An Integrated Approach to Demand and Capacity Planning in Outpatient Clinics

Navid Izady*

Cass Business School, City, University of London, UK

Abstract

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. A compact policy formulation is proposed that incorporates major operational levers suggested in the literature. It combines a slot publication policy, which specifies the pattern under which slots are released to the booking system, with an expediting policy that adjusts the daily workload of advance patients. Relying on a wide range of numerical experiments, a heuristic search method is developed for finding the joint publication and expediting policies, minimizing the cost of overtime slots whilst ensuring a waiting and an access constraint is met. Several managerial insights are derived using a combination of illustrative and real data, highlighting the importance of taking an integrated approach towards the operational levers captured by our policy formulation.

Keywords: OR in health services, Outpatient Clinics, Demand and Capacity Planning, Queues

1. Introduction

Outpatient medical facilities must typically serve 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, but also to those that have implemented "advanced access". In the former, a few slots in each day are reserved for patients with urgent medical needs and the rest is available to routine patients for advance booking (Murray and Berwick 2003). In the latter, although the primary objective is to offer every patient a same-day appointment regardless of urgency, certain patient groups such as commuters highly value the flexibility to book appointments for future days (Pope et al. 2008). With both delivery modes, clinics must therefore decide how much of

*Corresponding author
Email address: navid.izady@city.ac.uk (Navid Izady)

their daily capacity should be allocated to "advance booking" patients, and how much be left open in anticipation of "same-day" demand. Clinics may also decide on certain days to serve more pre-booked patients than originally scheduled for those days by "expediting" some patients, i.e. bringing their appointments forward, aiming to alleviate the excessive backlogs caused by a temporary surge (decline) in demand (supply). In addition to these capacity planning decisions, clinics may exercise some control over the demand streams. Adjusting the appointment scheduling window, i.e. how far in advance a patient can schedule an appointment with a provider, is an important way of achieving this especially when there is less flexibility in panel size selection (Liu 2015). In this paper, we develop a methodology that guides clinics in making such strategic capacity and demand planning decisions in an integrated manner, and investigate the efficiency gains of such integration.

We assume advance booking patients schedule their appointments through an electronic booking system (EBS), and same-day patients must be served on the day they arise. There is a wide range of EBS's used in primary and specialty clinics in different countries, e.g. *ZocDoc* in the US (Zocdoc 2015), and *ZorgDomein* in the Netherlands (Dixon et al. 2010). A prime example of an EBS implemented on a large scale is that of the *e-Referral Service* (e-RS) (NHS Digital 2017) deployed in the UK National Health Service (NHS), which has motivated this study. It enables routine patients referred to specialty clinics to book their first outpatient appointment online. The e-RS is linked to providers' digital records containing information about their appointment slots, including the timings, the clinicians providing them, and whether they are publishable on the e-RS on not. The free and publishable time slots on each day are released to the e-RS a given number of days in advance as specified by the appointment scheduling window (or the "polling range" as referred to in the e-RS 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, and second the patient books the most convenient slot from a menu of available slots displayed by the e-RS for her chosen provider. The slots not released to the e-RS will be offered to urgent referrals and walk-in patients on the day they arise. Most other EBS's enjoy the same basic functionalities as explained above for the e-RS.

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 (due to, e.g., 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 e-RS, for instance, the patient can click on the "Defer to Provider" option to inform the provider, who would subsequently contact the patient to book her an appointment on a day typically beyond the polling range. 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 and potentially longer delays 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 propose a novel and compact policy formulation that combines a *slot publication* policy with a backlog-dependent *expediting* 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 may be viewed as a combination of same-day reservation level and appointment scheduling window studied separately in the literature. The expediting policy, on the other hand, specifies the number of additional pre-booked patients the clinic must serve in each time period (by bringing some appointments forward), depending on the size of backlog at the start of the period. It is motivated by the "advance scheduling with expediting" paradigm proposed by Truong and Ruzal-Shapiro (2014).

Implementation of the slot publication policy is straightforward as EBS's typically have the flexibility of adjusting both dimensions. Implementing the expediting policy is more of a challenge as not all patients may agree to be expedited. Those patients, as argued in Truong and Ruzal-Shapiro (2014), can simply opt out. Others however may see it as an opportunity to be seen earlier, possibly on a day that would not have been available when they originally booked their appointments. Therefore, it is unlikely to be difficult to find the required number, which as we show later in our case studies rarely exceeds one patient per time period, when there is a large backlog (which is when the policy is activated). There is also some administrative cost for the clinic to reschedule patients. Our models enable the clinics to quantify the benefits of expediting policy, hence implement it only if the benefits outweigh the costs. Besides, appointment rescheduling is already a common practice in many clinics, e.g. for filling the slots vacated by patients' cancellations (Truong and Ruzal-Shapiro 2014) or backfilling the slots remaining empty on each day (NHS Digital 2016).

The combination of slot publication and expediting 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 daily number of required overtime slots whilst ensuring that advance patients' access and waiting time requirements are met. Overtime slots are necessary when the total number of advance and same-day patients on a day exceeds the number of regular slots. The access requirement restricts the average number of advance flexible patients forced to switch provider as a result of no slots being available online. The waiting time constraint ensures the average time between the request for an appointment and the earliest available slot does not exceed a given threshold.

To search for the optimal expediting and publication policies, first we develop a queueing model that represents the evolution of appointment backlog given specific policies. A major feature of our queueing model is that it does not require patients to take the earliest available slot. In fact, in clinics supported by EBS's patients may take any of the slots displayed on the system based on their preferences. This along with patient expediting violates the first-come first-serve (FCFS) discipline, and also creates "holes" in the backlog. We show that our model is still accurate in these circumstances as long as a mild condition is met. As such our model is able to capture the dynamics of backlog under EBS's without detailed information on patients' preferences. Next, relying on extensive numerical experiments, we partially characterize the structure of optimal policies, leading to the development of a fast heuristic search for systems with a bi-level expediting policy. The accuracy and reliability of policies obtained from the heuristic are confirmed through comparison with complete enumeration and simulation.

Using our models, we conduct a series of numerical experiments to derive managerial insights. The first insight is that there is substantial value in adopting an integrated approach towards the three operational levers captured by slot publication and expediting policies. In particular, we observe that setting the two dimensions of slot publication policy in a joint rather than sequential manner is likely to result in large efficiency savings. Further, we show that applying the expediting policy without fully revising the publication policy is less likely to create improvement. The second insight is that deployment of a backlog-dependent expediting policy would enable the clinics to release fewer slots to the EBS per unit of time but over a longer period. This reduces the overtime cost, as compared to the publication policy alone, whilst giving patients more choice over appointment days. The scale of reductions in overtime cost is likely to be large in a carve-out clinic where the majority of patients are advance booking. The administrative burden is also small as we observe that the optimal policy typically involves only one or two expedited visits when activated, and the frequency of its activation is small when the corresponding savings are large. In a typical advanced

access clinic, on the other hand, the savings from expediting policy are much smaller, especially when the ratio of advance booking patients is 25% as cited in Murray and Tantau (2000).

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 scheduling appointments in a single day. Examples include Koeleman and Koole (2012) and Cayirli et al. (2012). Multi-day scheduling, on the other hand, is concerned with allocating appointment requests arising in each day to future days, see Truong (2015). Our study is at a more strategic level as it seeks to provide a unifying framework for demand and capacity planning in mid to long-term periods based on predicted demand and supply patterns.

Most relevant to our study are the papers by Qu et al. (2007), Green and Savin (2008), Dobson et al. (2011), Liu and Ziya (2014), and Liu (2015). Given demand distribution and no-show probabilities, Qu et al. (2007) find the optimal proportion of slots held open for same-day visits so that the average number of daily consultations is maximized. Their focus is on a single day and so patients' waiting time is not captured by their model. Dobson et al. (2011) study the same problem but over multiple days (so that patients' waiting time can also be evaluated) and use a revenue maximization objective function. They do not, however, include no-shows in their model. Green and Savin (2008) identify the patient panel size in the context of advanced access so that a given percentage of patients can be offered a same-day appointment. They provide the first analytical model incorporating the impact of delay-dependent no-shows. Liu and Ziya (2014) simplify the models proposed in Green and Savin (2008) by excluding the impact of no-shows rescheduling their appointments, and jointly optimize panel size and overbooking level decisions. Applying models similar to those of Liu and Ziya (2014), Liu (2015) show that the optimal choice of the appointment window can lead to substantial efficiency gains when other demand management mechanisms are not available. We contribute to this literature by introducing the slot publication policy, capturing the joint impact of same-day reservation level and appointment scheduling window for the first time. In contrast with the aforementioned literature where patients' waiting time and/or numbers turned away are embedded in the objective function as cost elements, we also provide a more practical optimization framework by including these two measures as constraints. Moreover, following the reality of online booking systems we consider the possibility of some patients being offered an appointment directly

by the clinic when they cannot find an empty slot online.

A new paradigm named as advance scheduling with expediting is introduced in Truong and Ruzal-Shapiro (2014). In this paradigm, patients are given appointment dates at the time of request but may be expedited later depending on future arrivals. Using a dynamic programming formulation, Truong and Ruzal-Shapiro (2014) derive the structural properties of the optimal policy and propose an approximate algorithm for finding a policy matching these properties. Although our expediting policy is in the same spirit as that of Truong and Ruzal-Shapiro (2014), there is a fundamental difference in our modeling approach in that we seek to find the optimal expediting policy in anticipation of future demand, while the decisions in Truong and Ruzal-Shapiro (2014) are made dynamically over time. Our long-term view has enabled us to investigate the impact of expediting in conjunction with publication policy. We also provide insight on the size and activation frequency of expediting policy, and shed some light on the situations where it is likely to create efficiency savings.

Our heuristic algorithm requires efficient evaluation of major performance metrics for the appointment backlog. A number of queueing models are developed in the literature for this purpose, see, e.g. Green and Savin (2008), Creemers and Lambrecht (2010) and Kortbeek et al. (2014). We choose the discrete-time framework proposed in Izady (2015) due to its flexibility and better fit to the reality of outpatient clinics; it works with arbitrary demand distributions; captures the stochastic variability in supply of appointments caused by provider's slot cancellations; and incorporates the impact of delay-dependent no-shows (who may reschedule their appointments). See Green and Savin (2008) and Izady (2015) for empirical evidence on these features. As a result, our work is the first study that captures the collective impact of these features in joint optimization of slot publication and expediting policies. Our main methodological contribution is replacing the FCFS assumption widely made in the literature, in particular in all the papers cited above, with a milder condition that is likely to be valid in many clinics.

Compared with the study of Izady (2015), our work goes beyond pure performance evaluation by proposing a heuristic search for finding the publication and expediting policies in the presence of same-day demand. We also generalize the queueing model in Izady (2015) by making it state-dependent in both service capacity and arrival process so that the joint impact of publication and expediting policies can be evaluated. Furthermore, our work includes a series of carefully chosen numerical experiments generating useful management insight. Finally, we verify the reliability of the results obtained from the queueing model using simulation, where patient choice and some other features not considered in the queueing model are explicitly taken into account. The main limitation

of our work stems from its numerical nature. The structural properties upon which the heuristic search is built are derived through numerical experiments, and are also limited to bi-level expediting policies. This is however essential for studying integrated demand and capacity planning decisions while taking into account a wide range of realistic features.

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), e.g. days, numbered $1, 2, 3, \ldots$, 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 one of the available slots in the intervals *following* their arrival interval through the EBS. Advance booking patients who have already scheduled an appointment but not yet served in the clinic form the appointment backlog (or appointment queue).

Let (n, f) denote the slot publication policy of the clinic, where $n \leq r$ determines the number of slots pre-allocated to advance booking patients in each interval, and $f = pn$, where p is the polling range (We use the terms scheduling window and polling range interchangeably throughout.) If all f slots allocated to the EBS are filled, an advance booking patient is assumed to become a dedicated (flexible) patient with probability θ $(1-\theta)$, independently of everything else in the system. Following Jiang et al. (2012), we assume dedicated patients are booked into the first available slot beyond the scheduling window by the clinic, and flexible patients switch to a different provider.

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 by bringing their appointments forward. The number of *additional* advance patients served in a time period is specified by the expediting policy $e^{(i)}$, where i is the size of backlog in the beginning of that period before the service starts. Throughout we use the notation $x^{(i)}$ to show the dependence of a parameter x on the size of backlog i at the start of a period just before the service starts. Motivated by the advance scheduling literature suggesting that "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." (Truong 2015, p. 3), we assume that $e^{(i)}$ increases in i.

We assume during each time interval, a random number of slots is cancelled by the clinic. As pointed out in Murray (2007), providers' short-notice slot cancellations result in some variability in the supply of appointments, which could be even larger than the variability in demand for appointments. They are due to external factors such as clinicians' delays and absences (caused by sickness, attendance in professional meetings or serving patients in other clinics) or unavailability of equipments. If the cancelled slots have already been booked, new appointments should be scheduled for the affected patients. On the other hand, some advance patients (whose appointments are not cancelled by the clinic) may not show up for their appointments at all, or cancel their appointments too late for new patients to be replaced. We refer to both cases as patient no-show, and in line with empirical observations made in some clinics, e.g. as reported in Green and Savin (2008) and Gallucci et al. (2005), allow the the rate of no-shows to increase with patients' waiting time in the backlog. No-shows may or may not reschedule a new appointment.

All same-day patients turn up for their consultations. They are seen in open regular slots, plus overtime slots if needed. Regular slots remain open if they are not cancelled by the clinic, nor allocated to advance patients as a result of normal or expedited visits. Following Dobson et al. (2011), we assume all open regular slots will be used when same-day demand exists. In addition, although it may 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, we suppress that level of detail and assume all no-show slots will be wasted. This is also the assumption followed in Dobson et al. (2011).

The operational 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 clinic contacts patients with appointments in future periods to invite $e^{(i)}$ additional patients to have their visits in the current period; (iii) during period t, the clinic visits advance patients who turn up for their appointments (if their slots are not cancelled by the clinic), new patients take one of the slots available on the EBS (and beyond that if they are dedicated), and same-day patients are seen in the slots left open and overtime slots if necessary; (iv) at the end of period t, appointments are booked for no-shows who require a new appointment as well as those whose appointments were cancelled by the clinic.

To make our models work in the presence of patient choice and patient expediting, we make the following assumption.

Assumption 1. All the slots of a time period are filled before the service of that period starts as

long as any of the slots in future periods are booked.

Our interviews with booking managers in various clinics of our partner hospital suggest that in most clinics slots are usually filled before the start of each working day. The rescheduling of no-shows and cancellations, who are usually booked into the earliest available slots, also helps filling up the holes in the backlog. Moreover, in clinics with short polling range and/or highly variable demand where slots are more likely to remain empty, the clinics are advised to backfill them by expediting other patients so as to minimize their impact on system performance (NHS Digital 2016). As such, the assumption above is expected to hold in many clinics. It replaces the FCFS assumption typically made in the literature.

Our objective is to identify the combination of slot publication and expediting policies that minimizes the requirement to overtime slots whilst ensuring that sufficient access along with reasonably short waiting time is provided to advance booking patients. For given slot publication and expediting policies $(n, f, e^{(i)})$, let $O(n, f, e^{(i)})$ and $T(n, f, e^{(i)})$ denote the random numbers of required overtime slots and flexible patients turned away, respectively, per period in steady state. Let $W(n, f, e^{(i)})$ be the corresponding shortest waiting time offered to advance patients, i.e. the number of time units between the request for an appointment and the earliest available slot. The optimal policies are then identified through the optimization model

$$\begin{aligned} \min_{(n,f,e^{(i)})} \quad & \mathbb{E}\left[O(n, f, e^{(i)})\right] \\ \text{s.t.} \quad & \mathbb{E}\left[W(n, f, e^{(i)})\right] \leq q, \\ & \mathbb{E}\left[T(n, f, e^{(i)})\right] \leq b, \\ & n, f, e^{(i)} \in \mathbb{Z},\ e^{(i)} \text{ increasing in } i,\ e^{(i)} \leq (i-n)^+, \end{aligned} \tag{1}$$

where $\mathbb{Z}$ is the set of nonnegative integers, $q > 0$ ($b > 0$) is the maximum threshold for mean shortest waiting time offered (mean patients diverted per period), and $(x)^+ = \max\{x, 0\}$. We refer to the first (second) constraint in the model above as the waiting (access) constraint. The last constraint on $e^{(i)}$ is to ensure the number of expedited visits in each interval does not exceed the number of available patients in the following intervals. Note that we allow $e^{(i)} > r - n$, i.e. expedited visits can use overtime slots.

The optimization model in (1) is more relevant for practical applications than the models prevalent in the literature where all metrics are unified in a single cost-based objective function. It is because in these models cost values need to be identified for every day a patient is made to wait and/or for every flexible patient forced to switch provider, see e.g. Truong (2015) and Liu (2015).

The difficulties of finding the realistic estimates for such cost elements would hinder the application of models in practice.

To characterize the objective function in (1), let S, A and C represent the random numbers of same-day requests, advance booking requests, and provider slot cancellations, respectively, in a time interval. We assume that S and C have finite supports $\{0, 1, \ldots, s\}$ and $\{0, 1, \ldots, c\}$, respectively, with $c \leq r$. We assume S, A and C are mutually independent variables that are also independent and identically distributed (i.i.d) across time intervals with known distributions. Given $(n, f, e^{(i)})$, let X denote the resulting steady-state size of appointment backlog at the beginning of a time interval before service begins. We suppress the dependence of X on publication and expediting policies for brevity in our notations. The objective function in (1) is then represented as

$$\mathbb{E}\left[O(n, f, e^{(i)})\right] = \mathbb{E}\left[\left(\min\{X, (n + e^{(X)} - C)^+\} + S - r + C\right)^+\right], \tag{2}$$

where $\min\{X, (n + e^{(X)} - C)^+\} + S$ and $r - C$ give the total demand for and supply of slots in one time period, respectively. To evaluate the distribution of X as needed in above, as well as $\mathbb{E}\left[W(n, f, e^{(i)})\right]$ and $\mathbb{E}\left[T(n, f, e^{(i)})\right]$ in (1), we propose a model for the dynamics of appointment backlog in the next section, followed by its numerical analysis in the following section.

4. Dynamics of Appointment Queue

The focus here is only on advance booking patients as same-day requests do not influence the backlog. We develop a state-dependent discrete bulk service queue with slot cancellation and customer no-show to represent the evolution of backlog. Our queueing model is bulk service as a batch of customers is served in each interval; see Izady (2015) for further details. It is also a state-dependent model in service capacity, arrival process, and no-show probability. The dependence of service capacity to the size of backlog captures the impact of the expediting policy. The state-dependence of the arrival process represents the balking process in the queue, where flexible patients who find all the slots presented on EBS occupied do not join the system. Finally, the state-dependence of no-show probability is to allow for a backlog-dependent no-show rate.

Given $(n, f, e^{(i)})$, denote the size of appointment backlog in the beginning of interval t by X_t. The recursive equation

$$X_{t+1} = \left(X_t - (n + e^{(X_t)} - C)^+\right)^+ + D^{(X_t)} + A^{(X_t)}, \quad t = 1, 2, \ldots, \tag{3}$$

captures the evolution of X_t over time, where $n + e^{(X_t)} - C$ is the realized service capacity in period t, $D^{(X_t)}$ is the number of no-shows in period t who request a new appointment at the end of t so

rejoin the queue, and $A^{(X_t)}$ is the number of "accepted" requests for advance appointments in period t (including all dedicated requests plus flexible requests who find a slot available on the EBS). It is important to note that due to the discrete-time nature of Equation (3), it reflects the true numbers in the queue as long as Assumption 1 holds.

To characterize $A^{(i)}$ in Equation (3), for a time interval with i patients in the backlog at the start of the interval, let $m^{(i)}$ be the total number of slots available to new advance patients on the EBS, i.e. $m^{(i)} = \left(f - (i - n - e^{(i)})^+\right)^+$. We then have

$$A^{(i)} = \begin{cases} A, & A \leq m^{(i)}, \\ m^{(i)} + \sum_{j=1}^{A-m^{(i)}} I_j, & \text{otherwise}, \end{cases} \tag{4}$$

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 requests in a time interval, A, is larger than the total number of slots available in the following intervals, $m^{(i)}$, only the first $m^{(i)}$ requests plus the remaining requests that are dedicated, given by the sum in the equation, are satisfied. The proposition below gives the probability mass function (p.m.f.) of $A^{(i)}$ based on the p.m.f of A. Throughout for a non-negative discrete random variable Y, we denote its associated probabilities by $y_j = \mathbb{P}(Y = j)$. All proofs and a list of notations are given in the online appendix.

Proposition 1. *The pmf for $A^{(i)}$ is given by*

$$\begin{aligned} a_k^{(i)} &= \mathbb{P}(A^{(i)} = k) \\ &= \begin{cases} a_k \mathbf{1}_{m^{(i)}}(k) + \theta^{k-m^{(i)}} \sum_{l=\max\{k,m^{(i)}+1\}}^{\infty} \binom{l-m^{(i)}}{k-m^{(i)}} (1-\theta)^{l-k} a_l, & \text{if } i - e^{(i)} < f + n, \\ \theta^k \sum_{l=k}^{\infty} \binom{l}{k} (1-\theta)^{l-k} a_l, & \text{otherwise}, \end{cases} \end{aligned} \tag{5}$$

where $\mathbf{1}_y(x)$ is an indicator function equal to 1 for $x \leq y$.

To characterize $D^{(i)}$ in Equation (3), we need to relate the no-show rate of a patient to her waiting time in the queue. Following Green and Savin (2008), we define the no-show probability as an increasing function of the number of customers left behind by a departing patient, as a proxy for the number of customers observed by an arriving patient. The simulation experiments conducted in Green and Savin (2008) demonstrate that this approximation yields very reliable results for queue length given a FCFS discipline in booking appointments. More importantly, their simulation experiments confirm that the results remain reliable when advance patients take any of the available

slots with equal probability. We therefore adopt the same approach as in Green and Savin (2008). However, since in our discrete-time model departures occur in batches, we cannot assign different no-show probabilities to individual patients departing in a batch. As such, we take a conservative approach and assume that the no-show probabilities of all customers in a departing batch is the same as the first one in the batch; see Izady (2015) for more details on this. Specifically, let the function $\gamma\left((X_t-1)^+\right)$ represent the no-show probability of all departing patients at the end of interval t, and assume that each no-show requires a new appointment with a fixed probability $0 \leq \zeta \leq 1$, independently of anything else. This yields a re-show probability of $\zeta\gamma\left((X_t-1)^+\right)$ for each departing patient at the end of period t. The conditional random variable $D^{(X_t)}|X_t=i, C=k$ will therefore have a binomial distribution with parameters $\left(\min\{i,(n+e^{(i)}-k)^+\},\alpha^{(i)}\right)$ for $i=0,1,\ldots$ and $k=0,\ldots,c$ where $\alpha^{(i)}=\zeta\gamma\left((i-1)^+\right)$.

Appointment cancellation by patients is another feature commonly observed in outpatient clinics. Since the slots freed up by these cancellations can be used by other patients, we capture their impact by excluding them from the total number of requests A. Patients may also re-schedule their appointments, moving from one position in the backlog to another. This does not influence the queue length calculations as long as Assumption 1 is met. The model given in (3) would therefore capture major characteristics of a wide range of outpatient clinics supported by EBS's. The reliability of the model is confirmed in our simulation experiments presented later where patient preferences for appointment days and some other details are explicitly considered.

In the following section, we provide a numerical approach for obtaining the steady-state queue length probabilities $x_i=\mathbb{P}(X=i)=\lim_{t\to\infty}\mathbb{P}(X_t=i)$, using which we can evaluate the objective function and constraints in (1) for given slot publication and expediting policies.

5. Numerical Analysis of Appointment Queue

Define $n^{(i)}=n+e^{(i)}$ as the nominal service capacity available to advance requests, and let $\boldsymbol{x}=(x_0,x_1,\ldots,x_m)$ be the stationary distribution for the discrete-time Markov chain characterized by Equation (3), where m is a sufficiently large upper-bound for the numbers in the system. One can find the the stationary queue length probabilities by solving balance equations $\boldsymbol{x\phi}=\boldsymbol{x}$, with $\boldsymbol{\phi}=[\phi_{ij}]$ the transition probability matrix specified below.

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

acterized by (3) *are*

$$\phi_{ij} = \mathbb{P}(C \leq n^{(i)} - i - 1)\left(\alpha^{(i)}\right)^{j}\left(\beta^{(i)}\right)^{i-j}\sum_{l=(j-i)^+}^{j}\binom{i}{j-l}\left(\frac{\beta^{(i)}}{\alpha^{(i)}}\right)^{l}a_l^{(i)}$$
$$+\left(\alpha^{(i)}\right)^{j-i}\left(\beta^{(i)}\right)^{i-j}\sum_{k=n^{(i)}-i}^{c}\sum_{l=(j-i)^+}^{j-i+(n^{(i)}-k)^+}\binom{(n^{(i)}-k)^+}{j-i+(n^{(i)}-k)^+-l}\left(\frac{\beta^{(i)}}{\alpha^{(i)}}\right)^{l}\left(\alpha^{(i)}\right)^{(n^{(i)}-k)^+}a_l^{(i)}c_k, \tag{6}$$

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

Once the steady state probabilities are found, one can obtain the desired performance metrics. Since our ultimate objective is to embed the queueing model into a heuristic search where performance is evaluated for a range of policy parameters, an efficient method must be used for calculating steady-state probabilities. The computation time for solving balance equations however grows with system size m, making it impractical for large systems. Below we develop an alternative approach based on probability generating functions (PGF's) that is more efficient for systems with large m.

For a discrete random variable Y, define its PGF as $Y(z) = \sum_{j=0}^{\infty} y_j z^j$, which is known to be analytic for $|z| < 1$ and continuous for $|z| \leq 1$. To find the PGF for X, we make the following assumption.

Assumption 2.

- The sequences $\{n^{(i)}\}_{i=0}^{\infty}$, $\{\alpha^{(i)}\}_{i=0}^{\infty}$, and $\{A^{(i)}\}_{i=0}^{\infty}$ are eventually constant, i.e. there exists positive integers h_n, h_α and h_A such that $n^{(i)} = n^{(*)}$ for $i \geq h_n$, $\alpha^{(i)} = \alpha^{(*)}$ for $i \geq h_\alpha$, and $A^{(i)} = A^{(*)}$ for $i \geq h_A$.
- $a_0^{(*)} = \mathbb{P}(A^{(*)} = 0)$ and $c_0 = \mathbb{P}(C = 0)$ are positive,
- $n^{(*)} \leq h_n$.

The first assumption above naturally happens in practice for $n^{(i)}$ and $\alpha^{(i)}$. For $A^{(i)}$, it is easy to verify that setting $h_A = h_n$ for $f + n \leq h_n - e^{(h_n)}$ and $h_A = f + n + e^{(h_n)}$ for $f + n > h_n - e^{(h_n)}$ satisfies this assumption, with the p.m.f for $A^{(*)}$ identified by the second equation in (5). In the second assumption, $c_0 > 0$ may not always hold but it can be fixed by assigning infinitesimally small probabilities to zero cancellations and adjusting the remaining probabilities. For $A^{(*)}$, it is verified from (5) that $a_0^{(*)} > 0$ as long as $\theta < 1$, or $\theta = 1$ and $a_0 > 0$. The third assumption is the direct consequence of constraint $e^{(i)} \leq (i - n)^+$.

Under Assumption 2, we must have $\mathbb{E}\left[A^{(*)}\right] / \left(1-\alpha^{(*)}\right) < n^{(*)} - \mathbb{E}[C]$ 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 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 3. *Given $\mathbb{E}\left[A^{(*)}\right] / \left(1-\alpha^{(*)}\right) < n^{(*)} - \mathbb{E}[C]$, the PGF of the stationary queue length X is given by*

$$X(z) = \Bigg[z^{n^{(*)}} \sum_{i=0}^{h_n-1} A^{(i)}(z)\alpha(i,z)^i x_i \mathbb{P}(C \le n^{(i)} - i - 1) + z^{n^{(*)}} \sum_{i=0}^{h-1} A^{(i)}(z) z^i x_i \sum_{k=n^{(i)}-i}^{c} \left(\frac{z}{\alpha(i,z)}\right)^{-(n^{(i)}-k)^+} c_k - A^{(*)}(z)G(z) \sum_{i=0}^{h-1} z^i x_i \Bigg] \Big/ \left(z^{n^{(*)}} - A^{(*)}(z)G(z)\right), \quad (7)$$

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

$$G(z) = \sum_{k=0}^{c} z^{\min\{k, n^{(*)}\}} \left(1 - \alpha^{(*)} + \alpha^{(*)} z\right)^{\left(n^{(*)}-k\right)^+} c_k.$$

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 the accepted requests distribution, re-show probabilities, or nominal 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 equations 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 $\mathbb{E}\left[A^{(*)}\right] / \left(1-\alpha^{(*)}\right) < n^{(*)} - \mathbb{E}[C]$ and finite $\mathbb{E}\left[A^{(*)}\right]$, 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 the third condition in Assumption 2, however, $n^{(*)} \leq h_n \leq h$, so the number of equations would not be enough for finding the unknown probabilities. To resolve this, we use the first $h-n^{(*)}$ stochastic balance equations as the additional relations required. Solving these h simultaneous equations yields $x_0, x_1, \ldots, x_{h-1}$, which fully specify the PGF $X(z)$. From this PGF, we can obtain most of the important performance metrics.

The number of equations that needs to be solved for the PGF approach is $h+1$; one equation for finding the complex roots of the denominator of $X(z)$ and h simultaneous equations as explained above. 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 using a software package such as Maple. This is particularly fast when $A^{(*)}(z)$ is a polynomial function, i.e. when the random variable $A^{(*)}$ has a finite support, which naturally happens when arrivals are represented by an empirical distribution. 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 provide much advantage over the balance equations approach. In contrast, in systems with constant no-show probability the value of h is typically substantially smaller than m, making the PGF approach computationally much more efficient.

Having found the queue length probabilities through the PGF or balance equations approach, the following corollary shows how each term in the optimization model (1) is evaluated.

Corollary 1. *For the optimization model given in* (1)*,*

$$\begin{aligned}\mathbb{E}\left[O(n,f,e^{(i)})\right] = &\sum_{i=0}^{h_n-1}\sum_{k=0}^{c}\sum_{j=0}^{s}\left(\min\left\{i,\left(n^{(i)}-k\right)^+\right\}+j-r+k\right)^+ x_i c_k s_j \\ &+\mathbb{P}(X\geq h_n)\left(\mathbb{E}\left[\left(n^{(*)}+S-r\right)^+\right]\mathbb{P}\left(C\leq n^{(*)}\right)+\mathbb{E}\left[(S-r+k)^+\right]\mathbb{P}\left(C>n^{(*)}\right)\right),\end{aligned} \tag{8}$$

$$\mathbb{E}\left[T(n,f,e^{(i)})\right] = \mathbb{E}[A]-\mathbb{E}\left[A^{(X)}\right] = \mathbb{E}[A]-(1-\theta)\sum_{i;\, i-e^{(i)}<f+n}\left(m^{(i)}+\sum_{k=0}^{m^{(i)}}a_k(k-m^{(i)})\right)x_i+\theta\mathbb{E}[A], \tag{9}$$

$$\mathbb{E}\left[W(n,f,e^{(i)})\right] = \mathbb{E}[X]/n, \tag{10}$$

where $\mathbb{E}[X]$ *is obtained by evaluating the first derivative of* $X(z)$ *at* $z = 1$ *for the PGF approach, and by* $\mathbb{E}[X] = \sum_{i=1}^{m} ix_i$ *for the balance equations approach.*

Equation (10) above uses the size of backlog expressed in time intervals as a conservative measure of shortest waiting time offered to the patients. It is conservative as when there are holes in the backlog shorter waiting times are also available on the system. Note that the actual waiting times patients endure may be different; they will be shorter when patients are expedited, for example, or longer when cancellations occur.

6. Heuristic Search

The optimization model in (1) is intractable in its general form. This is due to many different functional forms that can be considered for piecewise function $e^{(i)}$. Various realistic features included in our model also adds to the complexity. To make the problem tractable, we restrict $e^{(i)}$ to bi-level functions represented as

$$e^{(i)} = \begin{cases} 0, & i < \Delta, \\ \Upsilon, & i \geq \Delta. \end{cases} \tag{11}$$

This simplification enables us to derive insight into the structure of optimal policy using numerical experiments, leading to the development of a heuristic search. Implementing a bi-level expediting policy is also straightforward in practice. Using (11), the optimization model in (1) can be written as

$$\begin{aligned} \min_{(n,\Upsilon,\Delta,f)} \quad & \mathbb{E}\left[O(n,\Upsilon,\Delta,f)\right] \\ \text{s.t.} \quad & \mathbb{E}\left[W(n,\Upsilon,\Delta,f)\right] \leq q, \\ & \mathbb{E}\left[T(n,\Upsilon,\Delta,f)\right] \leq b, \\ & (n,\Upsilon,\Delta,f) \in \mathbf{\Psi}, \end{aligned} \tag{12}$$

where

$$\mathbf{\Psi} = \left\{(n,\Upsilon,\Delta,f) \in \mathbb{Z}^4 : n \leq r,\ \Upsilon_{min} \leq \Upsilon, \Delta_{min} \leq \Delta \leq \Delta_{max}\right\}, \tag{13}$$

with $\Upsilon_{min} = \mathbb{E}\left[A^{(*)}\right] / \left(1 - \alpha^{(*)}\right) - n + \mathbb{E}[C]$, $\Delta_{min} = n + \Upsilon$ and $\Delta_{max} = \Delta_{min} + \Upsilon\infty$. The lower bound on Υ is to ensure the stability of appointment backlog, and the lower bound on Δ is to satisfy $n^{(*)} \leq h_n$ in Assumption 2 (Note that for the bi-level expediting policy represented in (11), $h_n = \Delta$ and $n^{(*)} = n + \Upsilon$.) The upper bound on Δ is to restrict Δ to n when $\Upsilon = 0$. We can enumerate over a sensibly truncated $\mathbf{\Psi}$ to find the optimal solution to the model in (12), but the computation

time would still be substantial. For a system with $r = 20$, for example, we observed that about one million different policy combinations need to be evaluated before the optimal solution is found. Here we develop a heuristic search method that reduces the number of computations to about one thousand for the same example, and proves to be highly accurate. We start with some monotonicity properties in Section 6.1, using which we narrow down the search space $\mathbf{\Psi}$ in the following section. The heuristic algorithm is outlined in Section 6.3.

6.1. Monotonicity Properties

The properties in the conjecture below are inferred numerically through conducting a substantial range of experiments using real and illustrative data, a sample of which is provided in the online appendix.

Conjecture 1. *For the optimization model in* (12)*,*

(a) mean overtime slots increases in Υ *and* f*,*

(b) mean overtime slots decreases in Δ *when no-show probability is constant,*

(c) mean shortest offered waiting time increases in Δ *but decreases in* n *and* Υ*,*

(d) mean patients' turned away increases in Δ *but decreases in* n *and* Υ*,*

(e) mean patients' turned away decreases in f *when no-show probability is constant,*

(f) mean shorted offered waiting time increases in f *when* $\Upsilon = 0$*.*

For (a), the impact of Υ is intuitive but for f it is because with larger values of f the backlog becomes busier, reducing the average number of slots released to the EBS that remain open thus increasing the average number of required overtime slots for same-day requests. For (b), increasing Δ would delay the allocation of additional capacity to advance patients, thus reduce the need to overtime slots when no-show probability is constant. With a delay-dependent no-show probability, however, increasing Δ would make the backlog and consequently no-show probability larger and so more no-shows come back for further appointments. Since re-shows are always given appointments, this reduces the number of slots left open that can be allocated to same-day patients, which may offset the impact of the extra slots becoming available to them as a result of later activation of the expediting policy. Properties (c) and (d) are intuitive. For (e), note that increasing f increases the number of slots available to advance requests, so the number of patients turned away would reduce

if no-show probability is constant. However, with a delay-dependent no-show, increasing re-show rate as a result of larger f may reduce the slots available to new patients and so mean patients diverted may actually increase. Property (f) is intuitive. It does not hold for $\Upsilon > 0$ becasue in this case the expediting policy may be activated more frequently with larger f values, reducing the size of backlog and so the offered wait. By properties (c) and (d) above, and the fact that Υ is unbounded from above, it is clear that the model in (12) always has a feasible solution.

6.2. Narrowing Down The Search Space

Note that the set $\mathbf{\Psi}$ given in (13) is wider than the feasible region. In this section, we replace $\mathbf{\Psi}$ with a subset of the feasible region without excluding the optimal solution. To this end, for given $n, \Upsilon, \Delta \in \mathbb{Z}$ with $\Delta \geq n + \Upsilon$, define $f(n, \Upsilon, \Delta)$ as the smallest f satisfying both waiting and access constraints, i.e.

$$f(n, \Upsilon, \Delta) = \min\{f \in \mathbb{Z} : \mathbb{E}\left[W(n, \Upsilon, \Delta, f)\right] \leq q,\, \mathbb{E}\left[T(n, \Upsilon, \Delta, f)\right] \leq b\},$$

assuming that $f(n, \Upsilon, \Delta) = \infty$ when such f does not exist, i.e. the set in the right hand side of above is empty. Similarly, for given $n, \Upsilon \in \mathbb{Z}$, define $\Delta(n, \Upsilon)$ as the largest Δ for which there exists a finite $f(n, \Upsilon, \Delta)$, i.e.

$$\Delta(n, \Upsilon) = \max\{\Delta \in \mathbb{Z} : f(n, \Upsilon, \Delta) < \infty,\, \Delta_{min} \leq \Delta \leq \Delta_{max}\},$$

assuming $\Delta(n, \Upsilon) = -\infty$ when such Δ does not exist. Finally, for given $n \in \mathbb{Z}$, define $\Upsilon(n)$ as the smallest Υ for which there exists a finite $\Delta(n, \Upsilon)$, i.e.

$$\Upsilon(n) = \min\{\Upsilon \in \mathbb{Z} : -\infty < \Delta(n, \Upsilon),\, \Upsilon_{min} \leq \Upsilon\},$$

assuming $\Upsilon(n) = \infty$ when such Υ does not exist. Note that, $\Delta(n, \Upsilon)$ ($\Upsilon(n)$) is in fact the largest (smallest) feasible Δ (Υ) for given n and Υ (n). In light of definitions above, we have the following proposition.

Conjecture 2. *Let $(\tilde{n}, \tilde{\Upsilon}, \tilde{\Delta}, \tilde{f})$ be the optimal solution to optimization model in* (12)*. Then, $(\tilde{n}, \tilde{\Upsilon}, \tilde{\Delta}, \tilde{f}) \in \mathbf{\Pi}$, where $\mathbf{\Pi}$ is a subset of the feasible region defined as*

$$\mathbf{\Pi} = \left\{(n, \Upsilon, \Delta, f) \in \mathbb{Z}^4 : n \leq r,\, \Upsilon(n) \leq \Upsilon, \Delta_{min} \leq \Delta \leq \Delta(n, \Upsilon),\, f = f(n, \Upsilon, \Delta)\right\}.$$

Furthermore, for $n, \Upsilon, \Delta \in \mathbb{Z}$ with $\Delta \geq n + \Upsilon$,

$$f(n+1, \Upsilon, \Delta) \leq f(n, \Upsilon, \Delta), \quad f(n, \Upsilon+1, \Delta) \leq f(n, \Upsilon, \Delta), \quad f(n, \Upsilon, \Delta) \leq f(n, \Upsilon, \Delta+1), \tag{14}$$

$$\Delta(n, \Upsilon) \leq \Delta(n+1, \Upsilon), \quad \Delta(n, \Upsilon) \leq \Delta(n, \Upsilon+1), \quad \Upsilon(n+1) \leq \Upsilon(n). \tag{15}$$

A proof for the conjecture above based on the properties given in Conjecture 1 is provided in the appendix. Conjecture 2 provides the foundation for designing a heuristic as explained below.

6.3. Algorithm

Conjecture 2 implies that in the absence of expediting policy, i.e. with $\Upsilon = 0$ and $\Delta = n$, the optimal slot publication policy can be obtained by finding the smallest feasible f for each value of n, and subsequently finding the pair (n, f) with the smallest objective function. However, we made an observation in our numerical experiments that the optimal n is always the smallest n meeting both constraints. This is because the increase in the objective function due to a larger n is more than the potential reduction with a smaller f. As a result we conjecture that the optimal slot publication policy is the one with smallest n and f satisfying waiting and access constraints. For joint slot publication and expediting policies, one can enumerate over $\mathbf{\Pi}$, having set an upper bound on Υ. However, in our numerical experiments we observed that for each $n \in \mathbb{Z}$, $\Upsilon(n)$ is the optimal $\Upsilon > 0$. The explanation is that the increase in the objective function due to a larger Υ is bigger than the potential reduction made by larger Δ and smaller f.

These observations and the results obtained in the two previous sections lead to the development of a heuristic search, the details of which is elaborated in the appendix. We provide an overview here. The search is divided into two sections, one searching for the optimal slot publication policy with $\Upsilon = 0$ and $\Delta = n$, and the other searching for the optimal joint slot publication and expediting policies. For the former, the algorithm evaluates waiting and access constraints for increasing values of n and f until both are met for the first time. For the latter, the algorithm iterates over all combinations of n, Υ, Δ, and f (with upper/ lower bounds set based on inequalities in Conjecture 2), as follows. In the most inner loop, the algorithm searches for the smallest f satisfying both access and waiting constraints for given n, Υ and Δ. Once such f is found, $f(n, \Upsilon, \Delta)$, the current optimal values, and $\Delta(n, \Upsilon)$ are updated and the same process is repeated for $\Delta + 1$. The search for Δ stops when for a given Δ no f can be found satisfying both constraints, i.e. $f(n, \Upsilon, \Delta) = \infty$. This is because by the last inequality in (14), $f(n, \Upsilon, \Delta + x) = \infty$ if $f(n, \Upsilon, \Delta) = \infty$ for $x \in \mathbb{Z}$. Once the search for Δ ends, the value of $\Upsilon(n)$ is updated to the current Υ and the search for Υ stops if a finite $\Delta(n, \Upsilon)$ is found. Otherwise the same process is repeated for $\Upsilon + 1$. Finally, the most outer loop considers all n values from 0 up to r.

7. Numerical Results

In this section, we apply the models developed in the paper to two different outpatient environments. One is based on representative data for an advanced access clinic, and the other is based on real data from a specialty clinic. The objective here is first to validate the structural properties conjectured in Section 6, second to derive management insight into the behaviour of optimal policies in different circumstances, and finally to validate the reliability of the heuristic search. We note that the accuracy of policies obtained from our heuristic are confirmed in all scenarios below through complete enumeration, thus we refer to these policies as optimal.

7.1. A Typical Advanced Access Clinic

Using the data provided in the literature, we set the exogenous parameters in our model so as to represent a typical advanced access clinic. In particular, following Liu (2015), we set $r = 20$ regular slots per day, and assume the total average demand (same-day plus advance booking) is 18, 19, or 20 patients per day, representing low, moderate or high workload, respectively. Following Green and Savin (2008), we assume all no-shows reschedule an appointment, i.e. $\zeta = 1$, and there is no slot cancellation by the clinic. To investigate the impact of no-show probabilities, we consider the parametric functions $\gamma(i) = 0.31 - (0.31 - 0.01)e^{-i/50n}$, $\gamma(i) = 0.51 - (0.51 - 0.15)e^{-i/9n}$, and $\gamma(i) = 1 - 0.5e^{-0.017i/n}$, proposed in Green and Savin (2008), Gallucci et al. (2005) and Kopach et al. (2007), to represent low, moderate and high delay-dependent no-show probabilities, respectively. Murray and Tantau (2000) report that about 25% of patients prefer to book appointments in advance, but we consider 35% and 45% as well. To investigate the impact of demand variability, we assume both advance and same-day booking requests follow either Poisson or Negative Binomial distributions, where the parameters of Negative Binomial are set so that its variance to mean ratio equals 2, as opposed to 1 for Poisson. We set $q = 2, 3, 4$ days, and $b = 2.5\%, 5\%, 7.5\%$ of mean advance booking requests. We assume $\theta = 0.5$ in all scenarios to represent a balanced case.

Using a subset of data explained above, we present several plots in the online appendix supporting the conjectures made in Section 6 for clinics with delay-dependent no-shows. Below we investigate a series of phenomena through numerical experiments. We use the notation $\mathcal{D}_{\mathcal{N},\mathcal{T},\mathcal{R}}$ to represent each scenario, where $\mathcal{D} \in \{P, N\}$ denotes Poisson and Negative Binomial demand distributions, $\mathcal{N} \in \{GS, G, K\}$ denotes the no-show functions given in Green and Savin (2008), Gallucci et al. (2005) and Kopach et al. (2007), $\mathcal{T} \in \{18, 19, 20\}$ denotes the total average demand,

and $\mathcal{R} \in \{0.55, 0.65, 0.75\}$ denotes the same-day demand ratio. For each phenomenon, we only illustrate a selection of 486 scenarios investigated for brevity, and present p instead of f in the slot publication policy to simplify comparison with the relevant literature.

Joint vs. Sequential Optimization Given that n and f are considered separately in the literature, it is natural to assume one could obtain the optimal (n, f) through sequential optimization. In this experiment, we evaluate the efficiency gains from joint optimization of (n, f) as compared to sequential optimization. For sequential optimization, assuming $f = \infty$ we first find the value of n yielding the lowest average overtime cost whilst satisfying the waiting constraint. Then for the resulting n we find the smallest f satisfying both access and waiting constraints. The results given in Table 1 for two scenarios illustrate savings as large as 19.5% in overtime cost as a result of joint optimization of (n, f). This is because sequential optimization often over-estimates the value of n, hence results in higher overtime costs.

Table 1: Joint vs. sequential optimization of slot publication policy.

Scenario	b	q	Joint Opt.				Seq. Opt.				Saving
			(n,p)	$\mathbb{E}[O]$	$\mathbb{E}[W]$	$\mathbb{E}[T]$	(n,p)	$\mathbb{E}[O]$	$\mathbb{E}[W]$	$\mathbb{E}[T]$	
$P_{GS,19,0.75}$	2.5%	2, 4, 6	(5,3.2)	1.087	1.932	2.3%	(6,1.3)	1.154	0.890	2.3%	5.8%
	5.0%	2, 4, 6	(5,2.0)	1.052	1.445	4.5%	(6,1.2)	1.132	0.854	3.6%	7.0%
	7.5%	2, 4, 6	(5,1.4)	1.008	1.173	7.4%	(6,1.0)	1.099	0.807	5.6%	8.3%
$P_{G,20,0.55}$	2.5%	2, 4, 6	(12,1.3)	2.612	1.114	2.3%	(13,1.0)	2.686	0.893	2.0%	2.7%
	5.0%	2, 4, 6	(12,1.0)	2.451	0.992	4.3%	(13,1.0)	2.686	0.893	2.0%	8.7%
	7.5%	2, 4, 6	(11,1.1)	2.161	1.163	7.4%	(13,1.0)	2.686	0.893	2.0%	19.5%

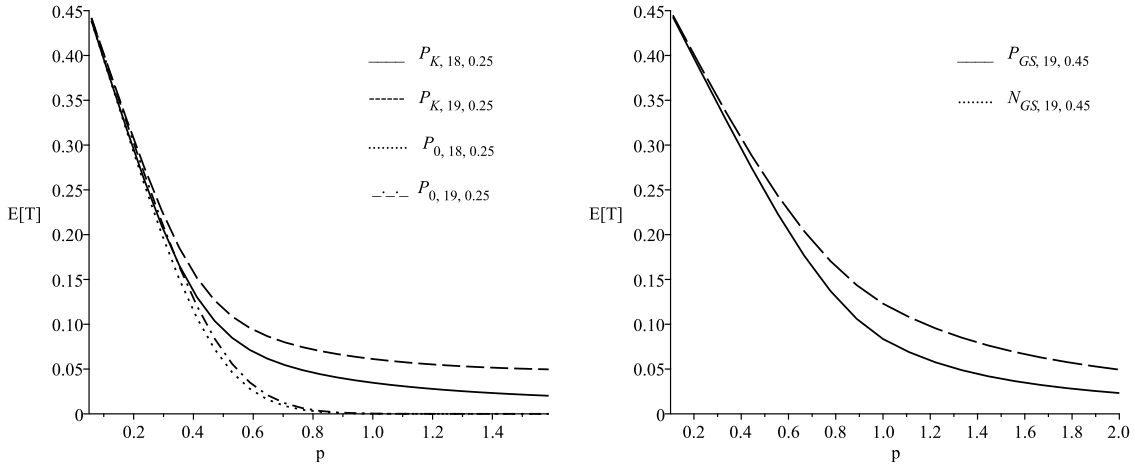
Sensitivity to q and b The results in Table 1 show that with optimal publication policy the mean number of patients diverted reaches a value close to the threshold b while the mean shortest offered wait falls substantially below the target q. This is because, on the one hand, the entire range of feasible values for waiting time and access are not covered due to the discrete nature of decision variables. On the other hand, both mean offered wait and mean overtime cost increase with f, pushing the optimal policy away from the waiting target. As a consequence of this, the slot publication policy is less sensitive to the parameter q than b as illustrated in Table 1.

Impact of Demand Mean Liu (2015) prove that the optimal scheduling window decreases with demand in their profit maximization model, where the clinic earns (incurs) a revenue (cost) for every patient visited (turned away). Here we investigate if this result remains valid in our formulation. To this end, we fix $q = 4$ days and $b = 5\%$, and present the optimal publication policies for different values of total demand and with different no-show functions for the scenarios

with $\mathcal{D} = P$ and $\mathcal{R} = 0.55$ in Table 2. The results in this table indicate that n is non-decreasing in total demand as expected. The impact of demand on p, however, is more complicated; it decreases with increasing demand as long as the optimal n increases, but increases otherwise. For example, when the total demand increases from 18 to 19 for the scenario with $\mathcal{N} = K$ in Table 2, the optimal n remains constant and the optimal p doubles. This contradicts the result in Liu (2015) stated above. To understand why, observe in Figure 1, panel(a) that $\mathbb{E}[T]$ reduces at a lower pace with respect to p when demand increases. The difference in rate of reductions is substantial when the impact of no-shows rescheduling appointments is captured as in our models, hence a larger p is needed to achieve the access constraint. When the impact of returning no-shows is excluded as in the models in Liu (2015), however, the difference in reduction rates is very small, hence the benefits of having a smaller p outweighs the cost of turning patients away.

Table 2: Optimal (n, p) for different total demand values and no-show functions when demand is Poisson.

Scenario / No-show Function	GS	G	K
$P_{\mathcal{N},18,0.55}$	(8,2.5)	(11,1.0)	(17,0.8)
$P_{\mathcal{N},19,0.55}$	(9,1.3)	(11,1.2)	(17,1.6)
$P_{\mathcal{N},20,0.55}$	(9,1.9)	(12,1.0)	(18,1.2)



(a) (b)

Figure 1: (a) $\mathbb{E}[T(17, 0, 17, p)]$ and (b) $\mathbb{E}[T(9, 0, 9, p)]$ as a function of p ($\mathcal{N} = 0$ represents the scenario with no-shows not rescheduling new appointments)

Impact of Demand Variability To investigate the impact of demand variability, we present optimal policies in Table 3 for the same scenarios as in Table 2 but with demand distributed as Negative Binomial. Table 3 also includes the corresponding percentage increase in optimal overtime

cost. The results show that the optimal n is non-decreasing in demand variability. The optimal p, however, decreases with demand variability if n increases, and increases otherwise. The former trend is intuitive but to explain the latter, we plot $E[T]$ in Figure 1, panel (b) as a function of p for both Poisson and Negative Binomial demand. This figure shows that when the demand variability increases, $E[T]$ reduces at a lower pace with respect to p, thus a larger p is needed for meeting the access constraint. The results in Table 3 also show that the optimal overtime cost increases when demand variability increases, but the extent of this increase is much smaller with highly delay-dependent no-show probabilities. This is because with a highly delay-dependent no-show probability, a large fraction of advance booking patients turn to no-shows, thus demand variability will have a smaller impact on performance.

Table 3: Optimal results for different total demand values and no-show functions when demand is Negative Binomial.

Scenario / No-show Function	GS	G	K
$NB_{\mathcal{N},18,0.55}$	(9,1.7), 117%	(11,1.4), 42%	(17,1.2), 4%
$NB_{\mathcal{N},19,0.55}$	(9,2.0), 119%	(11,1.9), 70%	(18,1.1), 1%
$NB_{\mathcal{N},20,0.55}$	(9,3.3), 45%	(12,1.4), 20%	(19,1.0), 1%

Impact of Expediting Policy For each scenario, we evaluate the average saving in mean overtime cost as a result of using joint publication and expediting policies compared with publication policy alone across all nine combinations of q and b. Our experiments show that when $\mathcal{R} = 0.75$ as suggested in Murray and Tantau (2000), the average saving falls below 2.5% for all scenarios. Given the additional challenges of expediting policy, its implementation may not then be justified. When $\mathcal{R}$ reduces to 0.55, however, the average saving increases reaching a maximum of 7.0% (for scenario $P_{GS,19,0.55}$), making it slightly more attractive for implementation.

7.2. A Specialty Clinic in the UK

We focus on a non-emergency glaucoma service delivered in this clinic on a referral basis (The emergency service is delivered in the accident and emergency department.) The service is provided over three days of the week. Routine patients book slots through the e-RS with waiting times that may exceed several weeks, and urgent patients must be seen within one week of referral. As such, we work with a weekly time unit. We obtained data from this clinic over a one year period (52 weeks) starting from July 2014, and extracted empirical distributions for exogenous variables in our model, i.e. numbers of new routine and urgent patients referred to the clinic (excluding those who cancel their appointments and return back to the primary care) as well as the slots cancelled

by the clinic. As summarised in Table 4, all three distributions are over-dispersed, highlighting the necessity of using general distributions. Table 4 also indicates that urgent referrals account for around 20% of total referrals to this clinic, a common ratio in clinics with carve-out delivery. We did not distinguish a strong delay-dependent no-show behaviour so a constant no-show probability $\gamma = 0.0627$ and re-scheduling probability $\zeta = 1$ obtained from the data is used in our analysis. A total of $r = 18$ regular slots are available for this service per week. We use $\theta \in \{0.1, 0.9\}$ in our experiments to consider both extremes. Several plots supporting the conjectures made in Section 6 are illustrated in the online appendix based on the data from the clinic.

Table 4: Summary statistics for weekly referrals and slot cancellations.

Measure	Routine Referrals	Urgent Referrals	Cancellations
Mean	14.280	3.326	1.098
Variance	48.683	4.724	3.050

Joint Slot Publication and Expediting Policy Table 5 illustrates the optimal slot publication policy, as well as the optimal joint publication and expediting policies, along with their corresponding performance for all combinations of $\theta = 0.1, 0.9$, $q = 2, 4, 6$ weeks and $b = 2.5\%, 5\%, 7.5\%$ of average weekly advance requests. The 'Saving^1" column in the table gives reductions in overtime cost as a result of joint optimization of (n, f) as compared to sequential optimization. The scale of reductions here are even larger than those observed in Table 1. Table 5 also shows that deployment of the expediting policy jointly with the publication policy enables the clinic to release one fewer slot per week, but a larger total number of slots, to the EBS as compared to the publication policy alone. Smaller n and larger f values result in a longer polling range (recall that $p = f/n$), giving patients more choice over their appointment weeks without compromising on the waiting time target. As illustrated in the "Saving^2 " column, the potential savings in overtime cost made by using the joint expediting and publication policies compared with publication policy alone could be as large as 18%. Furthermore, the deployment of expediting policy yields a wider spread of feasible points, making both mean offered wait and mean patients diverted close to their threshold values.

The interesting feature of the expediting policies represented in Table 5 is that they require only one expedited visit in all scenarios but one. The frequency of activating them, as indicated by the "$\mathbb{P}(X \geq \Delta)$" column, is also generally small when the corresponding savings are large. For example, in the scenarios in Table 5, they needs to be activated at most once every four weeks when the savings are 15% or more. This reduces the administrative cost of implementing the expediting

policy, making it even more attractive.

Table 5: Joint publication and expediting policies for the specialty clinic.

θ	b	q	Pub. Policy				Saving[1]	Joint Pub. and Exp. Policy						Saving[2]
			(n, f)	$\mathbb{E}[O]$	$\mathbb{E}[W]$	$\mathbb{E}[T]$		(n, f)	(Υ, Δ)	$\mathbb{E}[O]$	$\mathbb{E}[W]$	$\mathbb{E}[T]$	$\mathbb{P}(X \geq \Delta)$	
0.1	2.5%	2	(17,48)	1.899	1.615	2.4%	18.8%	(16,53)	(1,34)	1.748	1.990	2.5%	0.45	8.0%
		4	(17,48)	1.899	1.615	2.4%	0.0 %	(16,95)	(1,94)	1.583	3.901	2.5%	0.10	16.6%
		6	(17,48)	1.899	1.615	2.4%	0.0 %	(16,100)	(1,100)	1.577	4.128	2.5%	0.09	17.0%
	5.0%	2	(17,32)	1.752	1.260	4.8%	18.1%	(16,46)	(1,51)	1.431	1.988	5.0%	0.02	18.3%
		4	(16,48)	1.422	2.066	4.9%	23.2%	(15,76)	(1,58)	1.319	3.969	5.0%	0.60	7.3%
		6	(16,48)	1.422	2.066	4.9%	23.2%	(15,100)	(1,83)	1.311	5.525	5.0%	0.58	7.8%
	7.5%	2	(16,31)	1.314	1.438	7.3%	43.1%	(15,37)	(1,32)	1.178	1.914	7.5%	0.44	10.3%
		4	(16,31)	1.314	1.438	7.3%	22.4 %	(15,66)	(1,66)	1.113	3.655	7.4%	0.25	15.3%
		6	(16,31)	1.314	1.438	7.3%	22.4%	(15,85)	(1,86)	1.094	4.880	7.5%	0.20	16.7%
0.9	2.5%	2	(17,31)	1.898	1.699	2.5%	15.7%	(16,34)	(1,29)	1.779	1.987	2.5%	0.52	6.3%
		4	(17,31)	1.898	1.699	2.5%	0.0 %	(16,77)	(1,98)	1.584	3.991	2.5%	0.10	16.6%
		6	(17,31)	1.899	1.699	2.5%	0.0%	(16,100)	(1,134)	1.559	5.116	2.5%	0.05	17.9%
	5.0%	2	(17,14)	1.740	1.416	5.0%	15.0%	(16,22)	(1,50)	1.482	1.968	5.0%	0.13	14.9%
		4	(16,28)	1.423	2.455	4.9%	18.2%	(15,49)	(1,51)	1.322	3.964	5.0%	0.62	7.1%
		6	(16,28)	1.423	2.455	4.9%	18.2%	(15,80)	(1,83)	1.315	5.950	4.9%	0.59	7.7%
	7.5%	2	(16,10)	1.307	1.992	7.4%	31.7%	(15,12)	(2,36)	1.286	1.989	7.5%	0.29	1.6 %
		4	(16,10)	1.307	1.992	7.4%	18.7%	(15,32)	(1,73)	1.111	3.874	7.5%	0.25	15.0%
		6	(16,10)	1.307	1.992	7.4%	18.7%	(15,61)	(1,113)	1.092	5.930	7.5%	0.20	16.4%

Expediting Policy without Revising Publication Policy Suppose a clinic that is already meeting the access and waiting targets is planning to use the expediting policy in order to reduce the overtime cost. For example, consider the glaucoma clinic with $\theta = 0.9$, $b = 5.0\%$, and $q = 4$, and assume that the clinic is operating with the optimal publication policy, i.e. $(n = 16, f = 28)$ according to Table 5. Property (a) in Conjecture 1 suggests that with a fixed n and f, increasing Υ would only increase cost so implementing the expediting policy without revising publication policy would only make the situation worse. Suppose now that the clinic is flexible in revising the polling range. Our analysis show that the best policy in this case would be $(n = 16, \Upsilon = 1, \Delta = 137, f = 27)$, for which the saving would be less than 1.0%. However, if the clinic is flexible in revising both dimensions of publication policy, the results in Table 5 show that it can reduce the overtime cost by as much as 7.1%. This experiment further highlights the importance of adopting an integrated approach to publication and expediting policies.

Table 6: Simulated performance with optimal policies given in Table 5

b	q	Pub. Policy				Joint Pub. and Exp. Policy				Saving
		$\mathbb{E}[O]$	$\mathbb{E}[W]$	$\mathbb{E}[T]$	$\mathbb{E}[E]$	$\mathbb{E}[O]$	$\mathbb{E}[W]$	$\mathbb{E}[T]$	$\mathbb{E}[E]$	
	2	1.84,1.88	1.69,1.87	3.4%,2.9%	8.6%,8.0%	1.77,1.80	1.86,2.00	3.4%,2.9%	6.2%,6.1%	3.9%,4.2%
2.5%	4	1.84,1.88	1.69,1.87	3.4%,2.9%	8.6%,8.0%	1.57,1.58	3.75,3.98	3.0%,2.7%	1.8%,1.7%	14.9%,16.0%
	6	1.84,1.88	1.69,1.87	3.4%,2.9%	8.6%,8.0%	1.57,1.55	3.78,5.03	2.8%,2.9%	1.9%,1.3%	15.1%,17.7%
	2	1.68,1.74	1.36,1.55	6.0%,4.9%	11.3%,0.0%	1.39,1.46	1.97,2.00	5.9%,5.5%	4.8%, 5.0%	17.2%,15.8%
5.0%	4	1.38,1.40	2.12,2.54	5.9%,5.2%	4.4%,4.0%	1.34,1.33	3.49,3.89	5.0%,5.2%	0.9%, 0.8%	2.7%,5.3%
	6	1.38,1.40	2.12,2.54	5.9%,5.2%	4.4%,4.0%	1.31,1.32	4.94,5.76	5.1%,5.0%	0.3%,0.2%	5.2%,5.8%
	2	1.27,1.31	1.51,2.00	8.3%,7.4%	7.3%,0.0%	1.16,1.28	1.77,2.02	8.5%,7.4%	4.0%,0.0%	8.9%,2.6%
7.5%	4	1.27,1.31	1.51,2.00	8.3%,7.4%	7.3%,0.0%	1.11,1.11	3.32,3.98	7.8%,7.7%	0.8%,0.9%	12.9%,15.2%
	6	1.27,1.31	1.51,2.00	8.3%,7.4%	7.3%,0.0%	1.08,1.09	4.51,5.90	7.7%,7.4%	0.4%,0.2%	15.1%,16.7%

The first (second) number in each cell represents the result for $\theta = 0.1$ ($\theta = 0.9$).

7.3. Validation

To validate the reliability of policies obtained from the heuristic search, we develop a simulation model where additional features are included as follows. The order of booking activities in the simulation model is the same as described in Section 3. Given $\mathcal{F} \subset \{1, 2, \ldots, p\}$ the set of future intervals with empty slots on the EBS in a particular interval, the model assumes that an advance request books an appointment in interval j into the future with probability $w_j / \sum_{k \in \mathcal{F}} w_k$ where $w_j = 0.1(p - j + 1)$ for $j \in \mathcal{F}$. This is motivated by the multi-nomial logit probability model considered in Feldman et al. (2014), and represents the common situation where patients prefer later appointments less. If $\mathcal{F}$ is empty and the patient is dedicated, the first available slot beyond f is allocated to the patient. The model further assumes that when the expediting policy is activated, patients in the future intervals accept to be expedited with equal probabilities. Re-shows and cancellations are assumed to be allocated the earliest available slot as typically happens in reality. For delay-dependent no-shows, the simulation model uses the actual waiting time to estimate the no-show probability. Apart from the usual metrics, the simulation estimates $\mathbb{E}[E]$, where E represent the (random) ratio of empty slots out of n in an interval if future intervals are booked and zero otherwise. It indicates the extend to which Assumption 1 is violated. We conducted 20 replications of the simulation model, each over 50,000 time intervals, with optimal policies obtained from the heuristic algorithm. The 95% confidence intervals constructed from simulation results fell within $\pm 1\%$ of corresponding averages in all cases, suggesting a low estimation error. The simulated performance metrics for the glaucoma clinic are presented in Table 6.

Comparing the results in Table 6 with corresponding values in Table 5 reveals that waiting

constraint is met in all scenario and access constraint is violated by a maximum of one percentage point when policies derived from the heuristic are applied in a model with patient preferences included. It also shows that the efficiency gains from using a joint policy are replicated to a large extent, and the error in approximating mean overtime slots does not exceed 5%. The errors in overtime cost and patients diverted are due to Assumption 1 not being met in some intervals because of patient choice, as indicated by the $\mathbb{E}[E]$ values. This results in more slots being available to same-day requests and fewer advance patients served in those intervals. In general, $\mathbb{E}[E]$ and thus the queueing approximation error reduce when θ increases and/or expediting policy is deployed. For θ, it is because when it increases more advance patients book appointments, reducing the number of slots remaining empty. For expediting policy, it is because when it is used the optimal publication policy involves smaller n and larger f values, reducing the likelihood of slots remaining empty. Similar insights as above are obtained using the advanced access clinic data, which is significant given the approximation used for estimating no-show probability in the queueing model. These results indicate the reliability of the policies obtained from the heuristic search.

8. Conclusions

EBS's 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 more flexibility in managing their supply and demand. At a strategic level, this flexibility is achieved by enabling the clinics to set the number of slots allocated to different patient groups and determine the timing of their release to the booking system. Booking managers often rely on personal experience or general guidelines provided by the EBS's supplier to decide the appropriate values for these parameters. We proposed a structured approach capturing various sources of uncertainty in outpatient environments to facilitate this decision making. It includes an efficient performance evaluation model and a heuristic search algorithm. The performance evaluation model is based on a discrete-time queueing infra structure that neither requires a FCFS discipline nor detailed information on patient choice. We developed two different approaches, one using the common stochastic balance equations and the other probability generating functions, for finding the steady-state performance of the queueing model. The latter approach increases the computational speed substantially when no-show probability is constant. For example, we observed that computation time reduces by a factor of 4 to 6 when the PGF approach is applied with specialty clinic data as compared with balance equations. The heuristic search is also both efficient and accurate. Our numerical investigation

across a wide range of parameter instances shows that it finds the optimal solution by evaluating a maximum of 0.1% of policy combinations.

The reliability of the queueing approximation is investigated through comparison with a simulation model that incorporates further realistic features, in particular patient preferences for appointment days. A multi-nomial logit probability is assumed for patient choice in this model, giving larger probabilities to earlier appointments. This is a realistic assumption given the findings in the literature that patients in general prefer appointments that are sooner rather than later, unless the appointment on a later day is at a more convenient time (Feldman et al. 2014). The simulation results suggest reliable accuracy of the queueing model, especially given the highly variable demand and short polling ranges (imposed by tight targets set on access and waiting times) in the case studies.

We highlighted the value of adopting an integrated approach to demand and capacity planning, and proposed a unifying framework using a combination of slot publication and expediting policies for this purpose. We illustrated the efficiency gains from joint optimization of scheduling window and number of daily pre-bookable slots as compared to their sequential optimization, as well as the savings from joint consideration of slot publication and expediting policies as compared to each considered on their own. Given the wide range of realistic features included in our models, they can be used not only for generating top-level managerial insight but also as a decision support system for clinics using EBS's.

References

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.

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

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

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 âĂŤ a computational experience. *Journal of Systems Science and Systems Engineering* **20**(4) 475–494.

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.

Kopach, R., P. DeLaurentis, M. Lawley, K. Muthuraman, L. Ozsen, R. Rardin, H. Wan, P. Intrevado, X. Qu, D. Willis. 2007. Effects of clinical characteristics on successful open access scheduling. *Health Care Management Science* **10**(2) 111–124.

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* **25**(1) 128–142.

Liu, N., S. Ziya. 2014. Panel size and overbooking decisions for appointment-based services under patient no-shows. *Production and Operations Management* **23**(12) 2209–2223.

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.

Murray, M. F. 2007. Improving access to specialty care. *Joint Commission Journal on Quality and Patient Safty* **33**(3) 125–35.

NHS Digital. 2016. NHS e-Referral service managing urgent referrals for suspected cancer(2 Week Waits): Best Practice Guide. Retrieved November 27, 2017. URL `https://digital.nhs.uk/media/32683/Urgent-Referrals-for-Suspected-Cancer-2016/pdf/Urgent_Referrals_`

NHS Digital. 2017. NHS e-Referral service: managing and minimising appointment slot issues. Retrieved November 27, 2017. URL http://content.digital.nhs.uk/media/17208/Managing-and-Minimising-Appointment-Slot-Issues/pdf/Manag

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.

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.

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

Truong, Van-Anh, Carrie Ruzal-Shapiro. 2014. Optimal advance scheduling, with expediting. *submitted* .

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