. These thresholds may reflect similar performance targets imposed at a local or national level, or may simply represent the clinic desired performance objectives.

We apply our model to two sets of data, first a dataset obtained from a specialty clinic in the UK, and second the data for a magnetic resonance imaging (MRI) clinic in the US as reported in Green and Savin (2008). Our experiments show that deployment of simple allocation policies, that require only one or two additional consultations per week, could bring about substantial savings as opposed to serving exactly the same number of advance patients as scheduled. They also indicate that sequential optimization of the slot publication and dynamic allocation policies may lead to sub-optimal solutions, justifying our integrated approach. We further establish that, in the absence of dynamic allocation policies, the average number of flexible patients diverted with the optimal slot publication policy would reach its maximum threshold, while the average access time of advance patients would typically fall substantially below the corresponding maximum. When dynamic allocation policies are jointly considered with slot publication policies, however, both metrics reach their threshold values under optimality.

2. Literature Review

Our work is related to the appointment scheduling literature, see Cayirli and Veral (2003) and Gupta and Denton (2008) for comprehensive reviews. This literature can be divided to “intra-day” and “multi-day” scheduling. In intra-day scheduling, the focus is on a single day and intra-day measures such as patients’ office waits and providers’ idle/over-time. The majority of papers in this area seek to determine the optimal sequence of serving patients so that a combination of patients’ office wait and providers’ idle/over-time is minimized. Recent examples include Hassin and Mendel (2008), Klassen and Yoogalingam (2009), Koeleman and Koole (2012) and Cayirli et al. (2012). Some others investigate the optimal proportion of a day’s capacity that must be reserved for same-day requests in the context of advanced-access, see Qu et al. (2007) and Robinson and Chen (2010).

Multi-day scheduling is on the other hand concerned with access time. The focus of our research is also on multi-day scheduling. Truong (2015) divide this area into two main paradigms, “allocation scheduling” and “advance scheduling”. In allocation scheduling, “a wait list is maintained and patients are notified on the day of their appointments”, whereas in advance scheduling, “patients are given appointments in the future at the time of request”(Truong 2015, p. 1). See Gerchak et al. (1996), Ayvaz and Huh (2010), and Min and Yih (2014) for allocation scheduling, and Patrick et al. (2008) and Gocgun and Ghate (2012) for advance scheduling.

In a fundamental study, Truong (2015) show that for a broad class of advance scheduling problems with same-day and advance booking requests, the optimal scheduling policy can be exactly and efficiently constructed from a solution to an associated simple-to-calculate allocation scheduling problem. However the policy formulation proposed in their work as well as in other studies on advance and allocation scheduling does not fit the requirement of EBS’s. This is because in their formulation the demand for advance booking in the current time period must be fully known before a decision is made as to when each of the arriving requests must be met in the future, while with the EBS a menu of slots is presented to patients, from which they choose one slot as they arrive randomly over time. The studies of Liu et al. (2010) and Feldman et al. (2014) are more relevant in this sense, as they seek to find the optimal set of days that must be offered to patients in each time period. Their implementations are not straight forward however as the corresponding heuristic algorithms must be coded within the EBS. There are other studies on multi-day scheduling that do not fall within advance or allocation scheduling paradigms. In particular, Dobson et al. (2011) consider an outpatient setting where a fixed number of slots in each day must be reserved for urgent patients. They develop numerical models to find the optimal urgent reservation level as a function of cost parameters and the order in which routine and urgent patients call for appointments.

The impact of appointment scheduling window has not explicitly been considered in any of the papers cited above. However, using stylized $M/M/1$ queueing models, Liu (2015) show that optimal choice of the appointment window can lead to substantial efficiency gains, especially when other demand management mechanisms, such as setting the patient panel size, are not available. Overall, on the one hand, we introduce the slot publication policy to specify the pattern under which slots must be released to the EBS in mid to long-term periods. It captures the joint impact of the urgent reservation level and appointment scheduling window, and is also similar to the two dimensional policy proposed in Jiang et al. (2012). On the other hand, motivated by advance and allocation scheduling literature, we introduce the dynamic allocation policy to improve the efficiency. To the best of our knowledge, this is the first paper that proposes such a policy formulation for integrated demand and capacity management in outpatient clinics.

We find the (near-)optimal values of slot publication and dynamic allocation polices through enumeration. This is achieved by developing a queueing model for appointment backlog and proposing an efficient numerical method for evaluating its performance metrics. A number of queueing models are developed in the literature for appointment queues. Green and Savin (2008) propose $M/D/1$ and $M/M/1$ queues with backlog-dependent no-show probability to study panel size decisions. Creemers and Lambrecht (2010) and Kortbeek et al. (2014) develop two-time scale queueing models, representing both multi-day and intra-day performance. Izady (2015) propose a discrete-time bulk service model with cancellations and no-shows for appointment capacity planning. Their model is more flexible and fitting to the reality of outpatient clinics than the continuous-time models developed in Green and Savin (2008) and Creemers and Lambrecht (2010). Our queueing model is in fact a state-dependent generalization of the model proposed in Izady (2015). We make this model state-dependent so that both service capacity and arrival distribution can vary by the size of appointment backlog. This allows us to model the joint impact of slot publication and dynamic allocation policies. Jiang et al. (2012) also propose an $M/D/1$ queue with state-dependent arrival process but they present performance metrics only for the extreme cases where all patients are either flexible or dedicated. Dynamic allocation policies are not also considered in their work.

Our work differs from that of Izady (2015) in the following ways. First, we include both same-day and advance booking requests while Izady (2015) consider only the latter group. Second, the focus of Izady (2015) is only on the appointment capacity, whereas we consider the impact of scheduling window and a new dynamic allocation policy as well. Third, although both papers follow a probability generating function approach to evaluate the performance metrics, ours is more challenging due to the complexities caused by the state-dependence of our queueing model. Finally, the models in Izady (2015) are limited to performance evaluation, while we also propose a heuristic search process for finding optimal policies.

3. Problem Formulation

We assume the clinic provides services for two independent demand streams, one representing same-day requests and the other advance booking requests. Note that in the carve-out mode the urgency of care determines the group to which a patient belongs, while in the advanced access the preference of a patient is the major identifier. We divide the time axis into equally spaced intervals (periods), numbered $1,2,3,\dots$, and assume a nominal capacity of r regular slots is available in each interval. Same-day requests (or same-period requests, to be precise) must be met within their arrival periods, while advance booking requests book the first available slot in the intervals *following* their arrival intervals through the EBS. We define the appointment backlog (or appointment queue) in the beginning of a time interval before the service of that interval begins as the number of advance booking patients who have already scheduled an appointment but not yet served in the clinic. Note that a time period in our representation does not necessarily correspond to a working day. In fact, we observe that demand and supply patterns in some outpatient clinics vary substantially during a week but are relatively stable across the weeks, making a weekly time unit more appropriate for capacity and demand planning purposes.

Let (n,p) denote the slot publication policy of the clinic, where $0 < n \leq r$ determines the number of slots pre-allocated to advance booking patients in each interval, and p is the number of time units in advance that such slots are released to the EBS, i.e. the polling range. (We use the terms scheduling window and polling range interchangeably throughout.) Then $f = pn$ is the total number of slots available on the EBS for advance booking patients. If all f slots of the clinic are filled, an advance booking patient is assumed to become a dedicated patient with probability θ and a flexible patient with probability $1-\theta$, independently of everything else in the system. Dedicated patients will be offered appointments by the clinic, in a time period beyond the scheduling window, and flexible patients switch to a different provider. We consider a static slot publication policy over the planning horizon.

We assume in each time period, apart from serving patients scheduled for that period, the clinic may serve some of the patients scheduled for future periods. The number of *additional* advance patients served in a time period is specified by the dynamic allocation policy $e^{(i)}$, where i is the size of backlog in the beginning of that period before the service starts, and $e^{(i)}=0$ for $i=0,\dots,n$. Motivated by Truong (2015), we assume that $e^{(i)}$ increases in i. We further assume that $e^{(i)} \le r-n$ for all i, and $e^{(i)} \le i-n$ for $i \ge n$, i.e. the number of additional visits is restricted by the total number of regular slots and the number of patients scheduled to be served in future periods, respectively. $e^{(i)}$ additional patients are selected on a FCFS basis, and the schedule is updated at the end of each time period to fill up the slots released by these patients in a way that FCFS is preserved. This entails expediting some patients, to which we assume patients always consent.

We assume that in the beginning of each time interval before the service of that interval begins, a random number of slots in that interval is cancelled by the clinic as a result of providers' vacations, delays and absences. This results in some appointments being cancelled and subsequently rescheduled if the slots have already been booked. Furthermore, some advance patients may not show up for their appointments at all, or cancel their appointments too late to allow for new patients to be substituted. We refer to both cases as patient no-show, and in line with some empirical evidence reported in the literature, e.g. Gallucci et al. (2005), Green and Savin (2008) and Liu et al. (2010), assume that the rate of no-shows may increase with increasing appointment backlogs. We assume that at the end of each time interval, no-shows of that interval who request a new appointment as well as patients whose appointments are cancelled by the clinic in that interval are given new appointments at the end of the schedule.

All same-day patients turn up for their consultations. They are seen in the slots that remain unfilled and are not cancelled by the clinic, plus overtime slots if needed. Note that slots will remain unfilled if they are not booked by patients in advance and also not allocated to advance patients as a result of extra visits enforced by the dynamic allocation policy. We assume all unfilled slots will be used when same-day demand exists. This may not necessarily happen in reality due to late arrival of same-day requests. In addition, although it might be possible to schedule patients carefully during a time interval so that some same-day patients are served in the slots left unused by no-shows, see e.g. Zacharias and Pinedo (2014), we suppress that level of detail and assume all no-show slots will be wasted.

To summarize, the order of activities taking place in period t with i patients in the backlog at the start of the period is as follows: i) a maximum of n free slots on period $t+p$ is released to the EBS, (ii) the slots cancelled by the clinic are identified, (iii) patients scheduled to be seen on period t plus $e^{(i)}$ additional patients (selected on a FCFS order), who turn up for their appointments and their slots have not been cancelled, are served in the clinic, (iv) new patients take the slots available on the EBS (and beyond that if they are dedicated) on a FCFS order, (v) same-day patients are seen in the slots left unused, and overtime slots if necessary, (vi) The schedule is updated to fill up the slots released by additional patients served in period t, preserving the FCFS order, (vii) appointments are booked for re-shows and cancellations at the end of the schedule, (viii) the EBS is updated according to the schedule.

As an illustrative example, consider a clinic with one day as the time unit, $r = 4$ regular slots per day, and the slot publication policy $(n = 2, p = 4)$. Consider the dynamic allocation policy given below

$$e^{(i)} = \begin{cases} 0, & i \leq 5, \\ 1, & i > 5. \end{cases} \tag{1}$$

Figures 1 and 2 show the evolution of the appointment schedule for this clinic under two different assumptions. In Figure 1, we assume a FCFS discipline is followed as explained above. In Figure

2, we relax FCFS but with a condition that all the slots in the current day must be filled by the time the service of that day begins if any of the slots in the following days are taken.

In the diagrams of Figures 1 and 2, each row shows the status of the schedule in the current as well as following days in the beginning of the days specified by the vertical axis before the service begins. Patients are represented by numbers and slots by circles. Dotted circles represent the slots that are given to patients beyond the polling range. We assume there are no slot cancellations except for day four when two slots are cancelled. Patient 7 in Figure 1 and patient 9 in Figure 2 are assumed to miss their appointments and reschedule new ones. Below we provide a day-by-day explanation of the events in the diagram in Figure 1.

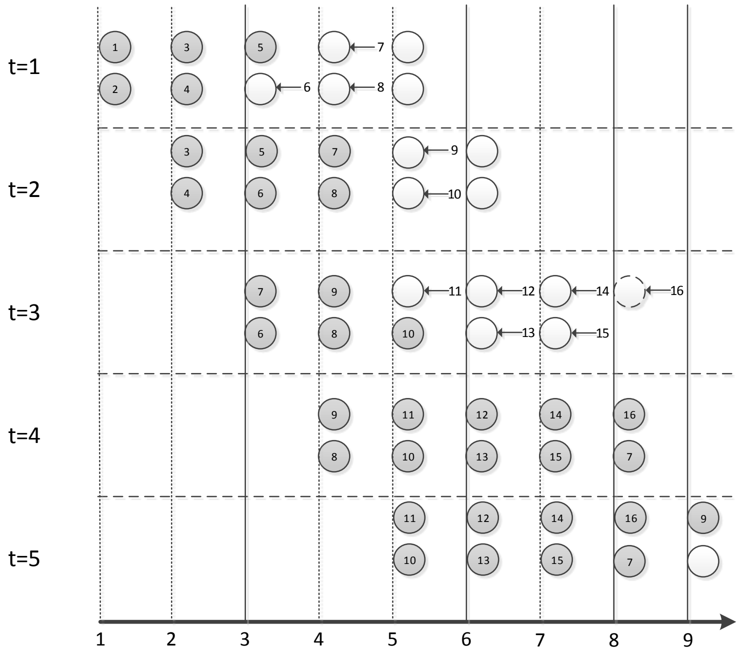
- At the beginning of day one ($t = 1$), two free slots on day five are released to the EBS. With five patients in the appointment queue, based on Equation (1), only two patients scheduled for this day are seen in the clinic. Three new advance patients arriving during the day, i.e. patients 6, 7, and 8, take the slots available on days three and four. There will be two slots left open to be used by same-day patients.

- At the beginning of day two ($t = 2$), two free slots on day six are released to the EBS. With six patients in the appointment queue, one additional patient must be seen in the clinic. Thus, apart from patients 3 and 4, the clinic serves patient 5 during the day. Note that EBS's are typically updated only once a day, often at midnight, so the slot freed up by patient 5 does not show up on the system immediately. As a result, two new advance patients 9 and 10 take the two slots available on day five. At the end of the day, the schedule (and the EBS) is updated by moving the appointments of patients 7 and 9 to days three and four, respectively, preserving the FCFS. There will only be one slot left open to be used by same-day patients.

- At the beginning of day three ($t = 3$), two free slots on day seven are released to the EBS. With five patients in the appointment queue, only two patients scheduled for this day must be seen in the clinic. Patient 6 is seen during the day but patient 7 does not turn up for her appointment. Five out of the six new advance patients, i.e. patients $11, \ldots, 15$, take the slots available on days five, six, and seven. Patient 16 does not find any slot available but as she turns out to be a dedicated patient, she is given an appointment on day eight. At the end of the day a new appointment is given to patient 7 on day eight. There will be two slots left open to be used by same-day patients.

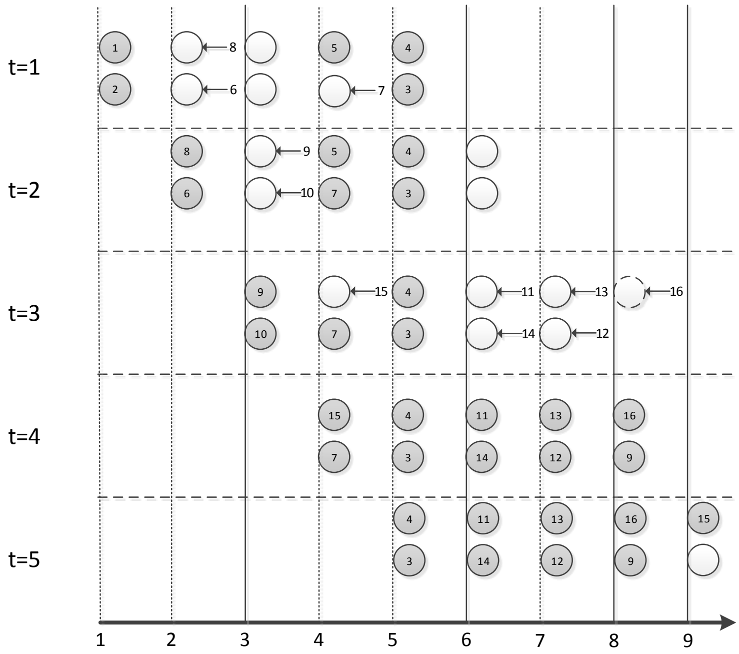
- At the beginning of day four ($t = 4$), there are no free slots available on day eight to be released to the EBS. With ten patients in the appointment queue, one additional patient must be seen in the clinic on this day. However, two of the slots are cancelled by the clinic, and so only patient 8 is seen during the day and Patient 9's appointment gets cancelled. The new advance requests during the day are flexible (as we assume) and are turned away. At the end of the day a new appointment is given to patient 9 on day 9. There will only be one slot left open to be used by same-day patients.

Figure 1 A simple illustration of evolution of the appointment schedule under FCFS policy.



Following a FCFS policy as above would lead to a schedule with no gaps which is ideal for queueing analysis. However it is not realistic as patients may not take the first available slot. Its implementation is also difficult as it would require many appointments to be rescheduled when additional visits are imposed by the allocation policy. In Figure 2, we illustrate the same system as the one in Figure 1 but with FCFS relaxed. Now patients may take any of the slots that are available in the days after their arrival days. There is also no need to change the entire schedule when additional patients are seen. For instance, on day two patient 5's appointment is brought forward by two days, assuming she is the first person in the queue who is happy to be seen earlier, but the appointments of the other patients, i.e. patients 7, 3 and 4, remain as before. Counting the numbered circles in Figures 1 and 2 shows that the sizes of appointment backlogs in the two systems are exactly the same, i.e. 5, 6, 5, 10 and 9 at the start of days 1, 2, 3, 4 and 5, respectively. This is because, although in Figure 2 there are gaps in the schedule on some days, the slots of each day are always filled before the service of that day begins, and so the length of queue changes by as much as it would under the FCFS policy. Similarly, clinics might have different strategies for rescheduling appointments for re-shows and cancellations. This would not cause any difficulty in queue length calculations as long as, when there are gaps in the schedule, the slots of each day are filled before the service of that day begins. This naturally happens in many days in clinics, and so our analysis would remain valid even if FCFS is not strictly followed.

Figure 2 A simple illustration of evolution of the appointment schedule without FCFS policy.



4. Dynamics of Appointment Queue

In this section, we view the dynamics of appointment backlog from a mathematical perspective in order to develop a model for performance evaluation of a clinic with given slot publication and dynamic allocation policies. For this purpose, we develop a state-dependent discrete bulk service queue with slot cancellation and customer no-show. The focus here is only on advance booking patients, and the objective is to obtain the steady-state distribution for the size of appointment backlog. Our queueing model is state-dependent in both service capacity and arrival process. The dependence of service capacity to the size of backlog captures the impact of the dynamic allocation policy. The state-dependence of the arrival process represents the balking process in the queue, where flexible patients who find all the slots within a given period of time specified by the scheduling window occupied, leave the system. Throughout for a non-negative discrete random variable Y, we denote its mean by μ_y, its variance by σ_y^2, and its associated probabilities by $y_j \triangleq \mathbb{P}(Y = j)$. We use the notation $X^{(i)}$ ($x^{(i)}$) to show the dependence of a random variable X (parameter x) on the size of backlog i.

Let $A^{(i)}$ denote the number of "accepted" requests, i.e. requests not turned away, for advance appointments during a time period when there are i patients in the queue in the beginning of that period before the service starts. This includes all dedicated requests as well as all flexible requests who find an appointment slot available. For now we assume this distribution is fully known but later in Section 6 we specify how it can be characterized based on the distribution for the overall number of advance requests, polling range and parameter θ. The requests may arrive at any point of time during an interval but they join the appointment queue at the end of that interval in order of their arrival. This corresponds to the formulation given in Section 3, where advance booking patients were assumed to take the first available slot in the intervals following their arrival intervals. A new arrival who finds i customers in the queue upon arrival will have a no-show probability $0 \leq \gamma^{(i)} < 1$, and subsequently each no-show will require a new appointment with a fixed probability $0 \leq \zeta \leq 1$, resulting in a re-show probability of $\delta^{(i)} \triangleq \zeta\gamma^{(i)}$. We assume $\gamma^{(i)}$ and thus $\delta^{(i)}$ increases in i.

The number of slots cancelled in an interval is represented by a random variable C, which is assumed to be independent from everything else in the model, and independent and identically distributed (i.i.d) across intervals. For simplicity, let C have a finite support $\{0,1,\ldots,\xi\}$ with $\xi \le n$. The nominal service capacity allocated to advance patients, which we refer to as "advance service capacity", is identified by the function $n^{(i)} \triangleq n + e^{(i)}$. When there are i customers in the queue at the beginning of a time interval, a batch of size i or $n^{(i)} - C$, whichever is smaller, would therefore be served during that interval. We assume no-shows remain in the queue and are served with other patients in the interval where they would have been served normally, and rejoin the backlog with probability ζ at the end of that interval.

Let X_t denote the size of appointment backlog in the beginning of a time interval t. The following recursive equation represent the evolution of appointment backlog,

$$X_{t+1} = \big(X_t - (n^{(X_t)} - C)\big)^+ + D^{(X_t)} + A^{(X_t)}, \quad t = 1,2,\ldots, \tag{2}$$

where $(x)^+ \triangleq \max\{x,0\}$ and $D^{(X_t)}$ denotes the number of re-shows at the end of period t. As explained above, we assume that no-show probabilities of arriving customers depend on the number of customers they see upon arrival in the queue. However, as we do not keep track of the size of appointment backlog at arrival epochs, we use the size of backlog at departure epochs, i.e. the end of time intervals, as a proxy as suggested in Green and Savin (2008) and Izady (2015). Note that the number of customers left behind by the first and last departing patients in a batch at the end of interval t will be $(X_t - 1)^+$ and $\big(X_t - (n^{(X_t)} - C)\big)^+$, respectively. With an increasing re-show function $\delta^{(i)}$, we take a conservative approach and assume re-show probabilities of all customers in a departing batch are the same as the first customer in the batch. This implies that the conditional random variable $D^{(X_t)} | X_t = i, C = k$ has a binomial distribution with parameters $\big(\min\{i, n^{(i)} - k\}, \alpha(i)\big)$ for $i = 0,1,\ldots$ and $k = 0,1,\ldots,\xi$ where $\alpha(i) \triangleq \delta^{(i-1)^+}$.

There is no closed from expression for the stationary queue length probabilities $x_i \triangleq \mathbb{P}(X = i) = \lim_{t\to\infty} \mathbb{P}(X_t = i)$ of the system identified by Equation (2). However, we can obtain these probabilities numerically as explained below. All the proofs are given in the Appendix.

5. Numerical Analysis of Appointment Queue

Let $\mathbf{x} \triangleq (x_0, x_1, \dots, x_m)$ be the stationary distribution for the discrete-time Markov chain characterized by Equation (2), where m is the maximum number of customers allowed in the system. We assume m is large enough so that the probability of having m customers in the system is sufficiently small. One can find the the stationary queue length probabilities by solving balance equations $\mathbf{x}\boldsymbol{\phi} = \mathbf{x}$, with $\boldsymbol{\phi} = [\phi_{ij}]$ the transition probability matrix specified below.

Proposition 1. *The transition probabilities $\phi_{ij} \triangleq \mathbb{P}(X_{t+1} = j | X_t = i)$ of the Markov chain given in (2) are*

$$\begin{aligned}\phi_{ij} = \mathbb{P}(C \leq n^{(i)} - i - 1)\alpha(i)^j \beta(i)^{i-j} & \sum_{l=max\{j-i,0\}}^{j} \binom{i}{j-l} \left(\frac{\beta(i)}{\alpha(i)}\right)^l a_l^{(i)} \\ + \alpha(i)^{j-i+n^{(i)}} \beta(i)^{i-j} & \sum_{k=n^{(i)}-i}^{\xi} \sum_{l=\max\{j-i,0\}}^{j-i-k+n^{(i)}} \binom{n^{(i)}-k}{j-i-k+n^{(i)}-l} \left(\frac{\beta(i)}{\alpha(i)}\right)^l \alpha(i)^{-k} a_l^{(i)} c_k, \quad (3)\end{aligned}$$

for $j \geq i - n^{(i)}$, and $\phi_{ij} = 0$ otherwise, where $\beta(i) \triangleq 1 - \alpha(i)$.

Once the steady state probabilities are found, one can easily obtain the desired performance metrics. Since our ultimate objective is to incorporate the queueing model into a numerical optimization model where performance metrics are calculated for a wide range of input parameters, we need to use an efficient method for calculating steady-state probabilities. The computation time for solving balance equations however grows with system size m, and thus an alternative approach based on probability generating functions (PGF's) as we explain below might be more efficient for systems with large m.

For a discrete random variable Y, define its PGF as $Y(z) \triangleq \sum_{j=0}^{\infty} y_j z^j$, which is known to be analytic for $|z| < 1$ and continuous for $|z| \leq 1$. To find the PGF for X, we make three new assumptions; A(i) the sequences $\{A^{(i)}\}_{i=0}^{\infty}$, $\{n^{(i)}\}_{i=0}^{\infty}$, and $\{\alpha^{(i)}\}_{i=0}^{\infty}$ are eventually constant, i.e. there exists positive integers h_A, h_n, and h_α such that $A^{(i)} = A^{(*)}$ for $i \geq h_A$, $n^{(i)} = n^{(*)}$ for $i \geq h_n$, and $\alpha^{(i)} = \alpha^{(*)}$ for $i \geq h_\alpha$; A(ii) $a_0^{(*)} \triangleq \mathbb{P}(A^{(*)} = 0)$ and $c_0 \triangleq \mathbb{P}(C = 0)$ are positive; and A(iii) $n^{(*)} \leq h_n$. A(i) naturally happens in practice. A(ii) may not hold in some situations but it can be fixed by assigning infinitesimally small probabilities to zero arrivals and zero cancellations and adjusting the remaining probabilities. As we shall see in Section 8, the restriction imposed by A(iii) is minimal in many practical situations.

Under A(i), A(ii), and A(iii), we must have $\mu_{A^{(*)}}/\left(1-\alpha^{(*)}\right) < n^{(*)} - \mu_C$ to achieve stability. The left side of this inequality gives the limiting value of the effective arrival rate, i.e. the average number of new accepted requests plus re-shows when the queue size i exceeds $\max\{h_A, h_n, h_\alpha\}$, and the right hand side gives the limiting value of average available capacity. The following proposition provides the PGF of the stationary queue length X.

Proposition 2. *Given $\mu_{A^{(*)}}/\left(1-\alpha^{(*)}\right) < n^{(*)} - \mu_C$, A(i), A(ii), and A(iii), the PGF of the stationary queue length X is given by*

$$X(z) = \left[z^{n^{(*)}} \sum_{i=0}^{h_n-1} A^{(i)}(z)\alpha(i,z)^i x_i \mathbb{P}(C \le n^{(i)} - i - 1) \right.$$
$$\left. + z^{n^{(*)}} \sum_{i=0}^{h-1} A^{(i)}(z) z^i x_i \sum_{k=n^{(i)}-i}^{\xi} \left(\frac{z}{\alpha(i,z)}\right)^{k-n^{(i)}} c_k - A^{(*)}(z)G(z)\sum_{i=0}^{h-1} z^i x_i \right]$$
$$/\left(z^{n^{(*)}} - A^{(*)}(z)G(z)\right), \quad (4)$$

where $h \triangleq \max\{h_A, h_n, h_\alpha\}$, $\alpha(i,z) \triangleq \beta(i) + \alpha(i)z$, and

$$G(z) \triangleq \left(1-\alpha^{(*)}+\alpha^{(*)}z\right)^{n^{(*)}} C\left(\frac{z}{1-\alpha^{(*)}+\alpha^{(*)}z}\right).$$

To obtain $X(z)$ for special cases where either $A^{(i)} = A$, $\alpha^{(i)} = \alpha$ or $n^{(i)} = n$ for all i, corresponding to situations where either accepted requests distribution, re-show probabilities, or advance service capacity is not state-dependent, we set ($A^{(*)} = A$, $h_A = 0$), ($\alpha^{(*)} = \alpha$, $h_\alpha = 0$), or ($n^{(*)} = n, h_n = n$), respectively in the equation above.

The PGF of the queue length distribution given above depends on h unknown probabilities, $x_0, x_1, \ldots, x_{h-1}$. The standard method for finding the unknown probabilities in a PGF is to solve a series of simultaneous equations obtained by substituting the zeros of the PGF denominator on or inside the unit circle in the numerator, see e.g. Kim et al. (2011). This is because the zeros of the denominator of a PGF on or inside the unit circle must be the zeros of the numerator too as otherwise the PGF would not be analytic. The Lemma below gives the number of complex zeros of the denominator of $X(z)$ on or inside the unit circle.

Lemma 1. *Given $\mu_{A^{(*)}}/\left(1-\alpha^{(*)}\right) < n^{(*)} - \mu_C$ and finite $\mu_{A^{(*)}}$, the equation*

$$z^{n^{(*)}} - A^{(*)}(z)G(z) = 0$$

has n^* complex solutions on or within the unit circle.

Hence, by the lemma above, the number of equations provided by the zeros of the denominator is n^*-1 ($z=1$ is one of the zeros which leads to a trivial equation), which combined with $X(1)=1$ would lead to $n^{(*)}$ equations. By A(iii), however, $n^{(*)} \leq h_n \leq h$, so the number of equations would not be enough for finding the unknown probabilities. This situation has rarely been encountered in the literature, the only example we are aware of being that of Powell and Humblet (1986). To resolve this, we use the first $h-n^{(*)}$ stochastic balance equations as the additional relations required. Combining all these equations leads to the following proposition.

Proposition 3. *The unknown probabilities $x_0, x_1, \ldots, x_{h-1}$ are found by solving the equation $\boldsymbol{\chi\rho} = \mathbf{y}$ for $\boldsymbol{\chi} \triangleq (x_0, x_1, \ldots, x_{h-1})$, where $\mathbf{y} \triangleq ((1-\alpha^{(*)})(n^{(*)}-\mu_C) - \mu_{A^{(*)}}, 0, \ldots, 0)$, and $\boldsymbol{\rho} \triangleq [\rho_{ij}]$ with*

$$\rho_{(i,0)} = \begin{cases} n^{(*)}(1-\alpha^{(*)}) + \beta(i)\Bigg(\mathbb{P}(C \leq n^{(i)}-i-1)(n^{(i)}-i) - n^{(i)} \\ \quad + \displaystyle\sum_{k=n^{(i)}-i}^{\xi} kc_k\Bigg) - \mu_C(1-\alpha^{(*)}) + \mu_{A^{(i)}} - \mu_{A^{(*)}}, & \text{for } i = 0,1,\ldots,h_n-1 \\ n^{(*)}(\alpha^{(i)}-\alpha^{(*)}) + \mu_C(\beta(i)+\alpha^{(*)}-1) + \mu_{A^{(i)}} - \mu_{A^{(*)}}, & \text{for } i = h_n, h_n+1, \ldots, h-1, \end{cases} \tag{5}$$

$$\rho_{(i,j)} = \begin{cases} z_j^{n^{(*)}} A^{(i)}(z_j)\Bigg(\alpha(i,z_j)^i \mathbb{P}(C \leq n^{(i)}-i-1) + \\ \displaystyle\sum_{k=n^{(i)}-i}^{\xi} z_j^{k+i-n^{(i)}} \alpha(i,z_j)^{n^{(i)}-k} c_k\Bigg) - z_j^i A^{(*)}(z_j)G(z_j), & \text{for } i = 0,1,\ldots,h_n-1 \\ z_j^i\left(A^{(i)}(z_j)\alpha(i,z_j)^{n^{(*)}} C\left(\dfrac{z_j}{\alpha(i,z_j)}\right) - A^{(*)}(z_j)G(z_j)\right) & \text{for } i = h_n, h_n+1, \ldots, h-1, \end{cases} \tag{6}$$

for $j = 1, 2, \ldots, n^{(*)} - 1$, *and*

$$\rho_{(i,j)} = \begin{cases} \phi_{(i,j-n^{(*)})} - 1 & \textit{for } i = j - n^{(*)} \\ \phi_{(i,j-n^{(*)})} & \textit{otherwise,} \end{cases} \tag{7}$$

for $j = n^{(*)}, \ldots, h-1$. $z_1, \ldots, z_{n^{(*)}-1}$ *are the complex roots of the denominator of* $X(z)$ *excluding* $z = 1$.

The complex roots $z_1, \ldots, z_{n^{(*)}-1}$ required for Equation 6 can be obtained by a software packages such as Maple. The fixed point iteration algorithm can also be used as in Kortbeek et al. (2014). The probabilities $x_0, x_1, \ldots, x_{h-1}$ fully specify the PGF $X(z)$, from which we can obtain most of the important performance metrics. One can also obtain the rest of the probabilities x_i, $i \geq h$, if needs be, by numerically inverting the PGF $X(z)$ (see Abate and Whitt 1992a,b on discrete (fast) Fourier transform method and Kim et al. 2011 on Taylor series expansion method).

The number of equations needed to be solved for the PGF approach is $h+1$, one equation for finding the complex roots of the denominator of $X(z)$ plus the simultaneous equations given in Proposition 3. For the stochastic balance equations approach on the other hand m equations must be solved. The choice between these two approaches would therefore depend on the time required for finding the complex roots of the denominator in the PGF approach as well as the value of h. The complex roots of the denominator can be found quickly if $A^{(*)}(z)$ is a polynomial function, i.e. when the random variable $A^{(*)}$ has a finite support. This naturally happens when an empirical distribution is used for representing arrivals. The computational speed would therefore depend on the value of h and m. In systems with state-dependent no-show probabilities, the value of h_α and thus h is often large so using the PGF approach may not be efficient. In contrast, in systems with constant no-show probability the value of h is typically substantially smaller than m, making the PGF approach computationally more efficient.

6. Accepted Requests Distribution

In this section, we explain how we can specify the probability mass function (p.m.f) for the number of accepted requests based on the overall number of advance requests. Assuming that the total number of advance booking requests in a time interval is given by the random variable A with a known p.m.f, we have

$$A^{(i)} = \begin{cases} A, \quad A \leq \left(f - (i-n)^+\right)^+, \\ \left(f - (i-n)^+\right)^+ + \displaystyle\sum_{j=1}^{A - \left(f-(i-n)^+\right)^+} I_j, \quad \text{otherwise}, \end{cases} \tag{8}$$

where I_j's are i.i.d random variables with Bernoulli distribution with success probability θ. In the equation above, when the total number of advance request in a time interval, A, is smaller than or equal to the total number of slots available in the following intervals (recall that advance requests cannot book slots in their arrival intervals), $(f - (i-n)^+)^+$, all appointment requests will be satisfied. When the total number of slots available is not enough, however, only the first $(f - (i-n)^+)^+$ requests plus the remaining requests that are dedicated, given by $\sum_{j=1}^{A-\left(f-(i-n)^+\right)^+} I_j$, are satisfied. Note that in the intervals where additional consultations take place due to the dynamic allocation policy, the slots freed up as a result do not show up on the EBS until the end of the interval when the entire schedule is updated, as explained in Section 3. This is why we have used n instead of $n^{(i)}$ in the equation above. If this is not the case, Equation (8) must be modified accordingly but this would make the analysis very complex. The proposition below gives the p.m.f. of $A^{(i)}$ for each i.

Proposition 4. *The pmf for $A^{(i)}$ is given by*

$$\begin{aligned} a_k^{(i)} &\triangleq \mathbb{P}(A^{(i)} = k) \\ &= \begin{cases} a_k \mathbf{1}_{\psi^{(i)}}(k) + \theta^{k-\psi^{(i)}} \displaystyle\sum_{l=\max\{k,\psi^{(i)}+1\}}^{\infty} \binom{l-\psi^{(i)}}{k-\psi^{(i)}} (1-\theta)^{l-k} a_l, & i = 0,1,\ldots,f+n-1 \\ \theta^k \displaystyle\sum_{l=k}^{\infty} \binom{l}{k} (1-\theta)^{l-k} a_l, & \textit{otherwise}, \end{cases} \end{aligned} \tag{9}$$

where $\mathbf{1}_y(x)$ is an indicator function equal to 1 for $x \leq y$, and 0 otherwise, and $\psi^{(i)} \triangleq (f - (i-n)^+)^+$.

The above characterizes $A^{(i)}$ as a state-dependent variable with distinct distributions for each of the values of $i = 0,1,\ldots,f+n-1$, and a single distribution for all values of $i \geq f+n$. As such, $h_A = f+n$, and $A^{(*)}$ has the p.m.f specified by the bottom-line equation given in (9). It is easily verified that the condition $a_0^{(*)} > 0$ is met as long as $\theta < 1$, or $\theta = 1$ and $a_0 > 0$.

7. Optimization Model

In the optimization model, we seek to identify the combination of slot publication and dynamic allocation policies that minimize the average cost of providing overtime slots whilst ensuring that the mean access time of advance booking patients and the mean number of flexible patients turned away in a time interval fall below maximum thresholds, q and b, respectively. Let S represent the random number of same-day requests in a time interval. We assume S has a finite support $\{0,1,\dots,\eta\}$, and is i.i.d. across time intervals. The clinic must provide overtime slots at a rate of o per slot when the number of same-day requests in a time period exceeds the number of slots available for them. The optimization model is

$$\begin{aligned}
\min_{(n,p,e^{(i)})} \quad & o\mathbb{E}\left[\left(\min\{X, n^{(X)}-C\}+S-r+C\right)^{+}\right] \\
\text{s.t.} \quad & \mu_W \leq q \\
& \mu_A - \mu_{A^{(X)}} \leq b,
\end{aligned} \tag{10}$$

where W is the access time of advance booking patients and $\mu_{A^{(X)}} \triangleq \mathbb{E}_{(A,X)}\left[A^{(X)}\right]$. The expressions $\min\{X, n^{(X)}-C\}+S$ and $r-C$ in the objective function calculate the total demand for and supply of slots in one time period, respectively. We refer to the first and second constraints above as the "access" and "service level" constraints, respectively.

Using the queueing model developed in Section 4 and its numerical solution in Section 5, we can find an approximate solution to the optimization problem above via numerical evaluation over a range of sensible values for n, p, and $e^{(i)}$. The following proposition shows how each term in the optimization model can be evaluated using the results of Section 5.

Proposition 5. *For the optimization model given in* (10)*,*

$$\begin{aligned}
\mathbb{E}\left[\left(\min\{X, n^{(X)}-C\}+S-r+C\right)^{+}\right] = & \sum_{i=0}^{h_n-1}\sum_{k=0}^{\xi}\sum_{j=0}^{\eta}\left(\min\{i, n^{(i)}-k\}+j-r+k\right)^{+} x_i c_k s_j \\
& + (1-\mathbb{P}(X<h_n))\sum_{j=0}^{\eta}\left(n^{(*)}+j-r\right)^{+} s_j,
\end{aligned} \tag{11}$$

$$\mu_{A(X)} = (1-\theta) \sum_{i=0}^{f+n-1} \left(\psi^{(i)} + \sum_{k=0}^{\psi^{(i)}} a_k (k - \psi^{(i)}) \right) x_i + \theta \mu_A, \tag{12}$$

and

$$\mu_W \triangleq \mathbb{E}[W] = \mu_X / \left(\mu_{A(X)} + \sum_{i=0}^{h_n - 1} \sum_{k=0}^{\xi} \alpha(i) \min\{i, n^{(i)} - k\} x_i c_k + (n^{(*)} - \mu_c) \left(\sum_{i=h_n}^{h-1} \alpha(i) x_i + \alpha^{(*)} (1 - \mathbb{P}(X \leq h-1)) \right) \right) - 1. \tag{13}$$

As illustrated in the equations above, the only queue length probabilities needed for evaluating the objective function and constraints in (10) are $x_0, x_1, \ldots, x_{h-1}$. With a PGF approach, these probabilities are obtained through the equations given in Proposition 3. μ_X is also obtained by evaluating the first derivative of PGF $X(z)$ at $z = 1$. With the stochastic balance equation approach, the entire range of probabilities are obtained, and so more complicated optimality criteria, e.g. defined based on tail probabilities rather than averages, can also be used.

8. Empirical Results

In this section, we apply the models developed in the paper to two different outpatient clinics, one an ophthalmology clinic in the UK, and the other an MRI clinic in the US as reported in study of Green and Savin (2008). The first case represents an example of a carve-out mode of delivery with a constant no-show probability and empirical distributions for urgent referrals, routine referrals, and clinic slot cancellations. The second case represents an example of advanced access delivery with delay-dependent no-show behaviour, no clinic slot cancellations, and Poisson distributions for advance and same-day requests.

8.1. A Specialty Clinic in the UK

We focus on the data obtained from a glaucoma service provided in this clinic over a one year period starting from July 2012. Using a weekly time unit, we infer empirical distributions for numbers of new routine and urgent patients referred to this service as well the slots cancelled by the clinic. As summarised in Table 1, all three distributions are over-dispersed. We did not observe a strong delay-dependent behaviour for no-show rate so a constant no-show probability $\gamma = 0.0627$

and re-scheduling probability $\zeta = 1$ obtained from the data is assumed in our analysis. Based on discussions with clinic consultants, we set $\theta = 0.90$, representing a highly dedicated demand stream with minimum substitute services available in the area. The total number of slots available for the glaucoma service is $r = 18$ slots per week. We set $o = \pounds 180$ (calculated as 1.8 times of the cost of providing a regular slot estimated at £100 by the clinic).

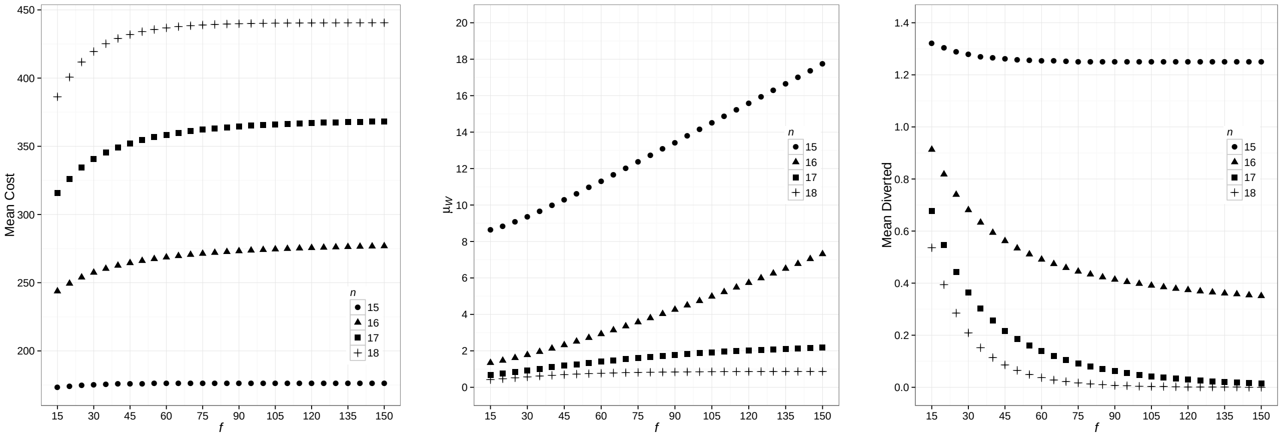
Table 1 Summary statistics for weekly referrals and slot cancellations.

Variable	Mean	Variance
Routine Referrals	13.942	48.683
Urgent Referrals	3.326	4.724
Slot Cancellations	1.098	3.050

In our first experiment, we assume $e^{(i)} = 0$ for all i so the number of advance patients seen every week is the same as the number scheduled (unless cancellations or no-shows occur). To find the optimal slot publication policy, we numerically evaluate the objective function and other measures required for the optimization model in (10) for all combinations of $n = 15, 16, 17, 18$ (we must have $n \geq 15$ to ensure stability) and $f = 15, 20, 25, \ldots, 150$. Note that instead of working directly with p, we use the total number of slots available to advance patients, f, in our experiments to have a fair comparison between different slot publication policies. Figure 3 illustrates the three major performance metrics we need for finding the optimal policies in terms of n and f. Panel (a) in this figure suggests that the average overtime cost is increasing in both n and f. The impact of n is intuitive but for f it is because with larger values of f the system becomes more congested. This reduces the average number of slots released to the EBS that remain unfilled, thus increasing the average number of required overtime slots for same-day requests. Panel (b) of Figure 3 shows that mean access time decreases (increases) with n (f), and panel (c) illustrates that the average number of patients turned away decreases with both n and f as expected.

In Table 2, the optimal policies and corresponding measures are displayed for all combinations of threshold values $q = 1, 2, \ldots, 8$ weeks and $b = 0.2, 0.4, 0.6, 0.8$ patients per week for the range of values considered for n and f. This table suggests that the optimal n is decreasing in both q and

Figure 3 **The average overtime cost (a), access time (in weeks) (b) and weekly number of patients turned away (c) as a function of n and f for the first experiment in Section 8.1.**



(a) (b) (c)

Table 2 **Optimal slot publication policies for the first experiment in Section 8.1.**

q	b	n	f	Mean Cost	μ_W	Mean Diverted	q	b	n	f	Mean Cost	μ_W	Mean Diverted
1	0.2	18	35	425	0.62	0.15	1	0.4	17	30	341	0.93	0.36
2	0.2	17	50	355	1.26	0.19	2	0.4	17	30	341	0.93	0.36
3	0.2	17	50	355	1.26	0.19	3	0.4	17	30	341	0.93	0.36
4	0.2	17	50	355	1.26	0.19	4	0.4	17	30	341	0.93	0.36
5	0.2	17	50	355	1.26	0.19	5	0.4	16	100	274	4.74	0.40
6	0.2	17	50	355	1.26	0.19	6	0.4	16	100	274	4.74	0.40
7	0.2	17	50	355	1.26	0.19	7	0.4	16	100	274	4.74	0.40
8	0.2	17	50	355	1.26	0.19	8	0.4	16	100	274	4.74	0.40
q	b	n	f	Mean Cost	μ_W	Mean Diverted	q	b	n	f	Mean Cost	μ_W	Mean Diverted
1	0.6	17	20	326	0.75	0.55	1	0.8	17	15	316	0.67	0.68
2	0.6	17	20	326	0.75	0.55	2	0.8	16	25	254	1.62	0.74
3	0.6	16	40	263	2.13	0.59	3	0.8	16	25	254	1.62	0.74
4	0.6	16	40	263	2.13	0.59	4	0.8	16	25	254	1.62	0.74
5	0.6	16	40	263	2.13	0.59	5	0.8	16	25	254	1.62	0.74
6	0.6	16	40	263	2.13	0.59	6	0.8	16	25	254	1.62	0.74
7	0.6	16	40	263	2.13	0.59	7	0.8	16	25	254	1.62	0.74
8	0.6	16	40	263	2.13	0.59	8	0.8	16	25	254	1.62	0.74

b. The optimal f is increasing in q, however it does not show a monotone behaviour with respect to b. An interesting observation is that with the optimal slot publication policy the mean number of patients diverted reaches its maximum threshold b (the small differences in Table 2 between b and "Mean Diverted" are due to step size of 5 we considered for f values), but the mean access time is typically far below its maximum q.

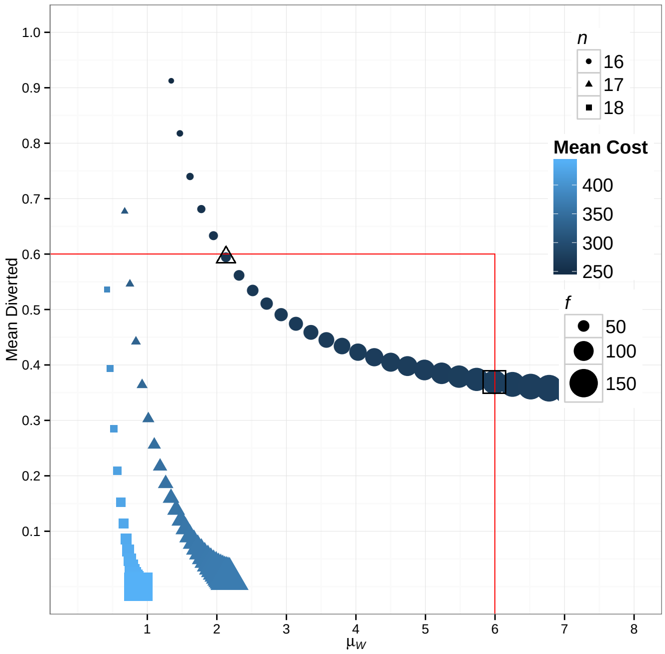
To explain this phenomenon, the average number of patients diverted is plotted against the average access time in Figure 4 for different values of n and f, with darker colour representing lower cost. For threshold values $q = 6$ and $b = 0.6$, for instance, the feasible points are the points inside the triangular area specified by the red lines and the two axes. Clearly the points associated with $n = 16$ are the lowest cost points in this region. Among them, the point corresponding to $f = 40$ (highlighted by a triangle) is the point with smallest f that satisfies the service level constraint. The mean access time for this point is 2.13 weeks, substantially below the six week threshold. Increasing f further, up until $f = 125$ (highlighted by a square), would make the access time closer to its threshold but at the expense of increasing cost. As such the optimal policy would be the one with $n = 16$ and $f = 40$, which does not reach the six week maximum wait. In general, due to the discrete nature of n, there will not necessarily be a feasible point on the top right corner of the triangular area where both mean access time and mean diverted reach their maximum thresholds. On the other hand, since both mean access time and mean cost increase with f, the optimal point would be a point on the top border of the triangular area where mean diverted takes its maximum, rather than on the right border where mean access time takes its maximum.

We use the empirical findings above to devise a search process that reduces the search space for finding the optimal n and f. For each value of n, starting with the smallest possible f, we evaluate the objective function and constraints for increasing values of f until the service level constraint is satisfied for the first time. If the access constraint is also satisfied, then we have a candidate optimum, whose cost must be compared with other candidates obtained by repeating the same process with other values of n. Otherwise there is not any feasible point for the corresponding n value. This leads to a considerable reduction in computation time for finding the optimal policy.

In the second experiment, we examine allocation policies jointly with publication policies. We consider publication policies (n, f) with $n = 14, 15, 16, 17$ and $f = 15, 20, \ldots, 150$, and restrict our experiments to bi-level allocation policies in the form below

$$e^{(i)} = \begin{cases} 0, & i < h_n, \\ \\ e^{(*)}, & i \geq h_n, \end{cases} \tag{14}$$

Figure 4 **Mean number of patients diverted versus μ_W for different values of n and f for the first experiment in Section 8.1.**

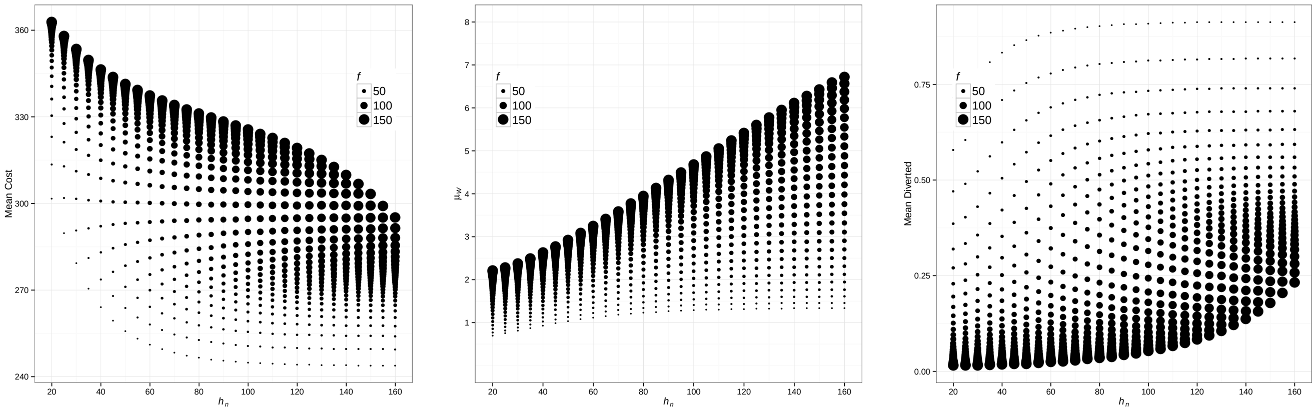


where $e^{(*)} = 1, 2, 3, 4$ (corresponding to $n^{(*)} = 15, 16, 17, 18$) and $h_n = 20, 25, 30, \ldots, 160$. Notice that the only restriction imposed by A(iii) in our example is that h_n should not be smaller that 18.

Figure 5 plots the performance metrics as a function of f and h_n for the case with $n = 16$ and $e^{(*)} = 1$. This figure shows that the impact of f on performance metrics is the same as in the first experiment. Increasing h_n clearly leads to longer access times and larger numbers of patients turned away but smaller overtime cost.

Table 3 displays the optimal joint policies and the associated performance metrics. For each combination of q and b, this table also gives the percentage saving gained by using a joint optimal policy, compared to the corresponding case with only the optimal slot publication policy (that was illustrated in Table 2). This suggests improvements up to 19 percent are likely to arise as a result of employing simple bi-level allocation policies, which in all scenarios except one require only one additional consultation per week when the queue length goes beyond h_n. Table 3 also suggests that both mean access time and mean patients diverted reach values close to their maximum thresholds. This is because, as illustrated in Figure 6 for the specific case of $n = 15$ and $e^{(*)} = 1$,

Figure 5 The average overtime cost (a), access time (in weeks) (b), and weekly number of patients turned away (c) as a function of f and h_n, for dynamic allocation policies with $n = 16$ and $e^{(*)} = 1$, for the second experiment in Section 8.1.



(a) (b) (c)

the introduction of a third parameters, i.e. h_n, leads to a much wider spread of the feasible points, compared to the case with only two parameters n and f. The final observation is that the optimal slot publication policies reported in Table 3 are different from corresponding policies given in Table 2, suggesting that sequential optimization of slot publication and dynamic allocation policies may lead to suboptimal solutions.

Based on empirical findings above, we devise a search process for finding the optimal f and h_n for given n and $e^{(*)}$, assuming a bi-level allocation policy. Starting with smallest possible f and h_n, we evaluate the objective function and constraints for increasing values of f until the service level constraint is met for the first time. If the access constraint is also met we have a candidate optimum, whose cost must be compared with other candidates obtained by repeating the same process with larger values of h_n. Otherwise, the search stops as increasing h_n further would only make the access time larger.

8.2. An MRI Clinic in the US

Based on the data provided in Green and Savin (2008), the daily average number of patients requesting a clinic appointment is $0.008v$ where v is the panel size of the clinic. There exists a total

Table 3 **Optimal joint slot publication and dynamic allocation policies for the second experiment in Section 8.1.**

q	b	(n,f)	h_n	$e^{(*)}$	Mean Cost	μ_W	Mean Diverted	Savings
1	0.2	(17,40)	45	1	371	0.92	0.18	13%
2	0.2	(16,60)	45	1	323	1.85	0.20	9%
3	0.2	(16,85)	75	1	314	3.00	0.18	12%
4	0.2	(16,100)	95	1	308	3.81	0.18	13%
5	0.2	(16,120)	120	1	304	4.90	0.19	14%
6	0.2	(16,140)	140	1	303	5.91	0.18	14%
7	0.2	(16,150)	150	1	303	6.44	0.18	14%
8	0.2	(16,150)	150	1	303	6.44	0.18	14%
1	0.4	(16,30)	20	1	330	0.94	0.39	3%
2	0.4	(16,45)	55	1	293	1.78	0.40	14%
3	0.4	(16,60)	80	1	284	2.54	0.40	17%
4	0.4	(16,85)	130	1	277	3.88	0.40	19%
5	0.4	(15,105)	25	1	273	4.99	0.40	1%
6	0.4	(15,120)	45	1	271	5.97	0.40	1%
7	0.4	(15,130)	60	1	270	6.70	0.40	2%
8	0.4	(15,145)	75	1	270	7.69	0.40	2%
1	0.6	(15,25)	25	2	297	0.99	0.55	9%
2	0.6	(16,40)	90	1	267	1.98	0.57	18%
3	0.6	(15,50)	35	1	253	2.72	0.59	4%
4	0.6	(15,70)	55	1	252	3.94	0.57	4%
5	0.6	(15,80)	70	1	248	4.72	0.60	5%
6	0.6	(15,95)	85	1	248	5.73	0.60	5%
7	0.6	(15,105)	95	1	248	6.40	0.59	5%
8	0.6	(15,120)	110	1	248	7.42	0.59	5%
1	0.8	(16,15)	30	1	279	0.81	0.77	12%
2	0.8	(15,30)	30	1	240	1.94	0.76	5%
3	0.8	(15,45)	50	1	233	2.93	0.76	8%
4	0.8	(15,60)	65	1	233	3.88	0.74	8%
5	0.8	(15,65)	75	1	227	4.37	0.80	11%
6	0.8	(15,65)	75	1	227	4.37	0.80	11%
7	0.8	(15,65)	75	1	227	4.37	0.80	11%
8	0.8	(15,65)	75	1	227	4.37	0.80	11%

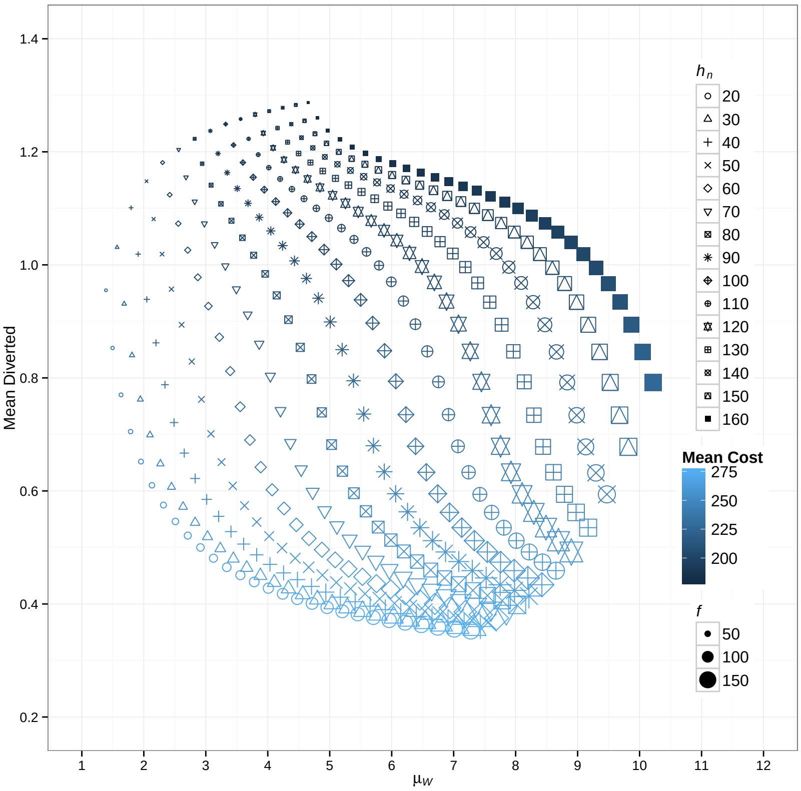
Saving column represents the percentage saving with the optimal joint policy compared with the best corresponding case with only the optimal slot publication policy

of 20 regular slots per day, and the no-show probability of a patient who finds i patients in the appointment backlog upon request for an appointment is estimated by the exponential function

$$\gamma(i) = 0.31 - (0.31 - 0.01)e^{-i/1000}. \tag{15}$$

All no-shows are assumed to reschedule an appointment, i.e. $\zeta = 1$. Considering all patients in a single group, Izady (2015) use the state-independent versions of the queueing model discussed in Section 4 to find the smallest panel sizes under which 75% of the patients can be offered a same-day appointment with various distributions for arrival requests. (The 75% figure reflects the 25% of

Figure 6 **Mean number of patients diverted versus μ_W for $n = 15$, $e^{(*)} = 1$, and different values of f and h_n for the second experiment in Section 8.1.**



patients who opt to be seen on a future day as observed in Murray and Tantau 2000.) In particular, they suggest when appointment requests follow a Poisson distribution, the minimum panel size that achieves the 75% same-day probability is 2337. However, they do not specify how the capacity should be divided between same-day and advance booking requests. They do not also consider the impact of scheduling window.

For $v = 2337$, we specify how many slots of each day must be reserved for same-day requests, and how long in advance the remaining slots must be released to the EBS for advance patients. We do not consider dynamic allocation policies in this section. In order to apply the model developed in this paper, we assume same-day and advance booking requests follow independent Poisson distributions with averages $0.008 \times 2337 \times 0.75 = 14.022$ and $0.008 \times 2337 \times 0.25 = 4.674$ patients per day, respectively. We assume there is no slot cancellation by the clinic, and following Green et al. (2006), we set $o = \$100$. We investigate three different scenarios with $\theta = 0.1, 0.5$, and 0.9.

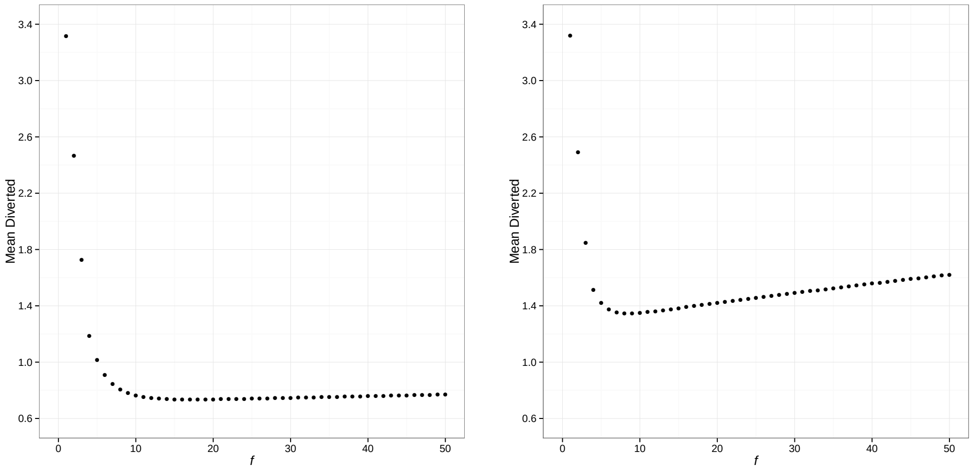
Simple analysis show that in order to achieve stability, we must have $n \geq 1$, $n \geq 4$, and $n \geq 7$, for each of these scenarios, respectively. We consider slot publication policies with n from the minimum possible up to 20 slots per day and $f = 1, 2, \ldots, 50$.

To show the impact of delay-dependent no-show probability, in panel (a) of Figure 7 we plot the average number of patients diverted versus f for $n = 4$ and $\theta = 0.1$. The plot shows that as f increases, the mean patients diverted initially decreases due to more slots being released to the EBS. But then as a result of larger queues, no-show probabilities increase leading to more patients coming back for further appointments, thus reducing the slots available for new patients and increasing the mean numbers diverted. The extent of this increase is more significant with larger no-show probabilities as illustrated in panel (b) of Figure 7, where the no-show probability function observed in Gallucci et al. (2005) as given below

$$\gamma(i) = 0.51 - (0.51 - 0.15)e^{-i/180}. \tag{16}$$

is used with other parameters the same as in panel (a).

Figure 7 **Average daily number of patients diverted versus f for $n = 4$, $\theta = 0.1$ and no-show function given in (a) Equation (15) and (b) Equation (16), for the MRI case study in Section 8.2.**



(a)

(b)

Despite different behaviour of the mean diverted measure with respect to f as observed above, we can apply the same search process for finding optimal values of n and f as the one proposed

Table 4 **Optimal slot publication policies and corresponding metrics for the MRI case study in Section 8.2.**

θ	q	b	(n,f)	Mean Cost	μ_W	Mean Diverted
0.1	1	0.01	(6,14)	106	0.19	0.008
	2, 3, 4, 5, 6, 7	0.01	(5,35)	99	1.5	0.01
	1	0.02	(6,13)	106	0.19	0.013
	2, 3, 4, 5, 6, 7	0.02	(5,28)	99	1.33	0.019
	1	0.03	(6,12)	106	0.18	0.020
	2, 3, 4, 5, 6, 7	0.03	(5,24)	98	1.20	0.030
	1	0.03	(6,11)	105	0.17	0.03
	2, 3, 4, 5, 6, 7	0.04	(5,22)	98	1.12	0.037
0.5	1	0.01	(6,13)	106	0.19	0.009
	2, 3, 4, 5, 6, 7	0.01	(5,34)	99	1.5	0.01
	1	0.02	(6,12)	106	0.18	0.015
	2, 3, 4, 5, 6, 7	0.02	(5,27)	99	1.33	0.02
	1	0.03	(6,11)	106	0.18	0.023
	2, 3, 4, 5, 6, 7	0.03	(5,24)	98	1.23	0.027
	1	0.03	(6,10)	105	0.16	0.03
	2, 3, 4, 5, 6, 7	0.04	(5,21)	98	1.12	0.038
0.9	1, 2, 3, 4, 5, 6, 7	0.01	(7,9)	111	0.06	0.007
	1, 2, 3, 4, 5, 6, 7	0.02	(7,8)	111	0.06	0.014
	1, 2, 3, 4, 5, 6, 7	0.03	(7,7)	111	0.05	0.027
	1, 2, 3, 4, 5, 6, 7	0.04	(7,7)	111	0.05	0.027

for systems with constant no-show probability in Section 8.1. This is because, for a given n, both mean cost and mean access time still increase with f, making the point with smallest possible f that satisfies both constraints a candidate optimum. The resulting optimal policies and relevant metrics are shown in Table 4 for $q = 1, 2, \ldots, 7$ days and $b = 0.04, 0.08, 0.12, 0.16$ patients per day. These results suggest that the optimal n decreases with q and increases with θ. As a consequence, it seems reasonable to allocate 25% of the daily capacity to advance booking requests, corresponding to their proportion out of the total demand, and reserve the rest for same-day booking as long as the maximum access threshold is not too tight and θ is not too large. For large values of θ and small values of q, however, more slots should be allocated to advance patients. Similar to the first experiment in Section 8.1, we also observe that optimal policies lead to mean patients diverted close to maximum thresholds and mean access times below maximum thresholds.

9. Conclusions

Online booking facilities have become an integral part of modern outpatient clinics. Apart from giving patients a greater choice over the location and time of their treatments, these systems give providers a great deal of flexibility in managing their supply and demand. To utilize this flexibility, an integrated approach that addresses demand and capacity planning decisions in a unified manner, considering both static mid-term decisions as well as dynamic day-to-day adjustments is developed in this paper. We did not include panel size as a decision variable explicitly in our policy formulation, but it can be easily considered by adapting the accepted request distribution based on the size of the panel.

We found the (near-)optimal values of the two policies by enumerating over a range of reasonable values, relying on a state-dependent queueing model for efficient computation of performance metrics. The probability generating function approach we developed for performance evaluation resulted in significant time savings, compared to the more common approach of using balance equations. In particular, we observed up to 20-fold reduction in calculation time with the PGF approach, enabling us to explore about forty different policy combinations in every hour. A search process was also proposed to further reduce the time required for finding the optimal policies.

The implementation of slot publication policy, as we formulated, is straightforward in almost all EBS's. The implementation of dynamic allocation policy would however require expediting some patients. This might prove challenging in practice, in particular if FCFS is to be followed strictly. Without FCFS the number of patients expedited in each time period would be limited to the number of additional consultations required for that period. As an alternative approach to expediting patients, clinics may fulfill the additional consultations by releasing the same number of additional slots to the online system. Since patients cannot book slots for the same period, these additional slots must be allocated to the next one or two periods, assuming such functionality is available in the EBS. Some of the efficiency of the allocation policy would then be achieved without the need for rescheduling patients.

Acknowledgments

Appendix. Proofs of Lemmas and Propositions

A. Proof of Proposition 1

For $j \geq i - n^{(i)}$, we have

$$\begin{aligned}
\mathbb{P}(X_{t+1}=j|X_t=i) &= \mathbb{P}\left[(i-(n^{(i)}-C))^+ + A^{(i)} + D^{(i)} = j|X_t=i\right] \\
&= \sum_{k=0}^{\xi} \mathbb{P}\left[A^{(i)} + D^{(i)} = j-(i+k-n^{(i)})^+|X_t=i, C=k\right] c_k \\
&= \sum_{k=0}^{\xi}\sum_{l=0}^{\infty} \mathbb{P}(D^{(i)} = j-(i+k-n^{(i)})^+ - l|X_t=i, C=k, A^{(i)}=l) c_k a_l^{(i)} \\
&= \sum_{k=0}^{\xi}\sum_{l=0}^{\infty} \binom{\min\{i, n^{(i)}-k\}}{j-(i+k-n^{(i)})^+ - l} \\
&\times \alpha(i)^{\left[j-(i+k-n^{(i)})^+ - l\right]} \beta(i)^{\left[\min\{i,n^{(i)}-k\}-j+(i+k-n^{(i)})^+ + l\right]} a_l^{(i)} c_k \\
&= \sum_{k=0}^{n^{(i)}-i-1} \sum_{l=\max\{j-i,0\}}^{j} \binom{i}{j-l} \alpha(i)^{j-l} \beta(i)^{i-j+l} a_l^{(i)} c_k \\
&+ \sum_{k=n^{(i)}-i}^{\xi} \sum_{l=\max\{j-i,0\}}^{j-i-k+n^{(i)}} \binom{n^{(i)}-k}{j-i-k+n^{(i)}-l} \alpha(i)^{j-i-k+n^{(i)}-l} \beta(i)^{i+l-j} a_l^{(i)} c_k \\
&= \mathbb{P}(C \leq n^{(i)}-i-1) \alpha(i)^j \beta(i)^{i-j} \sum_{l=max\{j-i,0\}}^{j} \binom{i}{j-l} \left(\frac{\beta(i)}{\alpha(i)}\right)^l a_l^{(i)} \\
&+ \alpha(i)^{j-i+n^{(i)}} \beta(i)^{i-j} \sum_{k=n^{(i)}-i}^{\xi} \sum_{l=\max\{j-i,0\}}^{j-i-k+n^{(i)}} \binom{n^{(i)}-k}{j-i-k+n^{(i)}-l} \left(\frac{\beta(i)}{\alpha(i)}\right)^l \alpha(i)^{-k} a_l^{(i)} c_k,
\end{aligned}$$

where $\alpha(i) = \delta^{(i-1)^+}$ and $\beta(i) = 1-\alpha(i)$. For $j < i - n^{(i)}$ the transition probability is equal to 0 as the number served in a time interval cannot exceed $n^{(i)}$. □

B. Proof of Proposition 2

From (2), we have

$$\begin{aligned}
\mathbb{E}[z^{X_{t+1}}] &= \mathbb{E}\left[z^{(X_t-(n^{(X_t)}-C))^+ + A^{(X_t)} + D^{(X_t)}}\right] \\
&= \sum_{i=0}^{\infty} A^{(i)}(z) \mathbb{E}\left[z^{(i+C-n^{(i)})^+ + D^{(i)}}|X_t=i\right] \mathbb{P}(X_t=i) \\
&= \sum_{i=0}^{\infty} A^{(i)}(z) \sum_{k=0}^{\xi} \mathbb{E}\left[z^{(i+k-n^{(i)})^+ + D^{(i)}}|X_t=i, C=k\right] c_k \mathbb{P}(X_t=i) \\
&= \sum_{i=0}^{\infty} A^{(i)}(z) \sum_{k=0}^{\xi} z^{(i+k-n^{(i)})^+} \alpha(i,z)^{\min\{i,n^{(i)}-k\}} c_k \mathbb{P}(X_t=i) \\
&= \sum_{i=0}^{\infty} A^{(i)}(z) \left[\sum_{k=0}^{n^{(i)}-i-1} \alpha(i,z)^i c_k \mathbb{P}(X_t=i) + \sum_{k=n^{(i)}-i}^{\xi} z^{i+k-n^{(i)}} \alpha(i,z)^{n^{(i)}-k} c_k \mathbb{P}(X_t=i)\right] \\
&= \sum_{i=0}^{\infty} A^{(i)}(z) \alpha(i,z)^i \mathbb{P}(X_t=i) \mathbb{P}(C \leq n^{(i)}-i-1) \\
&+ \sum_{i=0}^{\infty} A^{(i)}(z) z^i \mathbb{P}(X_t=i) \sum_{k=n^{(i)}-i}^{\xi} \left(\frac{z}{\alpha(i,z)}\right)^{k-n^{(i)}} c_k,
\end{aligned}$$

where $\alpha(i,z)=\beta(i)+\alpha(i)z$. Splitting the first and second sums in the last equality above at $i=h_n$ and $i=h$, respectively, where $h=\max\{h_A,h_n,h_\alpha\}$, we arrive at

$$\begin{aligned}\mathbb{E}[z^{X_{t+1}}]=\sum_{i=0}^{h_n-1}A^{(i)}(z)\alpha(i,z)^i\mathbb{P}(X_t=i)\mathbb{P}(C\le n^{(i)}-i-1)\\ +\sum_{i=h_n}^{\infty}A^{(i)}(z)\alpha(i,z)^i\mathbb{P}(X_t=i)\mathbb{P}(C\le n^{(*)}-i-1)\\ +\sum_{i=0}^{h-1}A^{(i)}(z)z^i\mathbb{P}(X_t=i)\sum_{k=n^{(i)}-i}^{\xi}\left(\frac{z}{\alpha(i,z)}\right)^{k-n^{(i)}}c_k\\ +A^{(*)}(z)G(z)\frac{1}{z^{n^{(*)}}}\sum_{i=h}^{\infty}z^i\mathbb{P}(X_t=i),\end{aligned}\quad(17)$$

where

$$G(z)=\left(1-\alpha^{(*)}+\alpha^{(*)}z\right)^{n^{(*)}}C\left(\frac{z}{1-\alpha^{(*)}+\alpha^{(*)}z}\right).$$

Note that for the last term in (17) we have used the assumption $n^{(*)}\le h_n$, resulting in $k=n^{(i)}-i=n^{(*)}-i\le h_n-i\le h-i\le 0$ for $i\ge h$. The same assumption gives $\mathbb{P}(C\le n^{(*)}-i-1)=0$ for $i\ge h_n$, so by removing the second sum in (17) and taking the limit of both sides as $t\to\infty$, we obtain

$$\begin{aligned}X(z)=\sum_{i=0}^{h_n-1}A^{(i)}(z)\alpha(i,z)^ix_i\mathbb{P}(C\le n^{(i)}-i-1)+\sum_{i=0}^{h-1}A^{(i)}(z)z^ix_i\sum_{k=n^{(i)}-i}^{\xi}\left(\frac{z}{\alpha(i,z)}\right)^{k-n^{(i)}}c_k\\ +A^{(*)}(z)G(z)\frac{1}{z^{n^{(*)}}}\left(X(z)-\sum_{i=0}^{h-1}z^ix_i\right).\end{aligned}$$

where $X(z)=\sum_{i=0}^{\infty}x_iz^i$. Solving the equation above for $X(z)$ yields

$$\begin{aligned}X(z)=\Bigg[z^{n^{(*)}}\sum_{i=0}^{h_n-1}A^{(i)}(z)\alpha(i,z)^ix_i\mathbb{P}(C\le n^{(i)}-i-1)\\ +z^{n^{(*)}}\sum_{i=0}^{h-1}A^{(i)}(z)z^ix_i\sum_{k=n^{(i)}-i}^{\xi}\left(\frac{z}{\alpha(i,z)}\right)^{k-n^{(i)}}c_k-A^{(*)}(z)G(z)\sum_{i=0}^{h-1}z^ix_i\Bigg]\\ \Big/\left(z^{n^{(*)}}-A^{(*)}(z)G(z)\right).\end{aligned}$$

For special cases where the arrival distribution, no-show probability, or advance service capacity do not depend on the number of customers in the system, the proof is obvious. The only trick is to set $h_n=n$ for the last case to satisfy A(iii).

□

C. Proof of Lemma 1

We show $z^{n^{(*)}} - U(z)$ has $n^{(*)}$ zeros on and within the unit circle, where $U(z) \triangleq A^{(*)}(z)G(z)$. We use Theorem (3.2) given in Adan et al. (2006). By this theorem, we need to show (i) $U(z)$ is a PGF, (ii) $U(0) > 0$, (iii) $U(z)$ is differentiable at at $z = 1$, and (iv) $U'(1) < n^{(*)}$. To show that $U(z)$ is a PGF, we re-write $G(z)$ as $G(z) = \sum_{i=0}^{n^{(*)}} g_i z^i$ where

$$g_i = \sum_{k=0}^{\min\{i,\xi\}} \binom{n^{(*)} - k}{i-k} (\alpha^{(*)})^{i-k} (1-\alpha^{(*)})^{n^{(*)}-i} c_k$$

for $i = 0, 1, \ldots, n^{(*)}$. This is obtained by expanding and re-arranging the terms in $G(z)$ as below

$$\begin{aligned}
G(z) = & \\
&= (1-\alpha^{(*)} + \alpha^{(*)} z)^{n^{(*)}} C\left(\frac{z}{1-\alpha^{(*)} + \alpha^{(*)} z}\right) \\
&= \sum_{k=0}^{\xi} z^k c_k \left(1-\alpha^{(*)} + \alpha^{(*)} z\right)^{n^{(*)}-k} \\
&= \sum_{k=0}^{\xi} \sum_{j=0}^{n^{(*)}-k} z^k c_k \binom{n^{(*)}-k}{j} (\alpha^{(*)} z)^j (1-\alpha^{(*)})^{n^{(*)}-k-j} \\
&= \sum_{k=0}^{\xi} \sum_{j=0}^{n^{(*)}-k} c_k \binom{n^{(*)}-k}{j} (\alpha^{(*)})^j (1-\alpha^{(*)})^{n^{(*)}-k-j} z^{k+j} \\
&= \sum_{k=0}^{\xi} \sum_{i=k}^{n^{(*)}} \binom{n^{(*)}-k}{i-k} (\alpha^{(*)})^{i-k} (1-\alpha^{(*)})^{n^{(*)}-i} c_k z^i \\
&= \sum_{i=0}^{\xi} \sum_{k=0}^{i} \binom{n^{(*)}-k}{i-k} (\alpha^{(*)})^{i-k} (1-\alpha^{(*)})^{n^{(*)}-i} c_k z^i + \sum_{i=\xi+1}^{n^{(*)}} \sum_{k=0}^{\xi} \binom{n^{(*)}-k}{i-k} (\alpha^{(*)})^{i-k} (1-\alpha^{(*)})^{n^{(*)}-i} c_k z^i.
\end{aligned}$$

Now Since $G(1) = 1$, and $g_i \geq 0$, $G(z)$ is a PGF, and thus $U(z)$ is also a PGF. For (ii), note that $U(0) = a_0^{(*)}(1-\alpha^{(*)})^{n^{(*)}} c_0$ which is positive since $a_0^{(*)}$, $1-\alpha^{(*)}$ and c_0 are all positive by assumption. Since $G(z)$ is always differentiable as it is a finite sum, condition (iii) holds when $\mu_{A^{(*)}}$ is finite. For (iv), note that $U'(1) = \mu_{A^{(*)}} + \mu_C(1-\alpha^{(*)}) + n^{(*)}\alpha^{(*)} < n^{(*)}$ by stability condition. □

D. Proof of Proposition 3

Equation $\boldsymbol{\chi}\boldsymbol{\rho} = \mathbf{y}$ is obtained from combining three sets of equations: (i) the equation derived from $X(1) = 1$ which forms the first column in matrix $\boldsymbol{\rho}$ as given in (5), (ii) the $n^{(*)} - 1$ equations obtained by replacing the zeros of the denominator in (4) that lie on or within the unit circle in the numerator, which form the next $n^{(*)} - 1$ columns in matrix $\boldsymbol{\rho}$ as given in (6), (iii) the first $h - n^{(*)}$ stochastic balance equations that form the last $h - n^{(*)}$ columns in matrix $\boldsymbol{\rho}$ as given in (7).

Equation set (i) is derived from $X(1)=1$ using the l'Hopital's rule. Let $X(z)=N(z)/D(z)$. We first evaluate the derivative of $N(z)$ at $z=1$

$$\begin{aligned}\frac{d}{dz}N(z)\,|_{z=1}&\\ &=\sum_{i=0}^{h_n-1}x_i\mathbb{P}(C\le n^{(i)}-i-1)\left(n^{(*)}+\mu_{A^{(i)}}+i\alpha(i)\right)+\sum_{i=0}^{h-1}\sum_{k=n^{(i)}-i}^{\xi}x_ic_k\left(n^{(*)}+\beta(i)(k-n^{(i)})+i+\mu_{A^{(i)}}\right)\\ &-\sum_{i=0}^{h-1}x_i\left(\mu_{A^{(*)}}+\mu_G+i\right).\end{aligned}$$

Since $\mu_G=\frac{d}{dz}G(z)|_{z=1}=n^{(*)}\alpha^{(*)}+(1-\alpha^{(*)})\mu_C$, we get

$$\begin{aligned}\frac{d}{dz}N(z)\,|_{z=1}&\\ &=\sum_{i=0}^{h_n-1}x_i\mathbb{P}(C\le n^{(i)}-i-1)\left(n^{(*)}+\mu_{A^{(i)}}+i\alpha(i)\right)+\sum_{i=0}^{h-1}\sum_{k=n^{(i)}-i}^{\xi}x_ic_k\left(n^{(*)}+\beta(i)(k-n^{(i)})+i+\mu_{A^{(i)}}\right)\\ &-\sum_{i=0}^{h-1}x_i\left(\alpha^{(*)}(n^{(*)}-\mu_C)+i+\mu_C+\mu_{A(*)}\right).\end{aligned}$$

Next we evaluate the derivative of $D(z)$ at $z=1$

$$\frac{d}{dz}N(z)\,|_{z=1}=n^{(*)}-\left(\mu_{A^{(*)}}+\mu_C(1-\alpha^{(*)})+n^{(*)}\alpha^{(*)}\right)=(1-\alpha^{(*)})(n^{(*)}-\mu_C)-\mu_{A^{(*)}}.$$

Now from $X(1)=N'(1)/D'(1)=1$, we obtain

$$\begin{aligned}\sum_{i=0}^{h_n-1}x_i\mathbb{P}(C\le n^{(i)}-i-1)\left(n^{(*)}+\mu_{A^{(i)}}+i\alpha(i)\right)+\sum_{i=0}^{h-1}\sum_{k=n^{(i)}-i}^{\xi}x_ic_k\left(n^{(*)}+\beta(i)(k-n^{(i)})+i+\mu_{A^{(i)}}\right)&\\ -\sum_{i=0}^{h-1}x_i\left(\alpha^{(*)}(n^{(*)}-\mu_C)+i+\mu_C+\mu_{A(*)}\right)=(1-\alpha^{(*)})(n^{(*)}-\mu_C)-\mu_{A^{(*)}}.&\end{aligned}$$

Combining the terms with coefficient x_i, we have

$$\begin{aligned}&\sum_{i=0}^{h_n-1}\Bigg(\mathbb{P}(C\le n^{(i)}-i-1)\left(n^{(*)}+\mu_{A^{(i)}}+i\alpha(i)\right)+\left(n^{(*)}-\beta(i)n^{(i)}+i+\mu_{A^{(i)}}\right)\mathbb{P}(C\ge n^{(i)}-i)\\ &\qquad+\sum_{k=n^{(i)}-i}^{\xi}\beta(i)kc_k-\alpha^{(*)}(n^{(*)}-\mu_C)-i-\mu_C-\mu_{A(*)}\Bigg)x_i\\ &+\sum_{i=h_n}^{h-1}\Bigg(\left(n^{(*)}-\beta(i)n^{(i)}+i+\mu_{A^{(i)}}\right)\mathbb{P}(C\ge n^{(i)}-i)+\sum_{k=n^{(i)}-i}^{\xi}\beta(i)kc_k-\alpha^{(*)}(n^{(*)}-\mu_C)-i-\mu_C-\mu_{A(*)}\Bigg)x_i\\ &\qquad=(1-\alpha^{(*)})(n^{(*)}-\mu_C)-\mu_{A^{(*)}}\end{aligned}$$

Since $\mathbb{P}(C\le n^{(i)}-i-1)+\mathbb{P}(C\ge n^{(i)}-i)=1$, and $\mathbb{P}(C\ge n^{(i)}-i)=1$ for $i\ge h_n$ (due to A(iii)), we have

$$\sum_{i=0}^{h_n-1}\Bigg(n^{(*)}+\mu_{A^{(i)}}+\mathbb{P}(C\le n^{(i)}-i-1)\left(i\alpha(i)+\beta(i)n^{(i)}-i\right)+i-\beta(i)n^{(i)}$$

$$
\left. + \sum_{k=n^{(i)}-i}^{\xi} \beta(i)kc_k - \alpha^{(*)}(n^{(*)} - \mu_C) - i - \mu_C - \mu_{A(*)} \right) x_i
$$

$$
+ \sum_{i=h_n}^{h-1} \left(n^{(*)} - \beta(i)n^{(*)} + i + \mu_{A^{(i)}} + \sum_{k=n^{(*)}-i}^{\xi} \beta(i)kc_k - \alpha^{(*)}(n^{(*)} - \mu_C) - i - \mu_C - \mu_{A(*)} \right) x_i
$$

$$
= (1 - \alpha^{(*)})(n^{(*)} - \mu_C) - \mu_{A^{(*)}}
$$

Simplifying the above we arrive at

$$
\sum_{i=0}^{h_n-1} \left(n^{(*)}(1 - \alpha^{(*)}) + \beta(i) \left(\mathbb{P}(C \leq n^{(i)} - i - 1)(n^{(i)} - i) - n^{(i)} + \sum_{k=n^{(i)}-i}^{\xi} kc_k \right) \right.
$$

$$
\left. - \mu_C(1 - \alpha^{(*)}) + \mu_{A^{(i)}} - \mu_{A(*)} \right) x_i
$$

$$
+ \sum_{i=h_n}^{h-1} \left(n^{(*)}(\alpha^{(i)} - \alpha^{(*)}) + \mu_C(\beta(i) + \alpha^{(*)} - 1) + \mu_{A^{(i)}} - \mu_{A(*)} \right) x_i = (1 - \alpha^{(*)})(n^{(*)} - \mu_C) - \mu_{A^{(*)}}
$$

Coefficients of x_i in the above equation constitute the first column in matrix $\boldsymbol{\rho}$, and the right hand side value will be the first element of vector $\mathbf{y}$.

For equation set (ii), we replace the zeros z_j of the denominator in (4) in the numerator to obtain

$$
z_j^{n^{(*)}} \sum_{i=0}^{h_n-1} A^{(i)}(z_j)\alpha(i,z_j)^i x_i \mathbb{P}(C \leq n^{(i)} - i - 1) + z_j^{n^{(*)}} \sum_{i=0}^{h-1} A^{(i)}(z_j) z_j^i x_i \sum_{k=n^{(i)}-i}^{\xi} \left(\frac{z_j}{\alpha(i,z_j)} \right)^{k-n^{(i)}} c_k
$$

$$
- A^{(*)}(z_j)G(z_j) \sum_{i=0}^{h-1} z_j^i x_i = 0,
$$

for $j = 1, 2, \ldots, n^{(*)} - 1$. Combining the coefficients of x_i in the above equation we have

$$
\sum_{i=0}^{h_n-1} \left(z_j^{n^{(*)}} A^{(i)}(z_j) \left(\alpha(i,z_j)^i \mathbb{P}(C \leq n^{(i)} - i - 1) + \sum_{k=n^{(i)}-i}^{\xi} z_j^{i+k-n^{(i)}} \alpha(i,z_j)^{n^{(i)}-k} c_k \right) - z_j^i A^{(*)}(z_j)G(z_j) \right) x_i
$$

$$
+ \sum_{i=h_n}^{h-1} z_j^i \left(A^{(i)}(z_j)\alpha(i,z_j)^{n^{(*)}} C\left(\frac{z_j}{\alpha(i,z_j)} \right) - A^{(*)}(z_j)G(z_j) \right) x_i = 0
$$

Coefficients of x_i in the above equation constitute column 2 to $n^{(*)}$ in matrix $\boldsymbol{\rho}$, and the corresponding elements in vector $\mathbf{y}$ will be zero.

For equation set (iii), we use the first $h - n^{(*)}$ stochastic balance equations as below

$$
\sum_i x_i \phi_{(i, j-n^{(*)})} - x_{j-n^{(*)}} = 0,
$$

for $j = n^{(*)}, \ldots, h-1$, where transition probabilities ϕ_{ij} are given in Equation (3). Since $\phi_{ij} = 0$ for $i > j + n^{(i)}$,

$$
\sum_{i,\, i \leq j - n^{(*)} + n^{(i)}} x_i \phi_{(i, j-n^{(*)})} - x_{j-n^{(*)}} = 0.
$$

From this, the coefficient of x_i for $i \neq j - n^{(*)}$ will be $\phi_{(i,j-n^{(*)})}$, and $\phi_{(i,j-n^{(*)})} - 1$ otherwise. These coefficients form the last $h - n^{(*)}$ columns in matrix $\boldsymbol{\rho}$. □

E. Proof of Proposition 4

Conditioning on A, we have

$$\begin{aligned}
a_k^{(i)} &\triangleq \mathbb{P}(A^{(i)} = k)\\
&= \sum_{l=0}^{\infty} \mathbb{P}\left(A^{(i)} = k | A = l\right) a_l\\
&= \sum_{l=0}^{\psi^{(i)}} \mathbb{P}(A = k, A = l) + \sum_{l=\psi^{(i)}+1}^{\infty} \mathbb{P}\left(\sum_{j=1}^{l-\psi^{(i)}} I_j = k - \psi^{(i)}\right) a_l\\
&= a_k \mathbf{1}_{\psi^{(i)}}(k) + \sum_{l=\psi^{(i)}+1}^{\infty} \mathbb{P}\left(\text{Binomial}\left(l-\psi^{(i)}, \theta\right) = k - \psi^{(i)}\right) a_l\\
&= a_k \mathbf{1}_{\psi^{(i)}}(k) + \sum_{l=\psi^{(i)}+1}^{\infty} \binom{l-\psi^{(i)}}{k-\psi^{(i)}} \theta^{k-\psi^{(i)}} (1-\theta)^{l-k} a_l\\
&= a_k \mathbf{1}_{\psi^{(i)}}(k) + \theta^{k-\psi^{(i)}} \sum_{l=\max\{k, \psi^{(i)}+1\}}^{\infty} \binom{l-\psi^{(i)}}{k-\psi^{(i)}} (1-\theta)^{l-k} a_l
\end{aligned}$$

for all nonnegative i and k where $\psi^{(i)} = (f - (i-n)^+)^+$. For $i \geq f + n$, the above simplifies to

$$a_k^{(i)} = \theta^k \sum_{l=k}^{\infty} \binom{l}{k} (1-\theta)^{l-k} a_l.$$

□

F. Proof of Proposition 5

For (11),

$$\begin{aligned}
&\mathbb{E}\left[\left(\min\{X, n^{(X)} - C\} + S - r + C\right)^+\right]\\
&= \sum_{i=0}^{h_n-1} \mathbb{E}\left[\left(\min\{i, n^{(i)} - C\} + S - r + C\right)^+ | X = i\right] x_i + \sum_{i=h_n}^{\infty} \mathbb{E}\left[\left(\min\{i, n^{(*)} - C\} + S - r + C\right)^+ | X = i\right] x_i\\
&= \sum_{i=0}^{h_n-1} \sum_{k=0}^{\xi} \sum_{j=0}^{\eta} \left(\min\{i, n^{(i)} - k\} + j - r + k\right)^+ x_i c_k s_j + \sum_{i=h_n}^{\infty} \mathbb{E}\left[\left(n^{(*)} - C + S - r + C\right)^+ | X = i\right] x_i,
\end{aligned}$$

where the second term of the last equality is because $n^{(*)} \leq h_n$ and so $\min\{i, n^{(*)} - C\} = n^{(*)} - C$ for $i \geq h_n$.

Due to independence of S and X, we then have

$$\begin{aligned}
&\mathbb{E}\left[\left(\min\{X, n^{(X)} - C\} + S - r + C\right)^+\right] =\\
&= \sum_{i=0}^{h_n-1} \sum_{k=0}^{\xi} \sum_{j=0}^{\eta} \left(\min\{i, n^{(i)} - k\} + j - r + k\right)^+ x_i c_k s_j + \mathbb{E}\left[\left(n^{(*)} + S - r\right)^+\right] (1 - \mathbb{P}(X < h_n)),
\end{aligned}$$

Expanding the expected value term above yields the result.

For (12),

$$\begin{aligned}
\mathbb{E}[A^{(X)}] &= \sum_{i=0}^{\infty} \mathbb{E}[A^{(i)}|X=i]x_i \\
&= \sum_{i=0}^{\infty} \left(\mathbb{E}[A^{(i)}, A \le \psi^{(i)}] + \mathbb{E}[A^{(i)}, A > \psi^{(i)}]\right) x_i \\
&= \sum_{i=0}^{\infty} \left(\mathbb{E}[A, A \le \psi^{(i)}] + \mathbb{E}[\psi^{(i)} + \sum_{j=1}^{A-\psi^{(i)}} I_j, A > \psi^{(i)}]\right) x_i \\
&= \sum_{i=0}^{f+n-1} \left(\mathbb{E}[A, A \le \psi^{(i)}] + \mathbb{E}[\psi^{(i)} + \sum_{j=1}^{A-\psi^{(i)}} I_j, A > \psi^{(i)}]\right) x_i \\
&+ \sum_{i=f+n}^{\infty} \left(\mathbb{E}[A, A \le 0] + \mathbb{E}[\sum_{j=1}^{A} I_j, A > 0]\right) x_i \\
&= \sum_{i=0}^{f+n-1} \left(\sum_{k=0}^{\psi^{(i)}} k a_k + \sum_{k=\psi^{(i)}+1}^{\infty} \left[\psi^{(i)} + (k-\psi^{(i)})\theta\right] a_k\right) x_i + \sum_{i=f+n}^{\infty}\sum_{k=1}^{\infty} k\theta a_k x_i.
\end{aligned}$$

Simplifying the above, we have

$$\begin{aligned}
\mathbb{E}[A^{(X)}] &= \sum_{i=0}^{f+n-1} \left(\sum_{k=0}^{\psi^{(i)}} k a_k + \psi^{(i)}(1-\theta) \sum_{k=\psi^{(i)}+1}^{\infty} a_k + \theta \sum_{k=\psi^{(i)}+1}^{\infty} k a_k\right) x_i + \theta\mu_A \sum_{i=f+n}^{\infty} x_i \\
&= \sum_{i=0}^{f+n-1} \left(\sum_{k=0}^{\psi^{(i)}} k a_k + \psi^{(i)}(1-\theta) \sum_{k=\psi^{(i)}+1}^{\infty} a_k + \theta \left(\mu_A - \sum_{k=0}^{\psi^{(i)}} k a_k\right)\right) x_i + \theta\mu_A \sum_{i=f+n}^{\infty} x_i \\
&= \sum_{i=0}^{f+n-1} \left((1-\theta)\sum_{k=0}^{\psi^{(i)}} k a_k + \psi^{(i)}(1-\theta) \sum_{k=\psi^{(i)}+1}^{\infty} a_k + \theta\mu_A\right) x_i + \theta\mu_A \sum_{i=f+n}^{\infty} x_i \\
&= (1-\theta) \sum_{i=0}^{f+n-1} \left(\sum_{k=0}^{\psi^{(i)}} k a_k + \psi^{(i)} \sum_{k=\psi^{(i)}+1}^{\infty} a_k\right) x_i + \theta\mu_A \sum_{i=0}^{f+n-1} x_i + \theta\mu_A \sum_{i=f+n}^{\infty} x_i \\
&= (1-\theta) \sum_{i=0}^{f+n-1} \left(\psi^{(i)} + \sum_{k=0}^{\psi^{(i)}} a_k(k-\psi^{(i)})\right) x_i + \theta\mu_A,
\end{aligned}$$

for $\psi^{(i)} = (f-(i-n)^+)^+$.

For (13), we apply the Little's law with average number in the system, μ_X, and effective arrival rate, $\mathbb{E}\left[A^{(X)} + D^{(X)}\right] = \mu_{A^{(X)}} + \mathbb{E}[D^{(X)}]$, to obtain the average number of time intervals a patient spends in the system, which subtracted by one gives the average waiting time. The expected number of re-shows, $\mathbb{E}[D^{(X)}]$, is calculated as below.

$$\begin{aligned}
\mathbb{E}[D^{(X)}] &= \sum_{i=0}^{\infty}\sum_{k=0}^{\xi} \mathbb{E}[D^{(i)}|X=i, C=k] x_i c_k \\
&= \sum_{i=0}^{\infty}\sum_{k=0}^{\xi} \alpha(i) \min\{i, n^{(i)} - k\} x_i c_k
\end{aligned}$$

$$\begin{aligned}
&= \sum_{i=0}^{h_n-1}\sum_{k=0}^{\xi}\alpha(i)\min\{i,n^{(i)}-k\}x_ic_k + \sum_{i=h_n}^{\infty}\sum_{k=0}^{\xi}\alpha(i)(n^{(*)}-k)x_ic_k \\
&= \sum_{i=0}^{h_n-1}\sum_{k=0}^{\xi}\alpha(i)\min\{i,n^{(i)}-k\}x_ic_k + \sum_{i=h_n}^{\infty}\alpha(i)x_i\left(\sum_{k=0}^{\xi}(n^{(*)}-k)c_k\right) \\
&= \sum_{i=0}^{h_n-1}\sum_{k=0}^{\xi}\alpha(i)\min\{i,n^{(i)}-k\}x_ic_k + (n^{(*)}-\mu_c)\left(\sum_{i=h_n}^{h-1}\alpha(i)x_i + \sum_{i=h}^{\infty}\alpha^{(*)}x_i\right) \\
&= \sum_{i=0}^{h_n-1}\sum_{k=0}^{\xi}\alpha(i)\min\{i,n^{(i)}-k\}x_ic_k + (n^{(*)}-\mu_c)\left(\sum_{i=h_n}^{h-1}\alpha(i)x_i + \alpha^{(*)}(1-\mathbb{P}(X\le h-1))\right)
\end{aligned}$$

□

References

Abate, Joseph, Ward Whitt. 1992a. The fourier-series method for inverting transforms of probability distributions. *Queueing Systems. Theory and Applications* **10**(1-2) 5–88.

Abate, Joseph, Ward Whitt. 1992b. Numerical inversion of probability generating functions. *Operations Research Letters* **12**(4) 245–251.

Adan, I. J. B. F., J. S. H. van Leeuwaarden, E. M. M. Winands. 2006. On the application of rouch's theorem in queueing theory. *Operations Research Letters* **34**(3) 355–360.

Ayvaz, Nur, Woonghee Tim Huh. 2010. Allocation of hospital capacity to multiple types of patients. *Journal of Revenue and Pricing Management* **9**(5) 386–398.

Cayirli, Tugba, Emre Veral. 2003. Outpatient scheduling in health care: A review of literature. *Production and Operations Management* **12**(4) 519–549.

Cayirli, Tugba, Kum Khiong Yang, Ser Aik Quek. 2012. A universal appointment rule in the presence of no-shows and walk-ins. *Production and Operations Management* **21**(4) 682–697.

Creemers, Stefan, Marc Lambrecht. 2010. Queueing models for appointment-driven systems. *Annals of Operations Research* **178**(1) 155–172.

Dixon, Anna, Ruth Robertson, Roland Bal. 2010. The experience of implementing choice at point of referral: a comparison of the netherlands and england. *Health Economics, Policy and Law* **5** 295–317.

Dobson, Gregory, Sameer Hasija, Edieal J. Pinker. 2011. Reserving capacity for urgent patients in primary care. *Production and Operations Management* **20**(3) 456–473.

Feldman, Jacob, Nan Liu, Huseyin Topaloglu, Serhan Ziya. 2014. Appointment scheduling under patient preference and no-show behavior. *Operations Research* **62**(4) 794–811.

Gallucci, G., W. Swartz, F. Hackerman. 2005. Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatric Services* **56**(3) 344–6.

Gerchak, Yigal, Diwakar Gupta, Mordechai Henig. 1996. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science* **42**(3) 321–334.

Gocgun, Yasin, Archis Ghate. 2012. Lagrangian relaxation and constraint generation for allocation and advanced scheduling. *Computers & Operations Research* **39**(10) 2323–2336.

Green, Linda V., Sergei Savin. 2008. Reducing delays for medical appointments: A queueing approach. *Operations Research* **56**(6) 1526–1538.

Green, Linda V., Sergei Savin, Ben Wang. 2006. Managing patient service in a diagnostic medical facility. *Operations Research* **54**(1) 11–25.

Gupta, Diwakar, Brian Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions* **40**(9) 800–819.

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

Izady, N. 2015. Appointment capacity planning in specialty clinics: A queueing approach. *Operations Research* **63**(4) 916–930.

Jiang, Houyuan, Zhan Pang, Sergei Savin. 2012. Performance-based contracts for outpatient medical services. *Manufacturing & Service Operations Management* **14**(4) 654–669.

Kim, Nam, Mohan Chaudhry, Bong Yoon, Kilhwan Kim. 2011. Inverting generating functions with increased numerical precision a computational experience. *Journal of Systems Science and Systems Engineering* **20**(4) 475–494.

Klassen, Kenneth J, Reena Yoogalingam. 2009. Improving performance in outpatient appointment services with a simulation optimization approach. *Production and Operations Management* **18**(4) 447–458.

Koeleman, Paulien M., Ger M. Koole. 2012. Optimal outpatient appointment scheduling with emergency arrivals and general service times. *IIE Transactions on Healthcare Systems Engineering* **2**(1) 14–30.

Kortbeek, Nikky, Maartje E. Zonderland, Aleida Braaksma, Ingrid M. H. Vliegen, Richard J. Boucherie, Nelly Litvak, Erwin W. Hans. 2014. Designing cyclic appointment schedules for outpatient clinics with scheduled and unscheduled patient arrivals. *Performance Evaluation* **80**(0) 5–26.

Liu, N. 2015. Optimal choice for appointment scheduling window under patient no-show behavior. *Production and Operations Management* **forthcoming**.

Liu, Nan, Serhan Ziya, Vidyadhar G. Kulkarni. 2010. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing & Service Operations Management* **12**(2) 347–364.

Min, Daiki, Yuehwern Yih. 2014. Managing a patient waiting list with time-dependent priority and adverse events. *RAIRO - Operations Research* **48** 53–74.

Murray, M., D. M. Berwick. 2003. Advanced access: reducing waiting and delays in primary care. *Journal of American Medical Association* **289**(8) 1035–40.

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

Patrick, Jonathan, Martin L. Puterman, Maurice Queyranne. 2008. Dynamic multipriority patient scheduling for a diagnostic resource. *Operations Research* **56**(6) 1507–1525.

Pope, Catherine, Jon Banks, Chris Salisbury, Val Lattimer. 2008. Improving access to primary care: eight case studies of introducing advanced access in england. *Journal of Health Services Research & Policy* **13**(1) 33–39.

Powell, Warren B., Pierre Humblet. 1986. The bulk service queue with a general control strategy: Theoretical analysis and a new computational procedure. *Operations Research* **34**(2) 267–275.

Qu, Xiuli, Ronald L. Rardin, Julie Ann S. Williams, Deanna R. Willis. 2007. Matching daily healthcare provider capacity to demand in advanced access scheduling systems. *European Journal of Operational Research* **183**(2) 812–826.

Robinson, Lawrence W., Rachel R. Chen. 2010. A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing & Service Operations Management* **12**(2) 330–346.

Truong, Van-Anh. 2015. Optimal advance scheduling. *Management Science* **61**(7) 1584–1597.

Zacharias, Christos, Michael Pinedo. 2014. Appointment scheduling with no-shows and overbooking. *Production and Operations Management* **23**(5) 788–801.

Zocdoc. 2015. About us. Retrieved April 04, 2015. URL `http://www.zocdoc.com/aboutus`.