Submitted to
manuscript (Please, provide the manuscript number!)

Appointment Capacity Planning in Specialty Clinics: A Queueing Approach

Navid Izady
School of Mathematical Sciences / Southampton Business School, University of Southampton, n.izady@soton.ac.uk

Specialty clinics provide specialized care for patients referred by primary care physicians, emergency departments, or other specialists. Urgent patients must often be seen on the referral day, while non-urgent referrals are typically booked an appointment for the future. To deliver a balanced performance, the clinics must know how much 'appointment capacity' is needed for achieving a reasonably quick access for non-urgent patients. To help identify the capacity that leads to the desired performance, we model the dynamics of appointment backlog as novel discrete-time bulk service queues, and develop numerical methods for efficient computation of corresponding performance metrics. Realistic features such as arbitrary referral and clinic appointment cancellation distributions, delay-dependent no-show behaviour and rescheduling of no-shows are explicitly captured in our models. The accuracy of the models in predicting performance as well as their usefulness in appointment capacity planning is demonstrated using real data. We also show the application of our models in capacity planning in clinics where patient panel size, rather than appointment capacity, is the major decision variable.

Key words: specialty clinics; appointment capacity; discrete bulk service queues; no-shows and cancellations

1. Introduction

Specialty clinics provide specialized and often complex care, including diagnosis and treatment, for patients who have been referred by primary care physicians, emergency departments (EDs), or other specialists. Urgent patients must often be seen on the referral day, while non-urgent referrals are typically booked an appointment for a later day. The waiting times for appointments however are often substantially long. For example, the average waiting time for the first outpatient specialty visit was 39 days in the UK National Health System (NHS) in 2011-12 (HES Report 2012), or the 2009 Survey on Physician Appointment Wait Times in the US suggests an average wait of longer than 20 days for non-urgent patients seeking specialty care. Apart from patient dissatisfaction, excessive delays might lead to adverse clinical outcomes, demoralization of specialty staff, and inefficiency and rework across the health system (Murray 2002, 2007),

In contrast with primary care offices where a significant proportion of providers' time is devoted to pre-booked outpatient visits, in specialty care clinics numerous other activities and duties 'compete' for the limited specialist time. In this competition, most often more 'urgent' streams of demand, including emergency surgery, urgent outpatient visits, and various on-call functions, take precedence over non-urgent outpatient visits. This causes the delays and inefficiencies in the system to be relegated to the lowest priority stream, the non-urgent outpatient stream, resulting in large backlogs of appointments. For instance, clinics often postpone pre-booked outpatient visits in favour of patients with emergency conditions when faced with shortage of staff. To deliver a balanced performance to non-urgent and urgent streams of demand, clinics must know how much 'appointment capacity' is needed for achieving a reasonably quick access for non-urgent patients. In this paper, we develop tractable queueing models that efficiently compute patients' waiting times and the size of appointment queues in terms of appointment capacity and other important system characteristics, and thus provide guidance in identifying the appointment capacity for achieving a given performance target.

We define the 'nominal' appointment capacity as the number of appointment slots made available to non-urgent patients per unit of time, e.g. a clinic session, a day or a week, and assume it is constant during a planning horizon. Our objective is to find the nominal appointment capacity such that the average or a percentile of patients' waiting times does not exceed a threshold limit. The need to specify the nominal appointment capacity is more highlighted in electronic referral and

Table 1 Summary of the Parameters For Clinics A, B, and C.

		Referral Distribution	Realized Capacity Distribution	No-show Probability	Re-scheduling Probability
Clinic A	Mean	60.40	69.40	0.156	0.70
	Var	363.97	417.55		
Clinic B	Mean	11.15	13.00	0.067	0.31
	Var	21.11	30.49		
Clinic C	Mean	99.45	99.94	0.071	0.99
	Var	479.61	624.24		

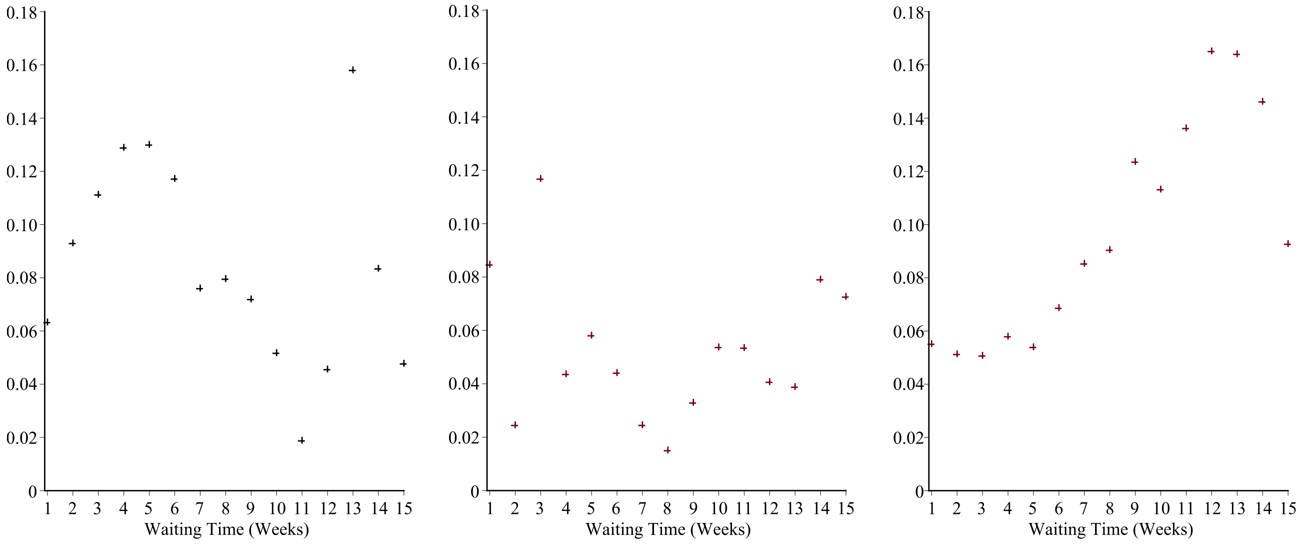
Time unit is one week

appointment booking systems, like the Choose and Book system implemented in the UK, where providers must decide in advance how many appointment slots to release for each particular day during a planning horizon. Once the nominal appointment capacity is set and patients are booked, however, the clinics may not always be able to realize the full capacity. It is often the case that, due to consultants' delays and absenteeism and/or an unpredicted increase in the emergency workload, providers have to cancel some pre-booked appointments. This results in some variability in the actual number of appointments supplied which, as pointed out by Murray (2007), could be even larger than the variability in the demand for appointments. On the other hand, while many patients are waiting, some do not turn up for their appointments, wasting the valuable clinic capacity. The likelihood of patients not showing up for their appointments sometimes increases with the amount of time they have waited in the appointment queue. Almost all the patients whose appointments are cancelled by the clinic as well as a fraction of no-shows must be given a new appointment, hence rejoin the appointment queue.

Using a weekly time unit, Table 1 illustrates some of the features explained above using the data obtained from three specialty clinics in the UK, referred to as clinics A, B, and C. The summary measures have been calculated over a one year period starting from March 2012 (excluding public holidays), during which there was no visible time-dependence effect. This table shows varying degrees of patient no-show and rescheduling probabilities, and a highly variable supply of appointments. Most importantly, it indicates a high degree of variability in the referral distribution, challenging the assumption of Poisson arrivals widely made in the literature.

Our contribution in this paper is developing three new queueing models that take the complexities discussed above into account. The first model considers no-shows and works with an arbitrary referral distribution but assumes that the realized appointment capacity is the same as the nominal capacity, i.e. no clinic cancellations. For this model, we derive the steady-state distribution for the size of the appointment queue (appointment backlog or waiting list) as well as patients' waiting time, taking rescheduling of no-shows into account. Our method for deriving the waiting time distribution is new and can be adapted to various rules with which no-shows' waiting times are calculated. To account for the variability observed in the supply of appointments, an element which has consistently been overlooked in the literature, we develop a second queueing model where the realized capacity is also an uncertain variable. Since in our clinic data, as illustrated in Figure 1, the no-show probability does not show a strong increasing trend with respect to waiting time, we assume a fixed no-show probability in the first two models. However, to extend the applicability of our models to situations where such trend does exist, as in the clinics studied by Green and Savin (2008) and Liu et al. (2010), we develop a third model where no-show probability is an increasing function of the size of appointment backlog.

Figure 1 Observed No-show Probability versus Waiting Time in Clinics A (left), B (middle), and C (right).



Our illustrative experiments with the first model provide insight on the impact of no-shows on performance under various referral distributions. In particular, they suggest that in clinics with referral distributions as variable as those illustrated in Table 1, a Poisson assumption for referrals could seriously under-estimate the capacity needed for achieving a given target. They further suggest that 'inflating' demand by the appropriate percentage based on the no-show and rescheduling probabilities would not fully capture the impact of no-shows returning to the system. This is because re-shows not only increase the mean but also influence the higher moments of the 'effective' arrival (i.e. new referrals plus re-shows) distribution, and hence adjusting the mean alone would not be enough. The impact of re-shows on the higher moments is not however always negative. For instance, we demonstrate that with a particular referral distribution if the traffic intensity is kept constant, increasing no-show probability would improve performance.

Our experiments with the second model using data from UK specialty clinics show that this model provides a highly accurate representation of the system performance, and that making further assumptions with regard to the referral and capacity distributions would lead to serious errors in performance evaluation. We also demonstrate how this model can be used in practice for finding the appointment capacity needed for achieving a given waiting time target. To illustrate the application of our third model, we use the data provided for an MRI clinic in the US in the study of Green and Savin (2008) where the decision variable is considered to be the patient panel size rather than appointment capacity. Our experiments show that as the variability in the daily number of appointment requests decreases, the transition to unmanageable backlogs occurs at larger values of panel sizes but at a faster pace, i.e. over a narrower interval of panel size values. We provide further guidance on the reliability of the estimates produced assuming appointment requests follow a Poisson distribution.

Throughout this paper, we use the term 'waiting time' to refer to the wait outside the clinic as opposed to 'office wait', and 'cancellation' to refer to the appointments cancelled by the clinic rather than patients. The models in the paper are developed having the UK specialty clinics in mind. However, as demonstrated by the example of the MRI clinic in the US, they can be applied to almost all outpatient settings. There also exist a wide range of other public and private sector delivery systems that provide services based on appointments to which the models developed here could be applied. All the proofs are given in the online appendices.

2. Literature Review

Appointment scheduling has drawn considerable attention during the last 50 years or so. See Cayirli and Veral (2003) and Gupta and Denton (2008) for broad overviews of the literature plus challenges and opportunities. This literature can roughly be divided into two main streams: 'advance' or 'offline' scheduling and 'sequential' or 'online' scheduling. In advance scheduling, a given number of pre-booked non-urgent patients are scheduled during a clinic session so that a weighted combination of patients' office wait and the server's utilization (over/idle time) is minimized. Recent examples of this literature include Hassin and Mendel (2008), Koeleman and Koole (2012) and Cayirli et al. (2012). In sequential scheduling on the other hand appointment requests arrive gradually over time and the scheduler has to fit each patient to one of the available slots. See, for example, Gerchak et al. (1996), Patrick et al. (2008), Liu et al. (2010), and Feldman et al. (2014). Our investigation here is at a more strategic level and can be considered as a prerequisite to advance and online scheduling: we find the capacity needed for pre-booked non-urgent patients so that a reasonably quick access can be guaranteed.

More similar to our research are the studies by Jiang et al. (2012), Creemers and Lambrecht (2010), Kortbeek et al. (2014), Green and Savin (2008), and Liu and Ziya (2014). Jiang et al. (2012) propose an $M/D/1$ queue - with Poisson (M) arrivals and deterministic (D) service times - with state-dependent balking as a suitable model for specialty clinics. However, they do not consider no-shows and provide only partial characterization of the model. Considering cyclic appointment systems, Creemers and Lambrecht (2010) and Kortbeek et al. (2014) develop two-time scale queueing models, representing the daily evolution of the appointment backlog as well as the minute by minute dynamics of the clinic sessions. Neither of these studies considers clinic appointment cancellations and rescheduling of no-shows. Green and Savin (2008) derive the queue length distribution for an $M/D/1$ and an $M/M/1$ model with backlog-dependent no-show probability to investigate patient panel size decisions in primary care. Their study is the only paper that explicitly considers the possibility of no-shows rejoining the queue. Liu and Ziya (2014) use simplified versions of the models proposed in Green and Savin (2008) to jointly optimize patient panel size and overbooking decisions.

Similar to Green and Savin (2008), we consider backlog-dependent no-show probability and rescheduling of no-shows. However, our study differs from Green and Savin (2008) in the following ways. First, our models are flexible in terms of the referral (arrival) distribution while the models developed in Green and Savin (2008) are restricted to Poisson arrivals. Second, for our first model we derive the waiting time distribution of patients while Green and Savin (2008) only derive the queue length distribution. Third, we capture the stochastic variability in the supply of appointments caused by appointment cancellations. Overall, this is the first paper that provides a joint consideration of arbitrary referral and appointment cancellation distributions as well as patient no-show and re-scheduling behaviours.

The above contributions have been facilitated by using discrete-time (DT) models rather than continuous-time (CT) models used by Green and Savin (2008), Creemers and Lambrecht (2010) and Jiang et al. (2012). Apart from simplifying the analysis to a great extent, DT models, we believe, provide a more realistic representation for the dynamics of the specialty clinics. This is because in CT models new arrivals that find the server idle will start their service immediately, while in specialty clinics there is always a time lag between referral of a patient and her visit in the clinic even if empty slots are available. This time lag is largely due to the communication delays between referring sources, specialty clinics, and patients, and to some extent to the administrative or clinical tasks that must be performed on referral requests before seeing patients in the clinic. Even in primary care clinics with advanced access policy (Murray and Tantau 2000), there is always a delay between a patient's request for an appointment and her arrival to the clinic. These delays are captured by DT models, given an appropriate choice of the time interval, since in these models by definition arrivals during each time interval cannot start their services until at least the beginning of the next interval.

The DT queueing models developed in this paper are *bulk service* models with constant service times and customer no-show: bulk service as a batch of customers, whose size depends on the clinic appointment capacity as well as the number in the queue, is served in every time interval, constant service time as the service of every batch takes exactly one time unit, and customer no-show as customers may not turn up for service and subsequently rejoin the queue.

DT bulk service models have a long history in the queueing literature. The first DT bulk service model appeared in the literature is due to Bailey (1954), who was also motivated by the queues in the specialty clinics. Assuming that all patients would turn up, he derived the queue length distribution for the case where arrivals (referrals) were Poisson and the service capacity was a given constant. The waiting time distribution for the same model was later obtained by Downton (1955), and Boudreau et al. (1962) extended the model to arbitrary arrival distributions. The extension to the situation where capacity is also a random variable was given by Jaiswal (1961). Over the years, various other extensions to the original bulk service model, e.g. Alfa (1982) and Singh (1971), as well as its computational aspects, e.g. Bruneel (1993) and Janssen and Leeuwaarden (2005), have been investigated. In fact, quoting from Leeuwaarden (2005), ”The work done on the discrete bulk service queue runs to a large extent parallel to the maturing of queueing theory as a branch of mathematics”. However, as far as we are aware, no-shows have not been considered in any of the papers related to DT bulk service models. As such, apart from practical importance for capacity planning, our models contribute to the wider theory of DT bulk service queues.

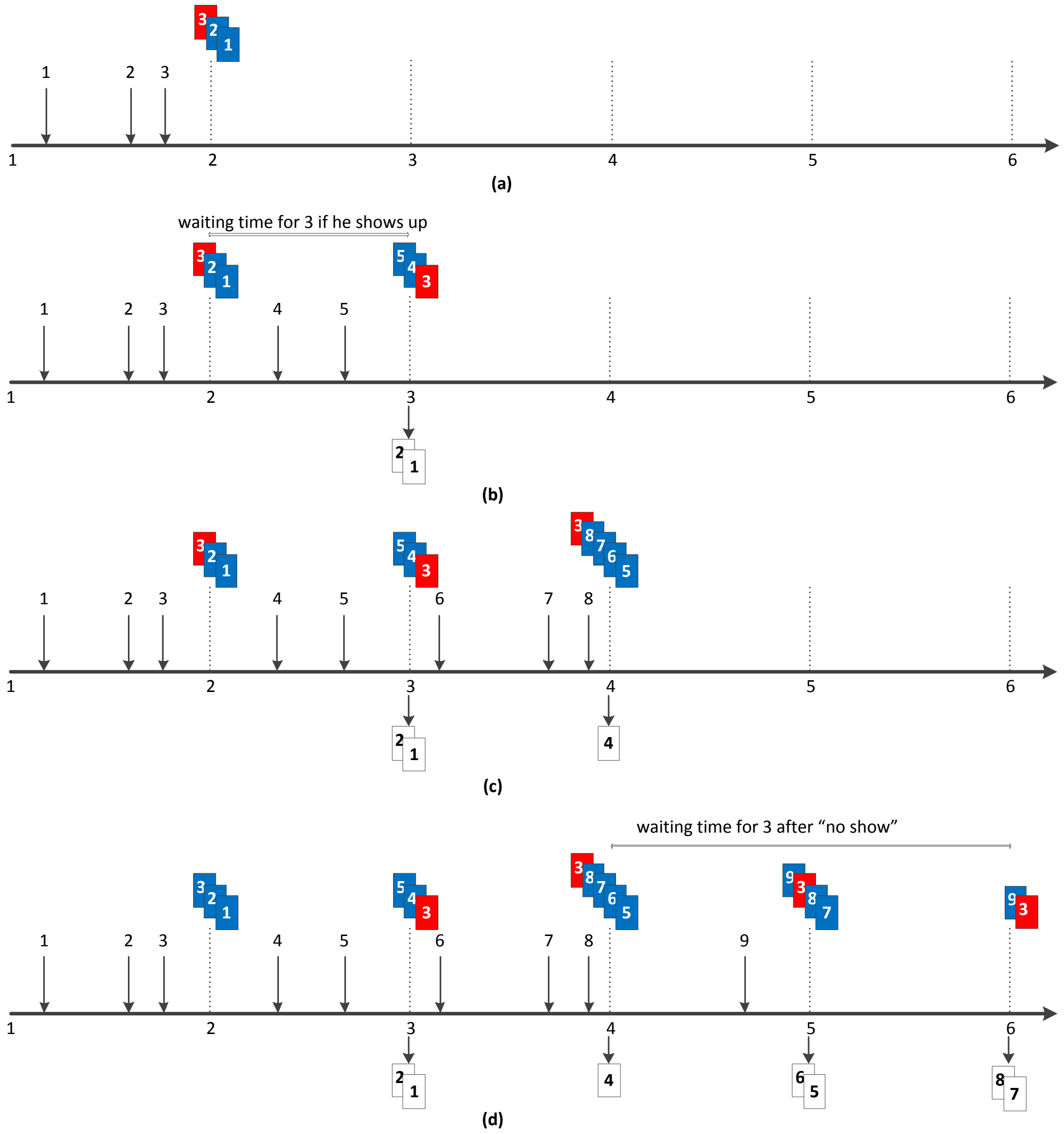
3. Modelling Framework

To illustrate the dynamics of the appointment backlog and waiting time measurements in specialty clinics, we start with a simple example (see Figure 2). Consider a specialty clinic that can always provide its nominal capacity of two appointments per unit time (which might be a day or several days). For simplicity assume the clinic starts empty. As illustrated in panel (a) of Figure 2, suppose three patients, 1,2, and 3, have been referred during the first time period. As explained in Section 2, these patients cannot start their visits until the beginning of the next time period even though the system is empty. The backlog will thus have three patients in the beginning of the second time interval. As illustrated in panel (b), patients 1 and 2 are served by the end of this time period (assuming both show up), two new patients are referred, and so three patients will be in the backlog in the beginning of the third time period. To measure the waiting times, we assume a clock starts at the end of the period when a patient is referred and stops in the beginning of the period when the patient is to be seen. The waiting times of patients 1, 2, and 3 will therefore be 0, 0, and 1 time unit assuming all show up.

But what happens when a patient does not make her appointment? According to the UK NHS waiting time calculation rules, "Where a patient fails to attend the first appointment after the initial referral that started their waiting time clock, their clock will be nullified (i.e. it is as if the referral never existed)" (Department of Health 2012). Then the provider has to decide whether it is more appropriate to return the patient back to the primary care, or a new appointment must be offered to the patient. In case of the latter, a new clock would start as soon as a new appointment is agreed and communicated to the patient.

To reflect the above situation, we assume no-shows either drop out of the system with a fixed probability, or rejoin the queue in the beginning of the next time period. Those who rejoin the queue are placed at the end of the backlog behind the most recent referrals, preserving their original order. To clarify this, focus on patient 3 in panel (b) of Figure 2, and suppose she misses her appointment in time three and does not drop out. She will then rejoin the queue in the beginning of time four, behind new referrals, 6, 7, and 8, as illustrated in panel (c). If patient 4 had missed her appointment and rejoined the backlog too, she would have been placed behind patient 3. A new waiting time clock starts on time four and stops on time six when patient 3 is supposed to turn up for her second chance, giving waiting time of two time units, as demonstrated in panel (d). The same process continues if the patient does not turn up for her appointment in time six.

Figure 2 A Simple Illustration of a Specialty Clinic.


(a)
(b)
waiting time for 3 if he shows up
(c)
(d)
waiting time for 3 after "no show"

The above example represents the complexities in modelling the specialty clinic system, in particular the difficulties of waiting time measurement for no-shows. The situation gets more complicated when we consider the variability in the supply of appointments, i.e. when clinic cannot match the pre-set nominal appointment capacity.

We model the system as a discrete-time bulk service model with no-shows as follows. The time axis is divided into equally-spaced intervals, numbered $1, 2, 3, \dots,$ where each interval corresponds to one or several days of clinic. The referrals can occur at any time during an interval but cannot start their service until at least the beginning of the next time interval, fitting our purpose as explained in Section 2. During each time interval, a maximum of N patients are served, implying a constant service duration of one time unit. In our first and third models, N is considered to be a deterministic variable while in the second model a probability distribution, based on the number of appointment cancellations, is assigned to it.

Throughout this paper, 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 $y_j \triangleq \mathbb{P}(Y=j)$. Furthermore, we define $Y(z) \triangleq \sum_{j=0}^{\infty} y_j z^j$ as the probability generating function (PGF) of Y. The PGF is known to be analytic for $|z|<1$ and continuous for $|z| \leq 1$.

For all models, we assume that each patient is given the first appointment available (first come first serve). We further assume that every patient does not show up for her appointment with probability $0 \leq \gamma < 1$. In our first two models, γ is a given constant, while in the third model it is an increasing function of the appointment backlog. Every no-show is assumed to rejoin the queue with a given probability $0 \leq r \leq 1$, independently of everything else in the model. The numbers of referrals in different time periods are assumed to be independent and identically distributed (i.i.d) as a non-negative discrete random variable R with $r_0 \triangleq \mathbb{P}(R=0) > 0$.

4. Model 1: Deterministic Capacity

In this model, we assume $\Pr(N=n)=1$, i.e. the appointment capacity is deterministic and equal to its nominal value n patients per unit time. Since the effective arrival rate is $\mu_R + \mu_R p + \mu_R p^2 + \cdots = \mu_R/q$ with $p=\gamma r$ and $q=1-p$, the stability condition is $\mu_R < nq$. We derive the queue length distribution in Section 4.1 and the waiting time distribution in Section 4.2.

4.1. Queue Length Distribution

Our objective here is to obtain the PGF of the stationary queue length distribution at the beginning of each time interval. Let X_t denote the queue length at the beginning of interval t just a moment before service begins for $t=1,2,\ldots$. Note that X_t is also the number of patients in the system (in service plus being queued) in the interval t after the service begins and before it ends. We have the following recursive equation

$$X_{t+1} = (X_t - n)^+ + D_t + R_t, \quad t = 1, 2, \ldots, \tag{1}$$

where $(x)^+ = \max\{x, 0\}$, R_t is the number of referrals during period t distributed as R, and D_t denotes the number of no-shows during period t that do not drop out, i.e. rejoin the queue. It is easy to see that $D_t \sim \text{Binomial}(\min\{n, X_t\}, p)$

Given the stability condition $\mu_R < nq$, let X be the stationary queue length with $x_j \triangleq \mathbb{P}(X = j) = \lim_{t\to\infty} \mathbb{P}(X_t = j)$. It proves convenient for the rest of our analysis to divide the PGF of X into two parts, $X_1(z)$ and $X_2(z)$, as defined below

$$\begin{aligned} X_1(z) &\triangleq \sum_{j=0}^{n-1} x_j z^j, \\ X_2(z) &\triangleq \sum_{j=0}^{\infty} x_{n+j} z^j, \end{aligned} \tag{2}$$

for complex $|z| \le 1$. Then $X(z) = X_1(z) + z^n X_2(z)$. The following proposition gives the PGF of X.

Proposition 1. *The PGF of the stationary queue length X in Model 1 is given by*

$$X(z) = R(z)\frac{z^n X_1(q+pz) - (q+pz)^n X_1(z)}{z^n - A(z)}, \quad |z| \le 1, \tag{3}$$

where $A(z) = (q+pz)^n R(z)$. Note that $A(z)$ is a PGF itself as it can be viewed as the PGF of a random variable A, defined as the convolution of R and a Binomial(n, p) *random variable.*

The PGF of the queue length distribution given in Proposition 1 depends on n unknown probabilities $x_0, x_1, \ldots, x_{n-1}$. However, these probabilities can be obtained using the zeros of the denominator in (3) that lie on or within the unit circle in the complex plane. The lemma below gives the number of zeros of the denominator.

Lemma 1. *Given $\mu_R < nq$ and finite μ_R, the equation $z^n - A(z)$ has n zeros on or within the unit circle.*

These zeros ($z = 1$ is an obvious one) can be obtained using one of the available software packages such as QROOT (Chaudhry 1991), or Maple. One can also use the fixed point iteration algorithm as in Kortbeek et al. (2014). Denoting the zeros by $z_0 = 1, z_1, \dots, z_{n-1}$, due to the convergence of $X(z)$ on and inside the unit circle and the fact that $R(z_k) \neq 0$, $k = 0, \dots, n-1$, (because otherwise $z_k = 0$ for some k which is a contradiction as $R(0) = r_0 \neq 0$ by assumption), we arrive at

$$z_k^n X_1(q+pz_k) - (q+pz_k)^n X_1(z_k) = 0, \quad k = 0, 1, \dots, n-1. \tag{4}$$

For $z_0 = 1$, the above equation has a trivial solution, but the normalizing condition $X(1) = 1$ provides an additional equation:

$$q \sum_{j=0}^{n-1} x_j (n-j) = nq - \mu_R,$$

where both sides of the equation give the expected unused service capacity. Combining the above with (4) and setting $z_k^n = A(z_k) = (q+pz_k)^n R(z_k)$ for $k = 1, \dots, n-1$ yield

$$\begin{cases} q \sum_{j=0}^{n-1} x_j (n-j) = nq - \mu_R, \\ \sum_{j=0}^{n-1} x_j \left((q+pz_k)^{n+j} R(z_k) - z_k^j (q+pz_k)^n \right) = 0, \quad k = 1, 2, \dots, n-1. \end{cases}$$

Since $(q+pz_k)^n \neq 0$ for all k (because otherwise $z_k = 0$ for some k which is clearly not a zero for the denominator assuming $r_0 > 0$ and $q > 0$), the above simplifies to

$$\begin{cases} q \sum_{j=0}^{n-1} x_j (n-j) = nq - \mu_R, \\ \sum_{j=0}^{n-1} x_j \left((q+pz_k)^j R(z_k) - z_k^j \right) = 0, \quad k = 1, 2, \dots, n-1. \end{cases}$$

This could be written in matrix form as $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{b} = (nq - \mu_R, 0, \dots, 0)^T$, $\mathbf{x} = (x_0, x_1, \dots, x_{n-1})^T$, and

$$a_{ij} = \begin{cases} (n-j+1)q, \quad i = 1, 1 \le j \le n, \\ R(z_{i-1})(q+pz_{i-1})^{j-1} - {z_{i-1}}^{j-1} \quad 2 \le i \le n, 1 \le j \le n. \end{cases}$$

Solving $\mathbf{Ax} = \mathbf{b}$ specifies the unknown probabilities in the PGF of X. Once the PGF is fully determined, one can find the remaining probabilities x_k, $k \geq n$, by numerically inverting the PGF $X(z)$ (see Abate and Whitt 1992a,b, and Abate et al. 1999 on discrete (fast) Fourier transform method, and Kim et al. 2011 on Taylor series expansion method). Alternatively, we can match the coefficients of both sides of

$$(z^n - A(z))X(z) = R(z)X_1(q+pz)z^n - A(z)X_1(z),$$

to obtain the remaining probabilities as

$$x_k = \frac{1}{a_0}\left(x_{k-n} - \sum_{i=1}^{k} a_i x_{k-i} - \sum_{i=0}^{k-n}\sum_{j=0}^{n-1} r_i x_j \binom{j}{k-i-n} (\frac{p}{q})^{k-i-n} q^j\right), \quad k \geq n,$$

where $a_i \triangleq \mathbb{P}(A = i) = \sum_{k=0}^{\min(i,n)} r_{i-k}\binom{n}{k} p^k q^{n-k}$.

For large systems, calculating the full range of x_k probabilities will be time-consuming. However, most of the important measures can be obtained from $x_0, x_1, \ldots, x_{n-1}$ probabilities. For instance, the corollary below gives the mean queue length and the PGF for the effective number of arrivals $E = R + D$, i.e. new referrals plus re-shows, in steady state.

Corollary 1. *For Model 1, we have*

$$\mu_X = \frac{q(1+p)\sum_{i=0}^{n-1} x_i(i^2 - n^2) + \mu_R^2 - \mu_R(2n+p) + qn^2(1+p) - \sigma_R^2}{2(\mu_R - nq)}, \tag{5}$$

and

$$E(z) \triangleq \mathbb{E}\left[z^E\right] = R(z)\left((q+pz)^n(1 - \mathbb{P}(X < n)) + X_1(q+pz)\right). \tag{6}$$

Application of the Little's law with effective arrival rate μ_R/q and mean queue length μ_X gives $\mu_X q/\mu_R$ as the mean waiting time in the system, which subtracted by one yields the mean waiting time in the queue. This may not be of much use for appointment capacity planning if the desired performance target is represented in terms of a percentile of the waiting time distribution rather than its mean. Besides, by treating re-shows as separate patients, the mean obtained in this way gives the average of all waiting times in the system including waiting times for missed appointments. Thus if the waiting times were calculated in a different way, e.g. if waiting times for missed appointments were excluded, it would not provide the correct mean. Below we develop a methodology for deriving the waiting time distribution which is flexible and can be used for a range of waiting time calculation rules.

4.2. Waiting Time Distribution

The major difficulty for calculating the waiting time distribution arises from the fact that the distribution of the number of patients in the queue seen by a new referral is not the same as the one seen by a re-show patient. To overcome this, we define $W^{(i)}$ as the conditional waiting time for the ith appointment for $i = 1, 2, \ldots$ given the first $(i-1)$ appointments are missed and rescheduled, and obtain its PGF in a recursive manner. Note that, as explained in Section 3, we assume that re-shows join the backlog at the end of their missed appointment periods, hence $W^{(i)}$ represents the time interval between a patient's referral and her first appointment periods for $i = 1$, and between the $(i-1)$th missed appointment and ith appointment periods for $i \geq 2$. We refer to $W^{(i)}$ as the ith conditional waiting time period (CWTP). Once the PGF of CWTP's are obtained, one could set up the appropriate equation that matches the waiting time rules used by the clinic to obtain the PGF of the overall waiting time. For instance, to match the UK NHS rules where waiting times for missed appointments are excluded, we set

$$W(z) \triangleq \mathbb{E}[z^W] = (1-p)\sum_{i=1}^{\infty} \mathbb{E}[z^{W^{(i)}}]p^{i-1}, \tag{7}$$

where W is the waiting time of a randomly chosen patient in steady state who eventually shows up. The analysis for finding the PGF of the CWTP's, i.e. $\mathbb{E}[z^{W^{(i)}}]$ for $i = 1, 2, \ldots$, is technical and presented in full detail in the online appendices.

5. Model 2: Stochastic Capacity

In this model, we assume the nominal appointment capacity is set to a constant value n but V_t appointments are cancelled in each time period t, independently of anything else, due to operational reasons such as consultants' delays and absenteeism or unpredicted high level of urgent workload.

This gives rise to the realized appointment capacity of $N_t = n - V_t$, where V_t are assumed to be i. i. d as a random variable V with finite support $\{0, 1, \ldots, n\}$ and $\mathbb{P}(V = 0) > 0$; hence, N_t will be i. i. d as a random variable N with finite support $\{0, 1, \ldots, n\}$ and $n_n \triangleq \mathbb{P}(N = n) > 0$.

In reality, the patients whose appointments are cancelled by the clinic must be given a new appointment within a short time period while their waiting time clock continues ticking. This in our modelling framework is equivalent to patients returning to the queue, not necessarily to the end of the queue as in the case of no-shows, but to the earliest possible slot. This makes the waiting time calculations more difficult than in Model 1, hence we only obtain the queue length distribution for this model. As in Model 1, denoting the queue length at the beginning of interval t before the service begins by X_t, we will have the following recursive equation

$$X_{t+1} = (X_t - N_t)^+ + D_t + R_t, \quad t = 1, 2, \ldots, \tag{8}$$

where $D_t \sim \text{Binomial}(\min\{N_t, X_t\}, p)$. The above equation implies that patients whose appointments are cancelled remain in the queue. The stability condition for this model is $\mu_R < \mu_N q$, where $q = 1 - p$ and $p = \gamma r$.

To obtain the PGF of X, the stationary queue length distribution, we first extend our definitions of $X_1(z)$ and $X_2(z)$ in (2) to account for the variability in the service capacity as follows.

$$X_{1,c}(z) = \sum_{j=0}^{c-1} x_j z^j,$$
$$X_{2,c}(z) = \sum_{j=0}^{\infty} x_{c+j} z^j,$$

for complex $|z| \leq 1$ and $c = 0, \ldots, n$, where $X_{1,0}(z)$ is assumed to be 0. Then $X(z) = X_{1,c}(z) + z^c X_{2,c}(z)$. We have the following proposition and Lemma.

Proposition 2. *The PGF of the stationary queue length distribution X in Model 2 is given by*

$$X(z) = R(z) z^n \frac{\sum_{c=0}^{n} X_{1,c}(q + pz) n_c - \sum_{c=0}^{n} \left(\frac{q+pz}{z}\right)^c X_{1,c}(z) n_c}{z^n - A(z)}, \quad |z| \leq 1, \, n \geq 1, \tag{9}$$

where $A(z) = R(z) \sum_{c=0}^{n} z^{n-c} (q + pz)^c n_c$. Note that $A(z)$ is a PGF itself as it can be viewed as the PGF of a random variable A, defined as the convolution of R and a discrete random variable G with probabilities $g_i = q^{n-i} \sum_{c=0}^{i} \binom{c+n-i}{c} p^c n_{c+n-i}$ for $i = 0, 1, \ldots, n$.

Lemma 2. *Given $\mu_R < \mu_N q$ and finite μ_R, the equation $z^n - A(z)$ has n zeros on or within the unit circle.*

Let $z_0 = 1, z_1, \ldots, z_{n-1}$ be the roots of $z^n - A(z) = 0$ on and within the unit circle. Following the same argument as in Model 1, we arrive at the following set of equations.

$$\sum_{c=0}^{n} X_{1,c}(q+pz_k)n_c - \sum_{c=0}^{n} \left(\frac{q+pz_k}{z_k}\right)^c X_{1,c}(z_k)n_c = 0, \quad k = 0,1,\ldots,n-1. \tag{10}$$

For $z_0 = 1$, the above equation has a trivial solution, but the normalizing condition $X(1) = 1$ provides an additional equation:

$$\sum_{j=0}^{n-1} x_j \sum_{c=j+1}^{n} (c-j)qn_c = q\mu_N - \mu_R,$$

where both sides of the equation give the expected unused service capacity. Expanding (10) and combining it with the above equation yield

$$\begin{cases} \sum_{j=0}^{n-1} x_j \sum_{c=j+1}^{n} (c-j)qn_c = q\mu_N - \mu_R, \\ \sum_{j=0}^{n-1} x_j \left((q+pz_k)^j \mathbb{P}(N \geq j+1) - {z_k}^j \sum_{c=j+1}^{n} \left(\frac{q+pz_k}{z_k}\right)^c n_c\right) = 0, \quad k = 1,2,\ldots,n-1, \end{cases}$$

which could be represented in matrix form as $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{b} = (q\mu_N - \mu_R, 0, \ldots, 0)^T$, $\mathbf{x} = (x_0, x_1, \ldots, x_{n-1})^T$, and $\mathbf{A} = [a_{ij}]$ with

$$a_{ij} = \begin{cases} \sum_{c=j}^{n} (c-j+1)qn_c, \quad i = 1, 1 \leq j \leq n, \\ (q+pz_{i-1})^{j-1} \Pr(N \geq j) - {z_{i-1}}^{j-1} \sum_{c=j}^{n} \left(\frac{q+pz_{i-1}}{z_{i-1}}\right)^c n_c, \quad 2 \leq i \leq n, 1 \leq j \leq n. \end{cases}$$

Having solved the above equation, similar to Model 1, one can obtain the remaining probabilities $x_k, k \geq n$ through matching the coefficients of both sides of Equation (9) as follows

$$\begin{aligned} x_k = \frac{1}{a_0}\Bigg(& x_{k-n} - \sum_{i=1}^{k} a_i x_{k-i} \\ & - \sum_{j=\max(k-2n+1,0)}^{k-n} \sum_{i=k-n-j}^{n-1} \sum_{c=i+1}^{n} r_j x_i n_c \left(\frac{q}{p}\right)^{j+n+i-k} \left[p^i \binom{i}{j+n+i-k} - p^c \binom{c}{j+n+i-k}\right] \\ & + \sum_{j=k-n+1}^{k} \sum_{i=0}^{k-j} \sum_{c=j+n+i-k}^{n} r_j x_i n_c p^c \binom{c}{j+n+i-k} \left(\frac{q}{p}\right)^{j+n+i-k} \Bigg), \end{aligned}$$

for $k \geq n$, where

$$a_i \triangleq \mathbb{P}(A=i) = \sum_{k=0}^{\min(i,n)} \sum_{c=0}^{k} r_{i-k} \binom{c+n-k}{c} p^c n_{c+n-k} q^{n-k}, \quad \text{for } i=0,1,\ldots.$$

The following corollary gives the mean queue length and the PGF for the effective number of arrivals $E = R + D$ in steady state.

COROLLARY 2. *For Model 2,*

$$\mu_X = \Bigg(q\left(-\mu_R - q^2\sigma_N^2 - q^2\mu_N^2 + q\mu_R - \sigma_R^2 + \mu_R^2\right) \sum_{i=0}^{n-1} \sum_{c=i+1}^{n} x_i n_c (i-c) \\ + q(q\mu_N - \mu_R) \sum_{i=0}^{n-1} \sum_{c=i+1}^{n} x_i n_c (-2i^2 + qi^2 + 2ci - qc^2) \Bigg) \Big/ \left(2(\mu_R - q\mu_N)^2\right) \quad (11)$$

and

$$E(z) \triangleq \mathbb{E}\left[z^E\right] = R(z) \sum_{c=0}^{n} \left[X_{1,c}(q+pz) + (q+pz)^c (1 - \mathbb{P}(X<c))\right] n_c. \quad (12)$$

6. Model 3: Backlog-Dependent No-Show

In this section, we generalize Model 1 to consider situations where no-show probability of each patient depends on how long she has waited in the queue. This is motivated by the study of Green and Savin (2008). They present empirical data from two clinics in the US, an MRI clinic and a mental health clinic, where no-show rates grow with the delays in the appointment queue. To capture this, they develop continuous-time $M/D/1$ and $M/M/1$ models with no-show probability defined 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.

Following the same argument as in Green and Savin (2008), we define no-show probability as an increasing function of the number of patients left behind by a departing patient. However, since in our discrete-time model departures occur in batches, we cannot differentiate among no-show probabilities of various members of a batch. As such, we develop bounds for performance metrics using the no-show probabilities of the first and last departing patient in a batch as the no-show probabilities of all patients in that batch. Specifically, let $\gamma((X_t-1)^+)$ $(\gamma((X_t-n)^+))$ represent the no-show probability of all departing patients at the end of interval t for the upper (lower) bound model, representing the more (less) congested system, and let $p(i) = r\gamma(i)$. We then have

$$X_{t+1} = min\{(X_t - n)^+ + D_t + R_t, k\}, \quad t = 1, 2, \ldots, \tag{13}$$

where k is the maximum system capacity and $D_t \sim \text{Binomial}(\min\{n, X_t\}, p((X_t - 1)^+))$ for the upper bound model, and $D_t \sim \text{Binomial}(\min\{n, X_t\}, p((X_t - n)^+))$ for the lower bound model. Note that for analytical tractability we have set a limit on the total number of patients allowed in the system for this model. Further note that the patients referred during a time interval are not counted in the no-show probability function of departing patients at the end of that interval.

The structure of the equation above makes it difficult to follow the PGF approach. Instead, we find the steady-state probabilities for X_t, denoted by $\mathbf{x} = (x_0, x_1, \ldots, x_k)$, by solving the stochastic balance equation $\mathbf{x} = \mathbf{x}\boldsymbol{\phi}$, where $\boldsymbol{\phi} = [\phi_{ij}]$ represents the one-stage transition probabilities as below.

Proposition 3. *The transition probabilities of the discrete-time Markov chain specified by* (13) *for the upper bound model are*

$$\phi_{ij} = \begin{cases} 0, \qquad j < i - n, \\ (p((i-1)^+))^{j-(i-n)^+} (q((i-1)^+))^{i-j} \sum_{l=(j-i)^+}^{j-(i-n)^+} \binom{\min\{n,i\}}{j-(i-n)^+-l} \left(\frac{q((i-1)^+)}{p((i-1)^+)}\right)^l r_l, \qquad i-n \leq j < k, \\ (q((i-1)^+))^{\min\{i,n\}} \sum_{l=k-i}^{k-(i-n)^+} r_l \sum_{m=k-(i-n)^+-l}^{\min\{i,n\}} \binom{\min\{i,n\}}{m} \left(\frac{p((i-1)^+)}{q((i-1)^+)}\right)^m, \qquad j = k, \end{cases} \tag{14}$$

where $q(i) = 1 - p(i)$. For the lower bound model, replace $p((i-1)^+)$ and $q((i-1)^+)$ with $p((i-n)^+)$ and $q((i-n)^+)$, respectively.

The bounds for performance metrics can then be calculated from steady-state probabilities.

7. Analysis and Results

In this section, we present some empirical results obtained from the models using both illustrative and real data. In Section 7.1, we use Model 1 with illustrative data to get some insight on the impact of no-shows on the system performance. In Section 7.2, we apply Model 2 to the real data obtained from clinics A and C to validate the model and also demonstrate the application of the model in practice for identifying the appointment capacity. In Section 7.3, we use the data provided by Green and Savin (2008) to investigate application of Model 3 in setting patient panel sizes.

7.1. Illustrative Analysis with Model 1

Throughout this section, we assume the clinic appointment capacity N is deterministic and equal to its nominal value n, all no-shows reschedule, i.e. $r = 1$, and the time unit is one day. To demonstrate the impact of various referral distributions, we consider three different distributions, Polya(b,α) (a Negative Binomial distribution with real stopping-time parameter b), Poisson(λ), and Binomial(m,α) where $b,\lambda > 0$, $0 < \alpha < 1$, and m is a positive integer. The PGF, mean and variance of these distributions are given below.

$$\begin{array}{llll}
\text{Polya:} & R(z) = \left(\dfrac{1-\alpha}{1-\alpha z}\right)^b, & \mu_R = \dfrac{\alpha b}{1-\alpha}, & \sigma_R^2 = \dfrac{\alpha b}{(1-\alpha)^2} \\
\text{Poisson:} & R(z) = \exp(\lambda(z-1)), & \mu_R = \lambda & \sigma_R^2 = \lambda \\
\text{Binomial:} & R(z) = (1-\alpha+\alpha z)^m, & \mu_R = m\alpha, & \sigma_R^2 = m\alpha(1-\alpha).
\end{array}$$

It is clear from above that the variance exceeds the mean for Polya distribution, equals the mean for Poisson distribution, and is smaller than the mean for Binomial distribution.

In our first experiment, as illustrated in Table 2, all parameters have remained constant except for the no-show probability. Note that parameters of the Polya (Binomial) distribution are set so that its squared coefficient of variation (SCV) is almost three times (a third) of that of the Poisson distribution. The corresponding plots for mean queue length and the 95th percentile of waiting time distribution are given in Figure 3. For waiting time calculations, we used Equation (7) and assumed that patients will make either the first, second, or third appointments. In Table 3, we have also represented the mean of the first, second, and third CWTPs for all test cases.

The first observation is that congestion increases substantially with the variability of the referral distribution. This, along with the fact that referral distributions are typically more variable than Poisson as illustrated in Table 1, indicates that the commonly made assumption of Poisson referrals

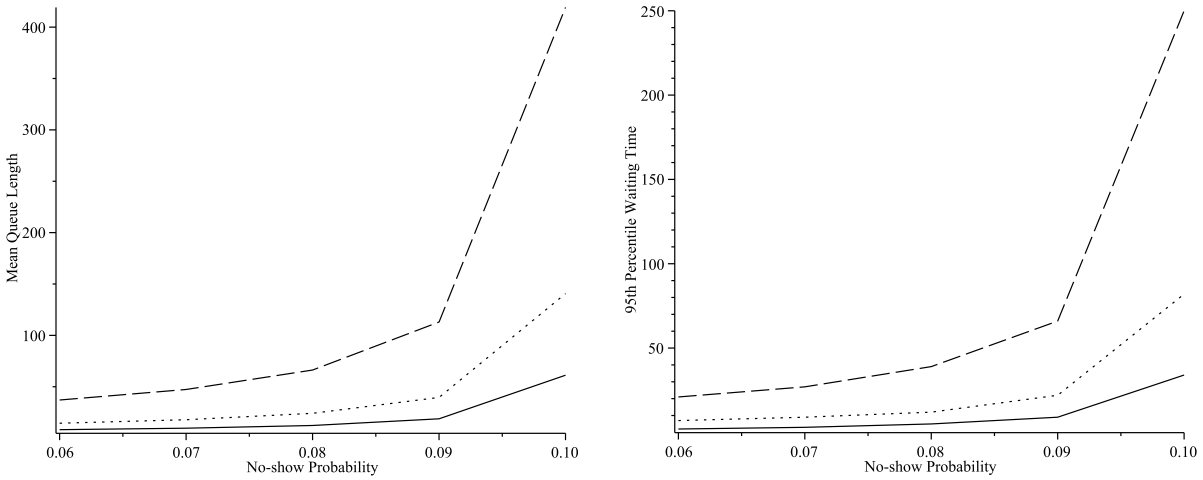
Table 2 Summary of the Parameters Used in the First Experiment in Section 7.1.

Characteristic	Test 1	Test 2	Test 3	Test 4	Test 5
Nominal Capacity			5		
Referral Average			4.482		
Referral SCV for Polya $(2, 0.6915)$			0.7231		
Referral SCV for Poisson(4.482)			0.2231		
Referral SCV for Binomial$(7, 0.6403)$			0.0802		
No-show Probability	0.06	0.07	0.08	0.09	0.1
Traffic Intensity	0.9536	0.9639	0.9743	0.985	0.996

could seriously under-estimate the capacity needed for achieving a given target. For instance, suppose that in a clinic with no-show probability of 0.09 achieving a waiting time's 95th percentile of 21 days is desirable. Dotted line in the right panel of Figure 3 suggests that assuming a Poisson referral distribution setting the capacity $n = 5$ should be enough. However, dashed line in the same plot suggests that this would lead to a waiting time's 95th percentile of 66 days if the referrals are actually distributed as a Polya distribution with the same average but three times larger SCV. The second observation is that the impact of no-show probability on the performance tends to be more significant with more variable referral distributions. The third observation is that according to Table 3, under our particular rescheduling policy, patients' waiting times increases when they miss their first appointments. The extent of this increase declines with increasing no-show probability and variability of the referral distribution. From the second appointment onwards, however, missing further appointments could lead to longer (for Poisson and Binomial referrals) or shorter (for Polya referrals) waiting times.

In the second experiment, we investigate the impact of varying no-show probabilities on systems with constant traffic intensities. In particular, we assume the referral average drops in response to an increase in the no-show probability such that traffic intensity remains at 0.98 in all cases. The list of the parameters used in this experiment as well as a summary of the results are given in Table 4. Note that the SCV for the Polya (Binomial) referral distribution is also fixed at 0.5 (around 0.12). The results show that with increasing no-show probability the performance degrades with Poisson and Binomial referrals and improves with Polya referrals. Since for each of these three cases, the mean for effective arrival distribution has remained constant at $\mu_R/q = 4.95$ and its SCV

Figure 3 Performance Metrics for the First Experiment in Section 7.1.



Note. Dashed line: Polya $(2, 0.6915)$, Dotted line:Poisson(4.482), Solid line: Binomial$(7, 0.6403)$

Table 3 Conditional Waiting Time Means for the First Experiment in Section 7.1.

Characteristic	Test 1	Test 2	Test 3	Test 4	Test 5
Binomial Referrals					
1st CWTP	0.69	0.97	1.48	2.78	11.33
2nd CWTP	1.26	1.53	2.05	3.36	11.86
3rd CWTP	1.35	1.64	2.13	3.47	12.06
4th CWTP	1.48	1.75	2.30	3.59	12.2
Poisson Referrals					
1st CWTP	2.01	2.67	3.90	6.99	27.18
2nd CWTP	2.55	3.24	4.47	7.59	27.74
3rd CWTP	2.60	3.29	4.56	7.65	27.79
4th CWTP	2.68	3.37	4.60	7.68	27.93
Polya Referrals					
1st CWTP	6.78	8.84	12.66	22.22	88.10
2nd CWTP	7.34	9.41	13.25	22.79	88.87
3rd CWTP	7.19	9.27	13.12	22.66	88.51
4th CWTP	7.06	9.14	13.01	22.57	88.53

has decreased with increasing no-show probabilities, this may suggest that the impact of re-shows goes beyond the first and second moments of effective arrival distribution.

The results given in Table 4 has two implications. First, estimating the impact of no-shows by inflating the demand according to the no-show and rescheduling probabilities would result in over/under-estimation. For instance, according to Table 4, estimating the performance of a system with $\mu_R = 4.018$, $n = 5$, $p = 0.18$ using a discrete bulk service model without no-shows but with adjusted referral rate equal to $\mu_R/q = 4.9$ would over (under)-estimate mean queue length by about 12 percent for Polya (Binomial) referrals.

Table 4 Summary of the Parameters and Results for the Second Experiment.

Characteristic	Test 1	Test 2	Test 3	Test 4
Appointment Capacity	5	5	5	5
Referral Average	4.9	4.606	4.018	2.45
No-show Probability	0	0.06	0.18	0.5
Traffic Intensity	0.98	0.98	0.98	0.98
Polya Referrals				
Referral Variance	12.005	10.6076	8.0722	3.0012
Referral SCV	0.5	0.5	0.5	0.5
Effective Arrival SCV	0.5	0.4533	0.3667	0.1786
Mean Queue Length	63.4598	61.2651	56.8859	45.2846
Poisson Referrals				
Referral Variance	4.9	4.606	4.018	2.45
Referral SCV	0.2041	0.2172	0.2489	0.4082
Effective Arrival SCV	0.2041	0.2034	0.1978	0.1555
Mean Queue Length	28.2599	29.6512	32.4329	39.8506
Binomial Referrals				
Referral Variance	2.8992	2.4845	1.9999	0.9494
Referral SCV	0.1207	0.1172	0.1239	0.1582
Effective Arrival SCV	0.1207	0.1150	0.1137	0.0926
Mean Queue Length	18.4212	18.5639	20.3569	25.1507

The SCV for effective arrvials has been calculated using Equation (6)

The second implication is that reducing no-show probability in response to increasing referral rate such that traffic intensity remains constant may or may not be enough for recovering performance depending on the referral distribution. To understand this, consider a clinic which in response to an increase in the referral rate from 4.018 to 4.606 has decided to invest in reducing the no-show probability from 0.18 to 0.06 in order to keep the traffic intensity and thus the congestion at the original level. The results presented in Table 4 suggest that this would not be enough for Polya referrals with the mean queue length being still higher than its original value, and hence further decrease in the no-show probability or an increase in the capacity is needed. On the other hand, for Poisson and Binomial referrals even a smaller improvement in the no-show probability would be enough for recovering the performance.

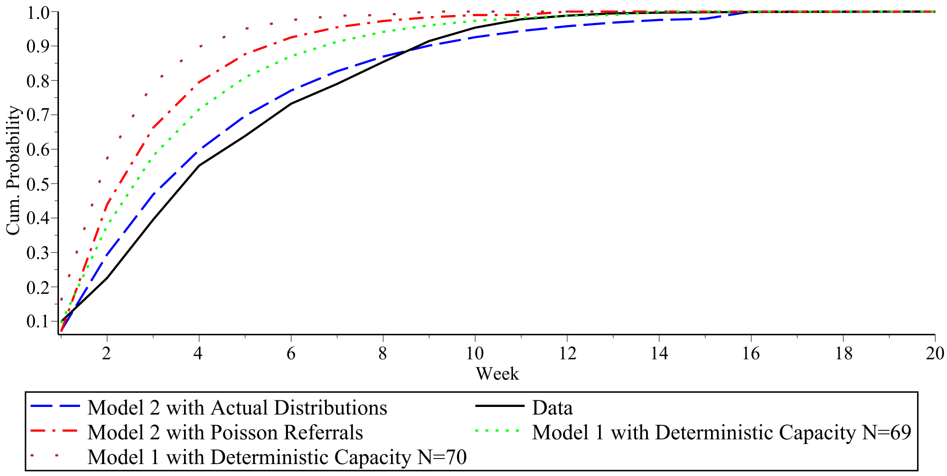
7.2. Realistic Analysis with Model 2

In this section, we first apply Model 2 to the historical data obtained from clinic A to validate the model. We then use Model 2 to investigate the appointment capacity requirement for patients referred to clinic C in order to achieve a specific performance target.

The input parameters required for our models are referral distribution, realized capacity distribution (or equivalently the nominal appointment capacity and the distribution for the number of appointments cancelled by the clinic), no-show probability, and rescheduling probability. A brief explanation on how these parameters are obtained from the data follows. For each appointment entry, we have access to a range of attributes including a unique patient ID, referral date, appointment date, and appointment outcome (attended, missed, cancelled by patient, cancelled by hospital). We work with a weekly time unit to reflect the communication delays and also to remove the day-of-week impact from our distributions. To obtain the referral distribution, we count the number of new referrals per week excluding those whose appointments are cancelled (by hospital or themselves) and return to primary care. This is mainly because they do not use any resources. Assuming that all slots are filled, adding up the number of missed and attended appointments in each week gives the realized capacity of that week. This added up with the number of appointments cancelled each week yields the nominal appointment capacity of that week. The no-show probability is obtained by dividing the total number of missed appointments by the total number of appointments supplied (attended plus missed) over the entire period. Those who miss their appointments and are not given further appointments are assumed to have returned to primary care; their proportion out of total missed appointments gives the complement of rescheduling probability.

For validation experiment, we use clinic A data over a one year period starting from March 2012, as summarized in Table 1. Weekly estimates for the realized and nominal appointment capacity are calculated in the way explained above. As the values obtained for nominal capacity vary slightly over time (the average nominal capacity is 74.3 appointments per week), we work directly with the realized capacity distribution as summarized in Table 1. Using this along with the actual referral distribution, no-show and re-scheduling probabilities in Model 2, we obtain the queue length distribution. The results are depicted in Figure 4, which show a very good match between model results, represented in terms of the number of appointment weeks as an estimate for waiting times, and the actual waiting time distribution obtained from the data. In this figure, we have also plotted the queue length distributions obtained from (i) Model 2 with realized capacity distribution but with Poisson referrals, (ii) Model 1 with actual referral distribution but with deterministic capacity $N = 69$ (the average realized capacity rounded down), and (iii) Model 1 with actual referral distribution but with deterministic capacity $N = 70$ (the average realized capacity rounded up). All of these over-estimate the performance, emphasizing the importance of combined modeling of arbitrary capacity and referral distributions.

Figure 4 **Comparison of the Queue Length Distribution Obtained from the Models and from the Data for Clinic A.**



We now use the appointment data for a class of non-urgent patients called routine patients referred to clinic C during the first six months of 2013 to plan for the following six months. Assuming that the referral and cancellation distributions and no-show and rescheduling probabilities remain the same, we aim to set the nominal appointment capacity so that 95% of the patients are seen within 6 weeks of referral, as requested by the clinic. A summary of the data extracted from the first six months of 2013 is given in Table 5.

The stability condition $\mu_R < (n - \mu_V)(1 - \gamma r)$ indicates that, with the existing parameters, the minimum value for the nominal appointment capacity n to achieve steady state is 122 slots per week which yields a traffic intensity of 0.99. Applying Model 2 with $n = 122$ and actual cancellation distribution, no-show and rescheduling probabilities would then give a 95th percentile of ten weeks

Table 5 Summary of the Parameters Extracted from Clinic C Data.

	Referral Distribution	Clinic Cancellation Distribution	No-show Probability	Re-scheduling Probability
Mean	103.93	8.59		
Var	570.13	47.36	0.076	0.996
SCV	0.053	0.64		

Time unit is one week

Table 6 Predicted Waiting Times's 95th Percentile for Clinic C.

No-show Probability		0.075	0.068	0.06
Current Referral Distribution	Current Cancellation Distribution	10	6	4
	New Cancellation Distribution	9	5	4
Poisson Referral Distribution	Current Cancellation Distribution	4	3	2
	New Cancellation Distribution	3	2	1

Time unit is one week

which does not satisfy the clinic target. Experimenting with larger values for nominal capacity shows that $n = 124$ is the smallest value that satisfies the waiting time target.

To investigate the impact of reducing cancellations, we set the nominal capacity to 114 but assume either 0 or 1 cancellation will occur with respective probabilities 0.558 and 0.412. This gives the same traffic intensity as the case with $n = 122$ but the waiting time's 95th percentile will drop to nine weeks as a result of less variability in the capacity distribution. We have performed a range of experiments with these two different cancellation distributions under different no-show probabilities and with actual as well as Poisson referrals. The results presented in Table 6 would provide further guidance for the clinic as to how the performance could be improved.

7.3. Panel Size Analysis with Model 3

So far the focus has been on specialty care clinics where appointment capacity is the major decision variable. In primary care clinics as well as some other outpatient facilities however patient panel size is also an important variable. Panel size is defined as the total number of patients who visit a particular doctor or practice on a regular basis. Finding the right panel size has been an active area of research over the recent years, see e.g. Altschuler et al. (2012), in particular in the US where healthcare reforms are expected to create millions of additional primary care visits per year (Hofer et al. 2011). In this section we show how the models we have developed, in particular Model 3 with delay-dependent no-show behaviour, can be used for determining optimal panel sizes.

We use the data provided by Green and Savin (2008) for an MRI clinic in the US. They fit the exponential function

$$\gamma(i) = \gamma_{max} - (\gamma_{max} - \gamma_{min})e^{-i/C}$$

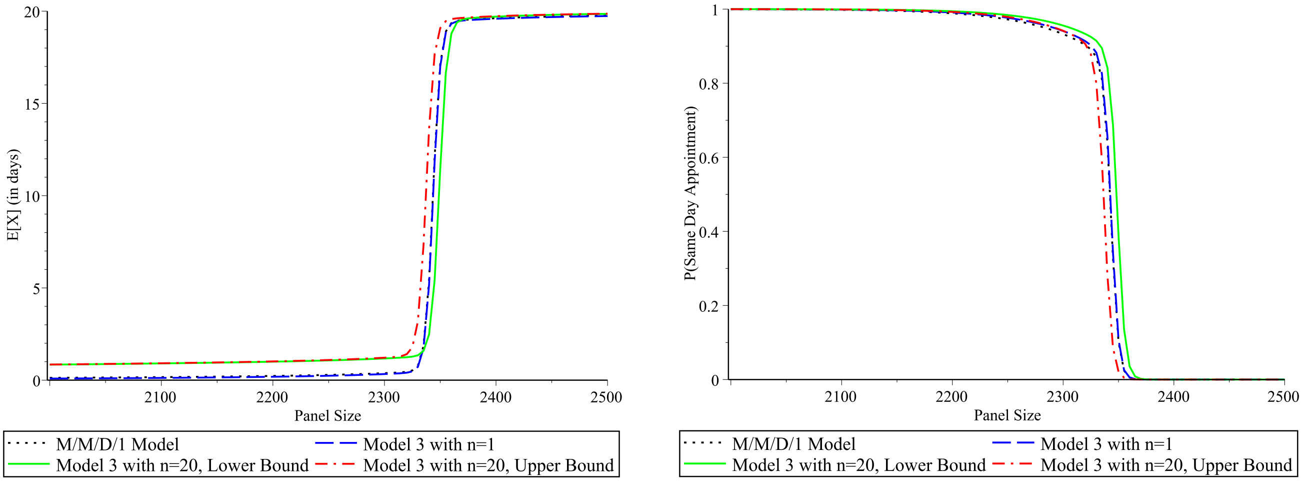
with $\gamma_{max} = 0.31$,$\gamma_{min} = 0.01$ and $C = 50$ days to the observed values for the no-show probability in terms of the appointment backlog i seen at the time of appointment request. They make the assumption that the number of appointment requests per day follows a Poisson distribution with parameter $0.008 \times s$, where s is the panel size, independently of the number of patients already in the system. They further assume that the clinic has a capacity of 20 visits per day, and the maximum system capacity is 400.

To apply Model 3, we need to choose an appropriate time unit. Anything between one service slot (1/20 of a day) up to an entire day, depending on the delay expected between request for an appointment and arrival to the clinic, sounds reasonable. For the former case, the system capacity $n = 1$, so both lower and upper bound models would produce the same results. For the latter however $n = 20$ and so the upper and lower bound models would be different. The performance measures we are interested in are the average number of patients in system, $\mathbb{E}[X]$, and the probability of obtaining a same-day appointment, which we define as $\mathbb{P}((X - n)^+ \leq 20)$.

We first compare the results obtained from Model 3 with $n = 1$ and $n = 20$ (both lower and upper bound cases) with those from $M/D/1/k$ model of Green and Savin (2008), assuming Poisson arrivals. The plots are presented in Figure 5 which show a perfect match between the continuous-time model and Model 3 with $n = 1$. The system however is more congested with $n = 20$ when panel sizes are small. This is due to the one day delay between request for an appointment and start of service in the clinic with daily time units, whose impact is more pronounced when the system is quiet.

Next we relax the Poisson assumption. To do so, we set the time unit to one service slot, i.e. $n = 1$, and assume the number of appointment requests per interval follows a discrete Weibull

Figure 5 **Expected Number of Customers in System (Left Panel), and Probability of Obtaining a Same-Day Appointment (Right Panel).**

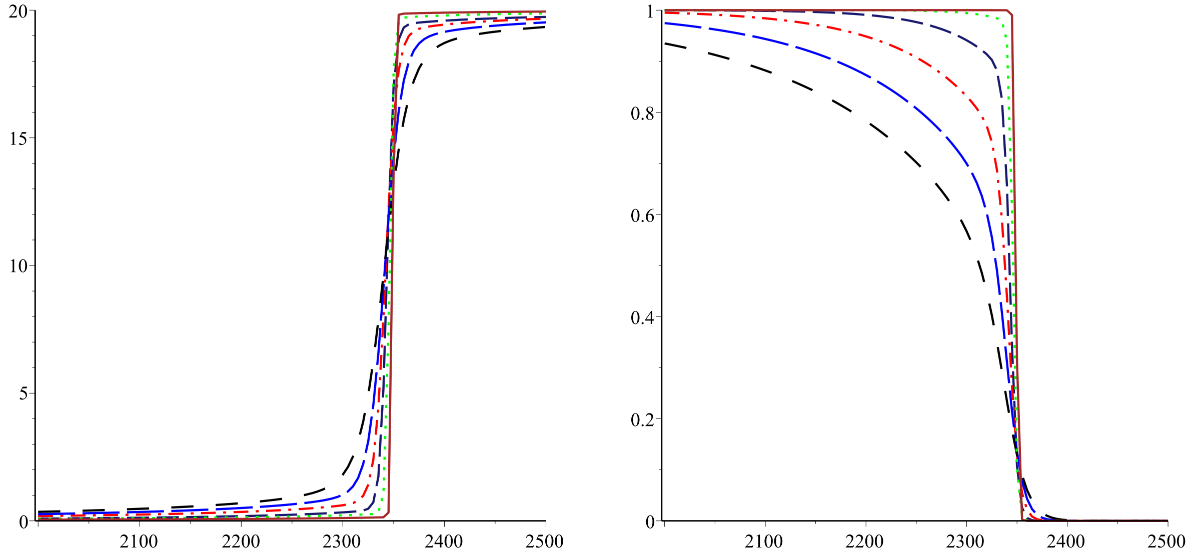


distribution proposed in Nakagawa and Osaki (1975) due to its flexibility. The probability mass function for this distribution is

$$\mathbb{P}(R=i)=\alpha^{i^b}-\alpha^{(1+i)^b}, \quad i=0,1,\dots$$

where $0<\alpha<1$ and $b>0$. We set the average number of appointment requests to $0.008s/20$ per interval as before, but experiment with standard deviations that are 0.5, 0.75, 1.25, 1.50, and 1.75 of Poisson standard deviation, i.e $\sqrt{0.008s/20}$. The results are illustrated in Figure 6. The plots show that as the variability of the arrival distribution decreases, transition to unmanageable backlogs occurs at larger values of panel sizes but at a higher pace, i.e. over a narrower interval of panel size values. In Table 7, we have also displayed the corresponding optimal panel sizes for achieving the same-day appointment probability of 0.75 needed for implementing advanced access policy (Murray and Tantau 2000). The table shows that a Poisson assumption would produce good estimates of panel sizes when the actual distribution is less variable than Poisson. For more variable distributions, however, the difference is more substantial. In particular, setting panel sizes based on a Poisson assumption, i.e. $s=2337$, when the distribution is actually 25%, 50% and 75% more variable than Poisson would lead to the same-day appointment probability of 0.54, 0.37, and 0.27, respectively.

Figure 6 **Expected Number of Patients in System (Left Panel), and Probability of Obtaining a Same-Day Appointment (Right Panel).**



Note. Discrete Weibull arrivals with standard deviations that are 0.50 (solid line), 0.75 (dotted line), 1.25 (dash-dotted line), 1.50 (long-dashed line) and 1.75 (space-dashed line) of that of Poisson, and Poisson arrivals (dashed line)

Table 7 **Panel Sizes Required for Achieving Same-Day Appointment Probability of** 0.75.

Distribution	Weibull	Weibull	Poisson	Weibull	Weibull	Weibull
Standard Deviation	$0.5\sqrt{\frac{0.008s}{20}}$	$0.75\sqrt{\frac{0.008s}{20}}$	$\sqrt{\frac{0.008s}{20}}$	$1.25\sqrt{\frac{0.008s}{20}}$	$1.5\sqrt{\frac{0.008s}{20}}$	$1.75\sqrt{\frac{0.008s}{20}}$
Panel Size	2348	2343	2337	2323	2280	2222

8. Conclusions

Specialty clinics operate in a complex environment where highly variable referral rates, frequent clinic appointment cancellations, small to medium no-show probabilities, and high rates of no-shows rescheduling are dominant features. In this paper, we develop queueing models that take these features collectively into account, and use them to find the appointment capacity needed for achieving performance targets set for non-urgent patients. Although the models developed here are stylized in nature (as they do not take patient choice into account), the experiments performed with real data demonstrate their accuracy. We also show the application of our models in making panel size decisions in clinics where appointment requests follow an arbitrary distribution and no-show probability increases with appointment backlog. The computation time for the case studies covered in this paper are reasonable, considering that capacity analysis is typically performed only once every few weeks. As such, we believe the three models developed here could be incorporated into a powerful and reliable tool for clinic managers, facilitating the process of capacity planning.

We end the paper by noting that further adjustment might be needed on the appointment capacity obtained from our models as the capacity left open for urgent streams of demand may not be sufficient, hence producing longer overtime hours and higher cancellations rates than expected. In fact, finding the 'optimal' value for appointment capacity, the value that jointly satisfies the requirement of both urgent and non-urgent patients without exceeding the permitted overtime hours, needs combined application of our model(s) and a daily-schedule evaluation model, as the one proposed in Koeleman and Koole (2012). Such a combined model might also help in joint optimization of patient panel size and appointment capacity.

Acknowledgments

The author thanks Dr. Dave Worthington for his comments on an early version of this paper. The author also thanks Prof. Sergei Savin for sharing his models and results. The author is very grateful to an associate editor and two referees for providing very useful comments that helped greatly improve the paper.

References

Abate, J., G. Choudhury, W. Whitt. 1999. An introduction to numerical transform inversion and its application to probability models. *Computational Probability*. Kluwer, Boston, 257–323.

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

Abate, J., W. Whitt. 1992b. Numerical inversion of probability generating functions. *Oper. Res. Lett.* **12**(4) 245–251.

Alfa, A. S. 1982. Time-inhomogenous bulk server queue in discrete time: A transportation type problem. *Oper. Res.* **30**(4) 650–658.

Altschuler, J., D. Margolius, T. Bodenheimer, K. Grumbach. 2012. Estimating a reasonable patient panel size for primary care physicians with team-based task delegation. *Ann. Fam. Med.* **10**(5) 396–400.

Bailey, N. T. J. 1954. On queueing processes with bulk service. *J. Roy. Statist. Soc. Ser. B-Statist. Methodology* **16**(1) 80–87.

Boudreau, P. E., J. S. Griffin, M. Kac. 1962. An elementary queueing problem. *Amer. Math. Monthly* **69**(8) 713–724.

Bruneel, H. 1993. Performance of discrete-time queueing systems. *Comp. Oper. Res.* **20**(3) 303 – 320.

Cayirli, T., K. K. Yang, S. A. Quek. 2012. A universal appointment rule in the presence of no-shows and walk-ins. *Production Oper. Management* **21**(4) 682–697.

Cayirli, T.a, E. Veral. 2003. Outpatient scheduling in healthcare: A review of literature. *Production Oper. Management* **12**(4) 519–549.

Chaudhry, M. L. 1991. *QROOR Software Package*. A&A Publications, 395 Carrie Crescent, Kingstone, Ontario, Canada.

Creemers, S., M. Lambrecht. 2010. Queueing models for appointment-driven systems. *Ann. Oper. Res.* **178**(1) 155–172.

Department of Health. 2012. Consultant-led referral to treatment waiting time guidance. `http://transparency.dh.gov.uk/2012/06/29/rtt-waiting-times-guidance/`. Accessed: 20/05/2013.

Downton, F. 1955. Waiting time in bulk service queues. *J. Roy. Statist. Soc. Ser. B-Statist. Methodology* **17**(2) 256–261.

Feldman, J., N. Liu, H. Topaloglu, S. Ziya. 2014. Appointment scheduling under patient preference and no-show behavior. *Oper. Res.* **62**(4) 794–811.

Gerchak, Y., D. Gupta, M. Henig. 1996. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Sci.* **42**(3) 321–334.

Green, L. V., S. Savin. 2008. Reducing delays for medical appointments: A queueing approach. *Oper. Res.* **56**(6) 1526–1538.

Gupta, D., B. Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE Trans.* **40**(9) 800–819.

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

HES Report. 2012. Hospital outpatient activity-2011-12. `http://www.hscic.gov.uk/hes`. Accessed: 20/05/2013.

Hofer, A. N., J. M. Araham, I. Moscovice. 2011. Expansion of coverage under the patient protection and affordable care act and primary care utilization. *Milbank Quarterly* **89**(1) 69–89.

Jaiswal, N. K. 1961. A bulk-service queueing problem with variable capacity. *J. Roy. Statist. Soc. Ser. B-Statist. Methodology* **23**(1) 143–148.

Janssen, A. J. E. M., J. S. H. van Leeuwaarden. 2005. Analytic computation schemes for the discrete-time bulk service queue. *Queueing Systems* **50**(2-3) 141–163.

Jiang, H., Z. Pang, S. Savin. 2012. Performance-based contracts for outpatient medical services. *Manufacturing Service Oper. Management* **14**(4) 654–669.

Kim, N., M. Chaudhry, B. Yoon, K. Kim. 2011. Inverting generating functions with increased numerical precision a computational experience. *J. Systems Sci. Systems Eng.* **20**(4) 475–494.

Koeleman, P. M., G. M. Koole. 2012. Optimal outpatient appointment scheduling with emergency arrivals and general service times. *IIE Trans. Healthc. Syst. Eng.* **2**(1) 14–30.

Kortbeek, N., M. E. Zonderland, A. Braaksma, I. M. H. Vliegen, R. J. Boucherie, N. Litvak, E. W. Hans. 2014. Designing cyclic appointment schedules for outpatient clinics with scheduled and unscheduled patient arrivals. *Performance Evaluation* **80**(0) 5–26.

Leeuwaarden, J. S. H. van. 2005. Queueing models for cable access networks. PhD thesis, Eindhoven University of Technology.

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

Liu, N., S. Ziya, V. G. Kulkarni. 2010. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing Service Oper. Management* **22**(2) 347–364.

Murray, M. 2002. Reducing waits and delays in the referral process. *Family Practice Management* **9**(3) 39–42.

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

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

Nakagawa, T., S. Osaki. 1975. The discrete weilbull distribution. *IEEE Trans. Rel.* **R-24**(5) 300–301.

Patrick, J., M. L. Puterman, M. Queyranne. 2008. Dynamic multipriority patient scheduling for a diagnostic resource. *Oper. Res.* **56**(6) 1507–1525.

Singh, V. P. 1971. Finite waiting space bulk service system. *J. Eng. Math.* **5**(4) 241–248.

Navid Izady is a lecturer (Assistant Professor) in Operational Research and Management Science jointly at the School of Mathematical Sciences and Southampton Business School at the University of Southampton. His research interests are stochastic modeling of service operations in particular healthcare services.