

Providing Timely Access to Medical Care: a Queueing Model

Linda V. Green, Sergei Savin

Graduate School of Business, Columbia University

Abstract

Many primary care offices and other medical practices regularly experience long backlogs for appointments. These backlogs are exacerbated by a significant level of last minute cancellations or “no-shows” which have the effect of wasting capacity. In this paper, we conceptualize such an appointment system as a single server queueing system in which customers who are about to enter service have a state-dependent probability of not being served and may rejoin the queue. We derive stationary distributions of the queue size assuming both deterministic as well as exponential service times and compare the performance metrics to the results of a simulation of the appointment system. Our results demonstrate the usefulness of the queueing models in providing guidance on identifying patient panel sizes for medical practices that are trying to implement a policy of “advanced access”.

1 Introduction and Literature Review

Difficulty in getting a timely appointment to see a physician is a very common problem. In a recent study, 33% of patients cited “inability to get an appointment soon” as a significant obstacle to care (Strunk and Cunningham 2002), and the Institute of Medicine 2001 report identified “timeliness” as 1 of the 6 key “aims for improvement” in its major report on quality of health care. For most patients, their primary care physician is their major access point into the health care system. Yet primary care practices often have long waits for appointments and may have difficulty in accommodating patients who have potentially urgent problems. As a result, patients experience delays in treatment and may be seen by someone other than their own physician, potentially leading to adverse clinical consequences, patient dissatisfaction, and loss of revenue for the practice. Large backlogs may require additional staff and resources to deal with patients trying to get appointments for the same day, and are often correlated with a high rate of cancellations or no-shows. No-show patients create a paradoxical situation where a physician is under-utilized while patients have long waits in getting appointments. There is a growing body of evidence indicating that no-shows are associated with substantial financial costs to primary care offices. Pesata *et al.* (1999) estimated a loss of over a million dollars resulting from 14,000 of missed appointments in a pediatric practice. Moore *et al.* (2001) looked at a family clinic where 31% of the appointments were missed or canceled and estimated that the corresponding loss was between 3% and 14% of annual revenues.

There have been a variety of approaches for reducing no-show rates: sending pre-appointment reminders; using financial penalties; and providing services to make it easier for patients to keep an appointment such as transportation vouchers and free or low-cost childcare (Tuso *et al.* 1999, Pesata *et al.* 1999). In recent years, a completely different approach to remedying the no-show problem is being adopted by a growing

number of primary care practices. This approach, known as advanced access, focuses on the inefficiencies built into many existing patient scheduling systems. As opposed to a “traditional” system where each physician’s daily schedule is fully booked in advance, or a “carve-out” model in which a fixed number of appointment slots are held open for urgent cases, the goal of the advanced access approach is to reduce delays by offering every patient a same-day appointment, regardless of the urgency of the problem. The fundamental idea behind advanced access is to “do all of today’s work today”, so that patients do not have to wait for appointments, practices do not waste capacity holding appointments in anticipation of same-day needs, and patients have a greater likelihood of seeing their own physician. Several success stories have documented the benefits of this approach in both managed care and fee-for-service environments including dramatically shorter waits, higher levels of continuity of care, less wasted capacity for the practice, and increased patient, staff, and physician satisfaction (Murray and Tantau 2000).

The ability to offer patients timely care, whether that means same day or within a couple of days, requires some minimum physician capacity relative to patient demand. Though advocates of advanced access stress that demand and supply must be “in balance”, they do not offer any means for physicians to identify an appropriate balance for their practice. In discussions with practitioners, we have found that one of the central questions in trying to implement an advanced access type of system is: What is a “manageable” panel size? Primary care practices and many specialty care practices, such as cardiology, have a “patient panel” – a set of patients who receive their care from the practice on some regular basis. So in these practices the primary lever to bring demand and supply into a relationship that is compatible with being able to offer short appointment dates is patient panel size.

In order to identify a panel size that will result in short waits for appointments with high probability, it is necessary to explicitly consider the impact of cancellations. Though

some patients cancel their appointments far enough in advance of their scheduled time to allow for a new appointment request to be substituted, many practices experience a high level of patients who cancel too late for this to happen or who simply do not show up at the scheduled time. In this paper, we will refer to both of these patient types as “no-shows”. Empirical evidence shows that the rate of no-shows increases with increasing appointment backlogs (Galucci *et al.* 2005), resulting in more unused appointment slots and wasted physician time. Therefore, instead of increased patient demand rate resulting in higher physician utilization, it may actually result in lower utilization levels. In addition, though some patients fail to appear at the appointed time because the original reason for the visit no longer exists, other no-shows are due to personal or work-related problems or the patient’s decision to seek treatment elsewhere rather than wait. In the latter situations, many no-shows schedule a new appointment with their original physician. This is even true when they have sought treatment elsewhere since it is common practice for clinics and emergency rooms to advise the patient to see their own physician as well.

In order to better understand the dynamics of such appointment systems, evaluate trade-offs, and provide guidance to identify panel sizes that are consistent with short patient backlogs and hence advanced access type of approaches, we develop two queueing models that explicitly capture the phenomenon of no-shows. Though a queueing model is an approximation of the actual dynamics, we show that the models can be used to provide reliable guidance on panel sizes that are consistent with advanced access. Despite the fact that about 2/3 of all primary care physicians work in group practices (National Ambulatory Medical Care Survey 2002), there are a number of studies that document the benefits of continuity of care (see e.g. Smoller 1992). These observations support the view that a patient should be seen, whenever possible, by his/her physician, and therefore, that a panel should be associated with an individual physician. Therefore, we conceptualize the appointment system as a single-server queue in which the probability of a no-show is

a non-decreasing function of the size of the backlog at the time of a patient's request for an appointment (arrival to the system). If a patient is a no-show, he/she remains in queue until the time at which service would have begun and the server begins an idle period for an interval of time equal to what would have been the service time. We assume that a given fraction of no-shows will reschedule after missing the appointment and so instead of leaving the system after the scheduled (but missed) service time, these patients join the end of the queue. Our model assumes that patients' requests for appointments follow a Poisson process. This is reasonable given the random nature of these requests and in fact the Poisson assumption has been used in other studies of outpatient appointment systems (see, e.g., Brahim and Worthington 1991). Because appointment slots are generally fixed and physicians strive to see a given number of patients per day, our primary model assumes deterministic service times. Under this assumption, we derive a recursion that can be used to efficiently compute performance measures such as the expected patient backlog and the probability of getting a same-day appointment. We demonstrate the reliability of these results by comparing them to those obtained from a simulation model of the patient appointment and backlog process which includes factors excluded from the queueing model. Because there is reason to believe that both of these models may underestimate total system variability and therefore result in overestimates of desirable panel sizes, we also consider the identical queueing model with i.i.d. exponential service times for which we can derive closed-form expressions. Our results indicate that for parameter values that are typical of many primary care practices, the two queueing models are very reliable in identifying a feasible range of panel sizes compatible with an advanced access approach.

Our work is related to several distinct literature streams.

The first is a set of papers on appointment scheduling practices, including those in health care settings. Mondschein and Weintraub (2003), Cayirli and Veral (2003) and Denton and Gupta (2003) provide comprehensive reviews of this literature. Most of the

papers dealing with health care appointments are focused on either the minimization of real-time, in-office patient delays or on the minimization of costs/maximization of profits for a primary care facility. An important subset of the appointment scheduling literature is comprised of papers focused on the optimal allocation of service capacity between different classes of patients, e.g. urgent vs. non-urgent patients (e.g., Gerchak *et al.* 1996, Gupta and Wang 2006). These papers provide a rather detailed modeling of the in-office waiting and service time dynamics. However, the analyzed optimal demand management decisions turn out to be sensitive to the values of model parameters which are hard to estimate in a practical setting. In particular, capacity allocation policies which reserve a fraction of service capacity to urgent cases (“carve out” policies) rely on the values of demand parameters for urgent and non-urgent cases alike, as well as on the estimates of the delay costs for the non-urgent patients and the overtime facility costs. For some time, practitioners (Murray and Tantau 2000, Murray and Berwick 2003) have pointed out this and other implementation challenges associated with “carve out” policies in support of the simpler advanced access approach.

There have also been many applications of queueing theory to healthcare systems (see, e.g. Preater 2001). The distinctive feature of our queueing model is the inclusion of no-shows. Only a few papers have dealt with this phenomenon. Mercer (1960, 1973) considered a queueing system in which customer arrivals are scheduled but may arrive late or not at all. Koole and Kaandorp (2006) developed a scheduling algorithm, which can incorporate no-shows, to optimize a weighted sum of the expected waiting times of customers, the idle time of the server, and the tardiness in the schedule. Most closely related to our work is a paper by Hassin and Mendel (2006) which considers a single-server queueing system with scheduled arrivals and Markovian service times and in which customers have a fixed probability of showing up. The paper focuses on identifying a schedule for a fixed number of customers which minimizes the sum of the expected customer waiting cost and

the expected server availability cost.

Our work draws on a literature stream of empirical papers estimating no-show rates in outpatient settings. The rates of missed appointments reported in these studies vary and appear to depend strongly on the type of health care service offered as well as on characteristics of the patient population. Hixon *et al.* (1999) found that more than a third of 468 surveyed US family practice residency programs reported an appointment no-show rate exceeding 20%. Moore *et al.* (2001) analyzed 4055 patient appointments scheduled at a family practice clinic over a period of 20 days and established that no-shows and canceled appointments occurred in around 31% of them. In addition, authors found that the clinic managed to recover only about 42% of the working schedule time left open by the no-shows and cancellations by using walk-ins and moving up the appointment times for other patients. Xakellis and Bennett (2001) report a similar figure of 25% of all appointments resulting in a no-show in a family practice clinic. Ulmer and Troxler (2006) studied seasonal variations of the no-show rates at a primary care practice and report the no-show rates ranging between 22% and 26.5%. A number of studies have focused on social and demographic factors behind patients' propensity for missing appointments: age, social and marital status, use of Medicaid, among others (Barron 1980; Tuso *et al.* 1999; Vikander *et al.* 1986; Specht and Bourget 1994; Weingarten *et al.* 1997). However, Galucci *et al.* (2005) is the only study we are aware of which looked at the effect of appointment delays on the resulting no-show rate. We will use the data reported in this study as well as the data we have obtained from the New York - Presbyterian MRI center to calibrate our model.

The rest of the paper is organized as follows. In section 2 we present results for an $M/D/1/K$ model in which a customer who is scheduled to begin service has a state-dependent probability of being a no-show resulting in an idle period for the server and, with a fixed probability, the customer rejoining the queue. We derive a recursion for

this model which can be used to obtain the stationary expected length of the patient backlog as well as the tail probabilities for the backlog. In section 3, we develop closed-form results for the same queueing model, but assuming i.i.d. exponential service times. Section 4 describes a simulation we developed to better capture the patient behavior in these type of appointment systems and section 5 contains the numerical results using all of these models with data from two healthcare facilities which experience backlog-dependent no-shows. Our findings and recommendations are summarized in Section 6.

2 An $M/D/1/K$ Queue with State-Dependent No-Shows

We model the single-physician practice with patient panel size N as a single server queueing system where patient demands for service form a Poisson process with rate λN . Though the customer pool is a finite source, we assume that N is sufficiently large that the arrival rate remains constant and is not dependent on the number of patients in service and in the appointment backlog. We further assume that there is a finite queue length, K , so that patients who arrive when the backlog is K are “lost”. We will use the terms “backlog” and “queue length” interchangeably to denote the total number of patients in the system, including the one being served. The “finite waiting room” assumption reflects the fact that when backlogs become excessive, patients will not make an appointment but will seek treatment elsewhere. The assumption of a finite queue size also makes the model analytically tractable, and we can choose a value for K for any specific setting that will not compromise the model’s usefulness. We assume that patients join the queue and are served in first-come, first-served order and that service times are deterministic with length T . The assumption of fixed service times is consistent with our goal of providing guidance

on patient panel sizes that result in short backlogs. Since physicians strive to see a fixed number of patients each day, idle time and delays during the day due to service time variability are inconsequential for this purpose.

In reality, patients often change their plans while being on the waiting list, generating reschedulings and no-shows. In the case of an appointment rescheduling, we assume that a patient moves from his/her current place on the waiting list to its end, and so doesn't affect the total length of the backlog. On the other hand, no-shows, or equivalently "last-minute" cancellations result in a service slot being unused. If a no-show patient doesn't reschedule then the backlog dynamics are unaffected. However as explained above, many no-shows do schedule a new appointment and hence generate an additional demand.

We are aware of only one empirical study on the functional form of the no-show rate. Galucci *et al.* (2005) report on the measurements of patient cancellation and no-show rates at a public mental health clinic at the Johns Hopkins Bayview Medical Center in Baltimore. The results of this study emphasize three features of the dependence of the rate of cancellations and no-shows on the backlog at the time when the patient receives an appointment:

a) there exists a non-negligible no-show rate even for same-day appointments (12% percent in this study);

b) the rate of no-shows monotonically increases (by about 12% per each extra day of waiting) with the backlog until it reaches a maximum (42%); and

c) the rate of no-shows stabilizes when it reaches this maximum value.

The general shape of the no-show function observed by Galucci *et al.* (2005) is consistent with data we have obtained for the MRI diagnostic facility at the New York-Presbyterian Medical Center. In particular, we found that the rate of cancellations and no-shows grew steadily with the backlog - from 4% for same-day appointments, to 11%,

17% and 29% for appointments scheduled 15, 30, and 45 days in advance, respectively - before finally stabilizing at 37% for appointments 60 days and longer. In our analysis below we use the following functional form for the no-show function which incorporates all three empirical features mentioned above in a simple and convenient way:

$$\gamma(k) = \gamma_{\max} - (\gamma_{\max} - \gamma_0) e^{-\frac{k}{\hat{C}}}. \quad (1)$$

Here k is the value of the appointment backlog at the time when a patient later exhibiting a no-show joins the backlog, $\gamma_0 \geq 0$ reflects the minimum observed no-show rate, $\gamma_{\max} \in (\gamma_0, 1]$ represents the maximum observed no-show rate, and C is a no-show backlog sensitivity parameter. Table 1 shows the best-fit values of the no-show function parameters \hat{C} , $\hat{\gamma}_0$, $\hat{\gamma}_{\max}$ for the data reported in Galucci *et al.* (2005) as well for the data we have obtained from the New York-Presbyterian Medical Center MRI facility.

Data Source	\hat{C} (days)	$\hat{\gamma}_0$	$\hat{\gamma}_{\max}$
Galucci <i>et al.</i> (2005)	9	0.15	0.51
MRI Facility at NYPMC	50	0.01	0.31

Table 1. Empirical estimates of no-show function parameters.

As estimates from Table 1 indicate, the actual parameters of the no-show function strongly depend on the type of medical environment: it is plausible that the patients of a mental health center are, in general, less reliable than the patients of an MRI diagnostic facility. Given the average no-show rates of 20% to 30% observed in primary care (Hixon *et al.* 1999, Moore *et al.* 2001, Xakelis and Bennett 2001, Ulmer and Troxler 2006), the MRI parameters appear to be more representative of a typical primary care environment than those observed in the mental care clinic.

As indicated above, the probability of a no-show is a function of the length of the backlog at the time at which the patient makes an appointment. To perfectly capture this empirical feature of the no-show rate would require a model with a system state

description which includes, in addition to the current backlog value, the backlogs observed at the arrival epochs of each of the patients currently waiting for service. For tractability, we approximate the no-show patient process as follows. No-shows are counted in the system state and are “served” during the unused service slot that they generate. At the end of a no-show service, a patient rejoins the backlog with a probability $r\gamma(k)$, where $\gamma(k)$ is a no-show rate which depends on k , the number of patients left behind by a departure, and r is the probability of a no-show rescheduling. Otherwise, the no-show leaves the system, and the backlog is reduced by one. These dynamics capture the critical feature of wasted service capacity observed in real medical practices that experience no-shows. The major advantage of this approximation is that it allows us to construct an analytical description of the transient behavior and an easy-to-compute recursion for the stationary state distribution of the resulting queue.

Below we extend the analysis of the $M/D/1/K$ system conducted by Garcia et al. (2002) to include the state-dependent patient no-show process. We first focus on the description of the transient evolution of the appointment backlog, and we will use the notation introduced in Garcia et al. (2002). In particular, for any time t we use $D(k, t, t + \Delta t)$ to denote the probability that a patient finishes his/her service in the time interval between t and $t + \Delta t$, leaving behind k patients in the appointment backlog, $0 \leq k \leq K - 1$. Similarly, we let $D(t, t + \Delta t) = \sum_{k=0}^{K-1} D(k, t, t + \Delta t)$ be the probability that there is an end-of-service for some patient in the interval $[t, t + \Delta t]$. Then, we define the corresponding departure rates as

$$d(k, t) = \lim_{\Delta t \rightarrow 0} \frac{D(k, t, t + \Delta t)}{\Delta t}, 0 \leq k \leq K - 1, \quad (2)$$

$$d(t) = \lim_{\Delta t \rightarrow 0} \frac{D(t, t + \Delta t)}{\Delta t}. \quad (3)$$

Departure rates (2) and (3) are instrumental in defining the dynamics of the appointment backlog. Let $p(k, t)$, $k = 0, \dots, K$ be the probability that the appointment backlog includes

k patients at time t and consider a set of time intervals $\Theta_n = \{t : (n-1)T \leq t < nT\}$, $n \in N$. For simplicity, consider the setting in which at time $t = 0$ the appointment system is empty¹. Then, during the time period $\Theta_1 = \{t : 0 \leq t < T\}$, there will be no patient departures, so that the appointment system will be evolving as a pure birth process. Using these arguments, Garcia et al. (2002) characterize the appointment backlog evolution for $t \in [0, T)$ for the initially-empty standard $M/D/1/K$ system, which coincides with the evolution for the same time period of the appointment system with potential no-shows which we analyze. For completeness, we report here the corresponding result:

Lemma 1 (Garcia et al. 2002)

Let $p(0,0) = 1$, $p(k,0) = 0$, $k = 1, \dots, K$, and $d(k,0) = 0$, $k = 0, \dots, K-1$. Then, at each time $0 \leq t < T$, all departure rates remain 0, and the probability distribution $p(k,t)$ obeys the following system of differential equations:

$$\begin{aligned} \frac{dp(0,t)}{dt} &= -\lambda N p(0,t), \\ \frac{dp(k,t)}{dt} &= -\lambda N p(k,t) + \lambda N p(k-1,t), \quad k = 1, \dots, K-1, \\ \frac{dp(K,t)}{dt} &= \lambda N p(K-1,t). \end{aligned} \tag{4}$$

Now, consider the time interval $\Theta_{\hat{n}} = \{t : (\hat{n}-1)T \leq t < \hat{n}T\}$ for $\hat{n} \geq 2$ and assume that the values of the backlog probabilities $p(k,t)$ and the departure rates $d(k,t)$ are known for any time $t \in \Theta_{\hat{n}-1} = \{t : (\hat{n}-2)T \leq t < (\hat{n}-1)T\}$. The following result describes the evolution of the appointment system in the time interval $\Theta_{\hat{n}}$.

Proposition 1

Let $\rho = \lambda NT$, and define

$$\alpha(k) = e^{-\rho} \frac{\rho^k}{k!}, \quad k \geq 0. \tag{5}$$

¹The case when at $t = 0$ an appointment system contains $B \geq 1$ patients can be considered using similar arguments (for details, see Garcia et al. 2002).

Then, for any time $t \in \Theta_{\hat{n}}$,

$$\begin{aligned}
d(0, t) &= p(0, t - T)\lambda N(1 - r\gamma(0))\alpha(0) + (1 - r\gamma(0))\alpha(0)d(1, t - T), \\
d(k, t) &= p(0, t - T)\lambda N((1 - r\gamma(k))\alpha(k) + r\gamma(k - 1)\alpha(k - 1)) \\
&\quad + (1 - r\gamma(k))\alpha(0)d(k + 1, t - T) \\
&\quad + \sum_{i=1}^k ((1 - r\gamma(k))\alpha(k + 1 - i) + r\gamma(k - 1)\alpha(k - i))d(i, t - T), \\
k &= 1, \dots, K - 2, \\
d(K - 1, t) &= p(0, t - T)\lambda N\left(\left(1 - \sum_{i=0}^{K-2} \alpha(i)\right)(1 - r\gamma(K - 1)) + r\gamma(K - 2)\alpha(K - 2)\right) \\
&\quad + \sum_{i=1}^{K-1} d(i, t - T) \\
&\quad \times \left(\left(1 - \sum_{j=0}^{K-1-i} \alpha(j)\right)(1 - r\gamma(K - 1)) + r\gamma(K - 2)\alpha(K - 1 - i)\right), \quad (6)
\end{aligned}$$

while the appointment backlog probabilities satisfy

$$\frac{dp(0, t)}{dt} = -\lambda Np(0, t) + d(0, t), \quad (7)$$

$$\frac{dp(k, t)}{dt} = -\lambda Np(k, t) - d(k - 1, t) + \lambda Np(k - 1, t) + d(k, t), \quad k = 1, \dots, K - 1, \quad (8)$$

$$\frac{dp(K, t)}{dt} = -d(K - 1, t) + \lambda Np(K - 1, t). \quad (9)$$

Proposition 1 outlines the numerical approach for recursively computing the time evolution of the appointment backlog for time periods Θ_n , $n \geq 2$, using the solution of (4). For each time period Θ_n , $n \geq 2$, equations (6) are used to compute the departure rates, using the corresponding departure rates and appointment backlog probability distribution for the previous time period $n - 1$. These computed departure rates are then used to solve the appointment backlog distribution equations (7)-(9). We note that while the form of the time evolution equations (7)-(9) for the system we consider is identical to the corresponding equations for an $M/D/K/1$ system, the actual expressions for the departure rates $d(k, t)$, as shown in (6) are very different in that they include the no-show function $\gamma(k)$. In particular, the results of Garcia et al. (2002) are a special case of (6)-(9) in which $\gamma \equiv 0$.

Our analysis allows us to define the following recursion for the stationary-state distribution of the appointment backlog:

Proposition 2

Define a recursion

$$\begin{aligned}
f(0) &= 1, \\
f(1) &= \frac{e^\rho}{1 - r\gamma(1)} - 1, \\
f(k+1) &= \frac{e^\rho}{(1 - r\gamma(k+1))} (f(k) - (1 - r\gamma(k+1))\alpha(k) - r\gamma(k)\alpha(k-1)) \\
&\quad - \frac{e^\rho}{(1 - r\gamma(k+1))} \left(\sum_{i=1}^k ((1 - r\gamma(k+1))\alpha(k+1-i) + r\gamma(k)\alpha(k-i)) f(i) \right), \\
k &= 1, \dots, K-2.
\end{aligned} \tag{10}$$

Then, the stationary backlog distribution, $\pi(k) = \lim_{t \rightarrow \infty} (p(k, t))$, $k = 0, \dots, K$, is given by

$$\begin{aligned}
\pi(0) &= \frac{1 - r\gamma(K)}{1 - r\gamma(K) + \rho \left(\sum_{i=0}^{K-1} f(i) \right) - r \sum_{i=1}^{K-1} (\gamma(K) - \gamma(i)) f(i)}, \\
\pi(k) &= \frac{(1 - r\gamma(K)) f(k)}{1 - r\gamma(K) + \rho \left(\sum_{i=0}^{K-1} f(i) \right) - r \sum_{i=1}^{K-1} (\gamma(K) - \gamma(i)) f(i)}, \quad k = 1, \dots, K-1, \\
\pi(K) &= 1 - \frac{(1 - r\gamma(K)) \left(\sum_{i=0}^{K-1} f(i) \right)}{1 - r\gamma(K) + \rho \left(\sum_{i=0}^{K-1} f(i) \right) - r \sum_{i=1}^{K-1} (\gamma(K) - \gamma(i)) f(i)}.
\end{aligned} \tag{11}$$

The results of Proposition 2 are generalizations of the stationary probability distribution for the standard $M/D/1/K$ queue derived in Garcia et al. (2002) for the case when patients exhibit state-dependent no-shows. The recursions for the stationary state probabilities are easily solved numerically. Figures 1 and 2 illustrate the expected appointment backlog and the fraction of patients who can be offered a same-day appointment computed using (11) as a function of patient panel size for a set of parameters based on our MRI data: $\lambda = 0.008$, $T = 1/20$ (in days), $\gamma_0 = 0.01$, $\gamma_{\max} = 0.31$, $C = 50$ (in days), $r = 1$, $K = 400$ appointment slots. For comparison, we also show the corresponding curves which

would result in the absence of no-shows. Several important observations can be made from these figures. First, the impact of cancellations is very significant, particularly when panel sizes exceed 2300 patients. Second, when cancellations exist, the expected patient backlog increases sharply for panel sizes greater than 2375 and quickly approaches the value of K . Similarly, at this same point, the probability of a patient getting a same-day appointment falls precipitously. This is of particular interest given the suggestion that an advanced access approach can be sustained in “typical” primary care practices which have panel sizes of 2500 patients (Murray and Tantau 2000). While such panel sizes seem to be manageable in the absence of no-shows, our results support observations made by several practitioners that a much smaller panel size may be required. For example, Figure 2 suggests that to ensure a 75% probability of same-day access, the panel size should be no more than 2360. A 75% level reflects the observation made by Murray and Tantau (2000) that about 25% of patients do not want a same-day appointment. As subsequently discussed, even smaller patient panel sizes are likely to be needed to support advanced access. Both figures highlight another important feature of appointment dynamics in the presence of no-shows: as the panel size increases, the transition to unmanageable backlogs occurs more rapidly, and over a narrower interval of panel size values than in the no-cancellation case. This phenomenon is due to the presence of a positive feedback effect that no-shows exert on the appointment backlog.

As mentioned previously, no-shows have been correlated with significant financial costs in medical practices. The major reason is wasted physician capacity. Since the percentage of no-shows grows with the patient backlog, physician utilization will eventually decrease as the backlog grows. This is illustrated in figure 3 which uses the MRI data to demonstrate how the actual utilization of a physician (expected fraction of time a physician spends serving “real” customers, *i.e.* not no-shows) depends on the value of the expected appointment backlog. We observe that for small expected backlog values, the utilization

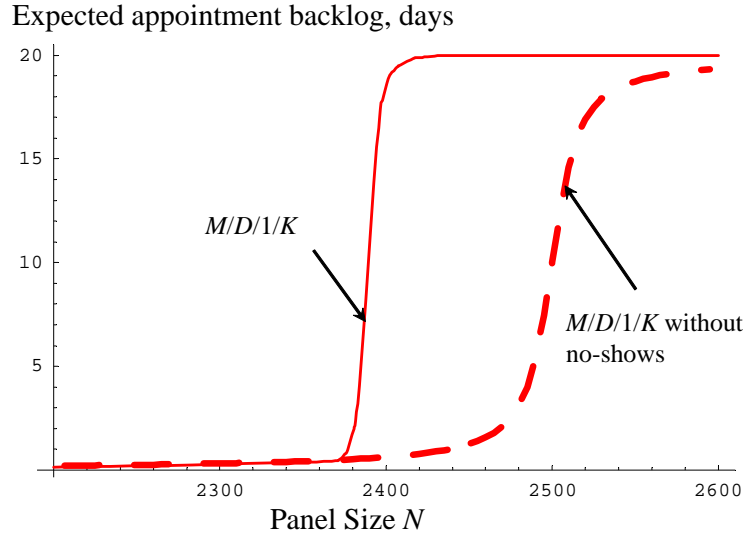


Figure 1: Expected appointment backlog as a function of the patient panel size for the $M/D/1/K$ model with and without no-shows ($\lambda = 0.008$, $T = 1/20$ (in days), $\gamma_0 = 0.01$, $\gamma_{\max} = 0.31$, $C = 50$ (in days), $r = 1$, $K = 400$ appointment slots).

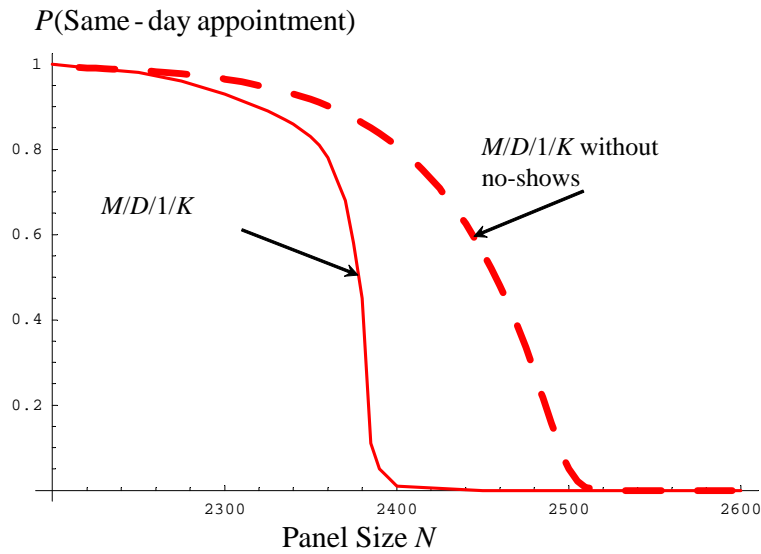


Figure 2: Probability of getting a same-day appointment as a function of the patient panel size for $M/D/1/K$ model with and without no-shows ($\lambda = 0.008$, $T = 1/20$ (in days), $\gamma_0 = 0.01$, $\gamma_{\max} = 0.31$, $C = 50$ (in days), $r = 1$, $K = 400$ appointment slots).

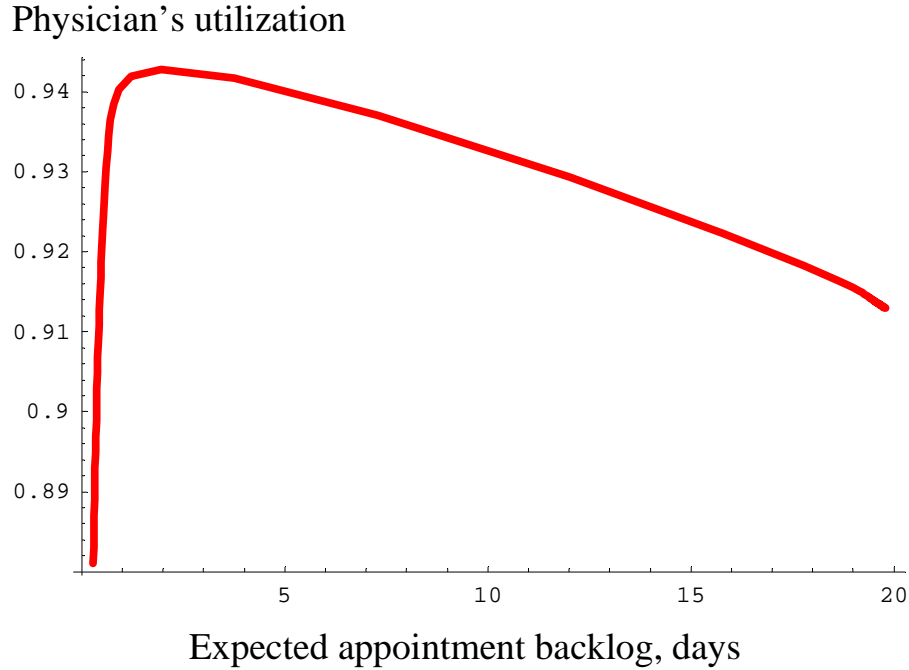


Figure 3: Physician's utilization as a function of the expected appointment backlog for the $M/D/1/K$ model with no-shows ($\lambda = 0.008$, $T = 1/20$ (in days), $\gamma_0 = 0.01$, $\gamma_{\max} = 0.31$, $C = 50$ (in days), $r = 1$, $K = 400$ appointment slots).

grows with the backlog. However, when the expected appointment backlog exceeds about 1.5 days, the utilization begins to drop - due to increasing levels of wasted physician time. This illustrates the observed paradoxical situation in which patients endure long waits for service while the physician is under-utilized. This also demonstrates that, as reported, an advanced access approach can increase physician utilization. Therefore, counter to common wisdom, by reducing its patient panel size and hence total demand for service, a medical practice can achieve higher revenue.

3 Alternate Model of Appointment Dynamics: $M/M/1/K$ Queue with State-Dependent No-Shows

The $M/D/1/K$ model introduced above may underestimate actual backlogs since it assumes that patients always take the next available appointment and are able to start service at the time of their request if the physician is available at that moment. In other words, the model assumes that there is no wasted physician capacity due to patient appointment preferences or patient travel times. Therefore, its estimates of maximum panel sizes to achieve a given backlog standard might be somewhat larger than can be actually achieved. In this section, we assume that service times are i.i.d. exponential random variables with expectation equal to T which allows us to obtain closed-form characterizations of the stationary state distribution for the appointment backlog. By introducing service variability into the model, we are hypothesizing that the resulting estimates for patient panel sizes will be useful in conjunction with those generated from the deterministic service model to establish ranges for reasonable patient panel sizes.

Let $p(k, t)$ be the probability that the appointment backlog B at time $t \geq 0$ is equal to k , and suppose, as before, that a patient completing service at time t was a no-show with probability $\gamma(k)$. As in the previous section, we assume that each no-show will be rescheduled with probability r . Then, the evolution of the probability distribution of the appointment backlog can be described by

$$\begin{aligned}
 p(k, t + dt) &= p(k, t) \left(1 - \lambda N dt - T^{-1} (1 - r\gamma(k-1)) dt \right) \\
 &\quad + p(k-1, t) \lambda N dt \\
 &\quad + T^{-1} p(k+1, t) (1 - r\gamma(k)) dt, \\
 k &= 1, \dots, K-1.
 \end{aligned} \tag{12}$$

(12) is equivalent to

$$\begin{aligned} \frac{dp(k, t)}{dt} &= -\left(\lambda N + T^{-1}(1 - r\gamma(k-1))\right)p(k, t) \\ &\quad + p(k-1, t)\lambda N + p(k+1, t)T^{-1}(1 - r\gamma(k)), \\ k &= 1, \dots, K-1, t \geq 0 \end{aligned} \quad (13)$$

Note that for $k = 0$ a similar analysis results in

$$\frac{dp(0, t)}{dt} = -\lambda N p(0, t) + p(1, t)T^{-1}(1 - r\gamma(0)), t \geq 0, \quad (14)$$

while for $k = K$ one gets

$$\frac{dp(K, t)}{dt} = -T^{-1}(1 - r\gamma(K-1))p(K, t) + p(K-1, t)\lambda N, t \geq 0, \quad (15)$$

For any time $t \geq 0$ the following normalization condition is required to hold:

$$\sum_{k=0}^K p(k, t) = 1. \quad (16)$$

The system (13)-(16) with appropriate initial conditions completely defines the evolution of appointment backlog in the presence of state-dependent patient no-shows. The expressions for the stationary probability distribution $\pi(k) = \lim_{t \rightarrow \infty} (p(k, t))$ corresponding to (13)-(16) satisfy

$$\begin{aligned} \left(\lambda N + T^{-1}(1 - r\gamma(k-1))\right)\pi(k) &= \pi(k-1)\lambda N + \pi(k+1)T^{-1}(1 - r\gamma(k)), \\ k &= 1, \dots, K-1, \\ \lambda N\pi(0) &= \pi(1)T^{-1}(1 - r\gamma(0)), \\ T^{-1}(1 - r\gamma(K-1))\pi(K) &= \pi(K-1)\lambda N, \end{aligned} \quad (17)$$

where

$$\sum_{k=0}^K \pi(k) = 1. \quad (18)$$

Closed-form expressions for the stationary probabilities satisfying (17)-(18) are described

in the following proposition.

Proposition 3

Let $\rho = \lambda NT$ and define

$$\bar{\gamma}(k) = \frac{1 - \left(\prod_{i=0}^{k-1} (1 - r\gamma(i)) \right)^{\frac{1}{k}}}{r}. \quad (19)$$

Then, the solution to (17)-(18) is given by

$$\pi(k) = \pi(0) \left(\frac{\rho}{1 - r\bar{\gamma}(k)} \right)^k, \quad (20)$$

where

$$\pi(0) = \left(1 + \sum_{l=0}^K \left(\frac{\rho}{1 - r\bar{\gamma}(l)} \right)^l \right)^{-1}. \quad (21)$$

Proposition 3 establishes the general form of the stationary distribution corresponding to the backlog dynamics (13)-(16). The closed-form nature of these stationary distribution expressions, in contrast with those in Proposition 2, allow for analytical study of any performance measure of interest. In particular, the expression for the expected backlog is given by

$$E[B] = \sum_{k=1}^K k\pi(k) = \left(1 + \sum_{l=0}^K \left(\frac{\rho}{1 - r\bar{\gamma}(l)} \right)^l \right)^{-1} \sum_{k=1}^K k \left(\frac{\rho}{1 - r\bar{\gamma}(k)} \right)^k, \quad (22)$$

while the expression for the probability that the stationary backlog does not exceed m is

$$P(B \leq m) = \sum_{k=0}^m \pi(k) = \left(1 + \sum_{l=0}^K \left(\frac{\rho}{1 - r\bar{\gamma}(l)} \right)^l \right)^{-1} \sum_{k=0}^m \left(\frac{\rho}{1 - r\bar{\gamma}(k)} \right)^k. \quad (23)$$

Figures 4 and 5 compare the curves for expected stationary backlog and the probability of a same-day appointment computed using (22) and (23) with the corresponding results for the $M/D/1/K$ model. Not surprisingly, the $M/M/1/K$ model predicts higher expected backlogs and lower probabilities of same-day appointments. Note that for the set of problem parameters considered here the performance measure curves produced by the

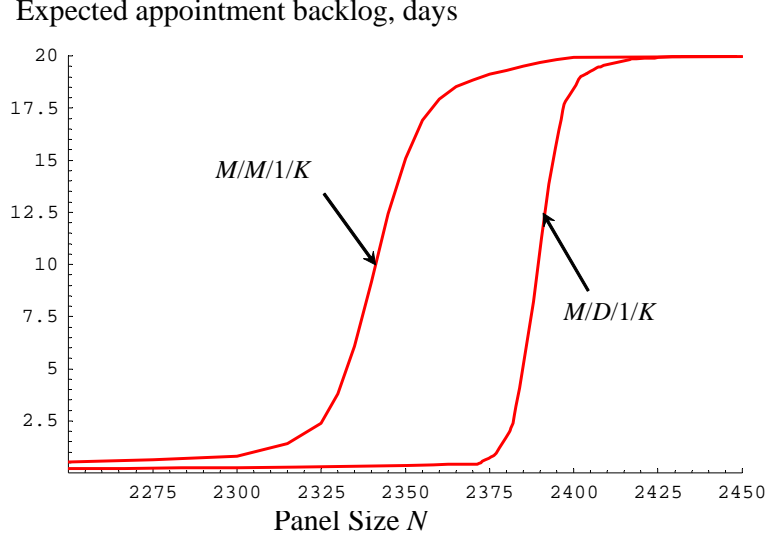


Figure 4: Expected appointment backlog as a function of the patient panel size for $M/M/1/K$ and $M/D/1/K$ models ($\lambda = 0.008$, $T = 1/20$ (in days), $\gamma_0 = 0.01$, $\gamma_{\max} = 0.31$, $C = 50$ (in days), $r = 1$, $K = 400$).

two queuing models, while qualitatively similar, do differ from each other in quantitative details. For example, both models exhibit a rapid growth of the expected backlog - but for the $M/M/1/K$ system it starts at a panel size of about 2325, while for the $M/D/1/K$ this growth becomes pronounced only for the panel sizes of about 2375. Similarly, the estimated panel size for which 75% of patients can be offered same-day appointments is 2363 using the $M/D/1/K$ dynamics and 2300 using the $M/M/1/K$ model.

In the next section, we describe a discrete-time simulation which captures additional real-life features of an appointment system.

4 A Simulation of the Patient Appointment System

Neither of the two queuing models developed in this paper capture all of the features of an actual patient appointment system which may impact the determination of patient

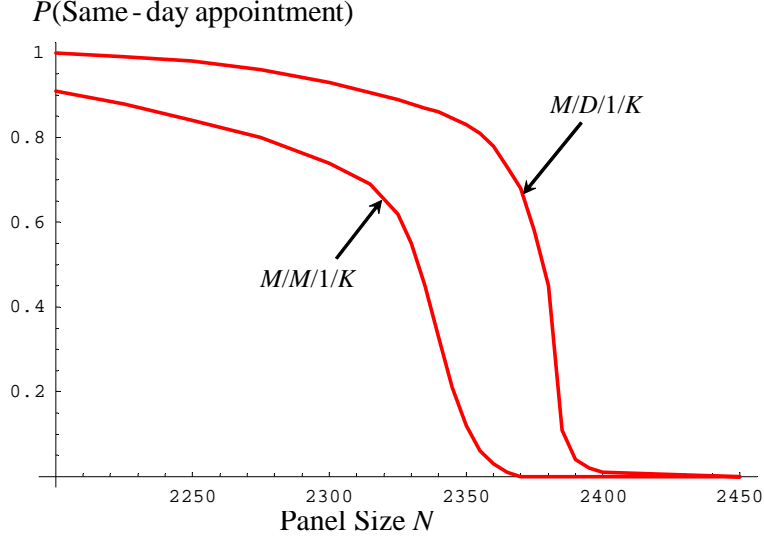


Figure 5: Probability of getting a same-day appointment as a function of the patient panel size for $M/M/1/K$ and $M/D/1/K$ models ($\lambda = 0.008$, $T = 1/20$ (in days), $\gamma_0 = 0.01$, $\gamma_{\max} = 0.31$, $C = 50$ (in days), $r = 1$, $K = 400$).

panel sizes that are consistent with a particular performance goal. To determine their reliability, we developed a simulation model that better conforms to the reality of patient appointment process dynamics.

The deterministic appointment duration T offers a convenient interval for measuring time. Specifically, in our model, the units of time are multiples of T . Let the “current” time period be designated by an integer index $n = 1, \dots, \mathcal{N}$, where \mathcal{N} is a large positive integer. At time n , we describe the state of the appointment system as a vector $\mathbf{s}^n = (a_{1n}, \tau_{1n}, a_{2n}, \tau_{2n}, \dots, a_{Kn}, \tau_{Kn})$, where a_{in} , $i = 1, \dots, K$ is a binary number indicating whether the i -th slot in the appointment schedule is occupied ($a_{in} = 1$) or vacant ($a_{in} = 0$), and $\tau_{in} < n$, $i = 1, \dots, K$ indicate, for slots occupied at time n ($a_{in} = 1$), at what time epoch the appointment for this slot was accepted (if $a_{in} = 0$, the value of τ_{in} is set to \mathcal{N} by default). In our state description, as in the queueing models, K is the maximum possible number of patients in the system. While the choice of K is somewhat arbitrary,

in our simulation study we use a value of $K = 400$ corresponding to a 20-day interval (*i.e.* 4 weeks) under the assumption of a typical appointment capacity of 20 patients per day (see, e.g. Murray and Berwick 2003).

We model the appointment acceptance dynamics as follows: when, in period n , a patient requests an appointment, the system offers him/her the list of vacant appointment slots. For expositional purposes it is convenient to define a list of empty slots at time n , $E_n = \{i = 1, \dots, K | a_{in} = 0\}$ with cardinality e_n . Let $E_n(j)$ to denote the j -th element of E_n (for example, in the set $E_n = \{3, 10\}$ of cardinality $e_n = 2$, $E_n(1) = 3$, and $E_n(2) = 10$). We assume that if E_n is empty, *i.e.*, that at time n there are no available appointments within a reasonably long appointment horizon, a patient requesting an appointment returns to the panel without receiving care from the provider. This would correspond to cases where a patient might use another provider, an emergency room or an urgent care center to obtain care. We assume that a patient accepts the first available slot $E_n(1)$ with probability p_f . Alternately, with probability $1 - p_f$, a patient picks, with equal likelihood, any slot from set E_n . In particular, if $e_n = 1$, the patient always takes the slot $E_n(1)$. This modeling approach to the appointment acceptance dynamics reflects the observation made by Murray and Tantau (2000) that in a typical practice there will always be a fraction of patients who opt to see a physician at a later date even if a same-day appointment is available. In the absence of empirical data on patient preferences regarding the date for non-same-day appointments, we use a simple uniform preference assumption. We note that the presence of preferences for later-date appointments increases the average time interval between the date when the appointment is accepted and the date when the actual care is administered and will therefore affect the average no-show rate.

Under the set of assumptions introduced above, the evolution dynamics of an appointment system can be described as follows. We consider the sequence of time epochs at which a service may potentially begin. Suppose that the appointment system is in state

$\mathbf{s}^n = (a_{1n}, \tau_{1n}, a_{2n}, \tau_{2n}, \dots, a_{Kn}, \tau_{Kn})$ at time n and let time n designate a moment of time when a patient in slot 1, if it is occupied, is about to be served. Between times n and $n+1$ (time interval of duration T) the number of new customers requesting an appointment is equal to k with probability $\alpha(k)$ defined by (5). In addition, if $a_{1n} = 1$, a patient about to be served is a no-show and requests a new appointment with probability $r\gamma(n - \tau_{1n})$. Thus, the total number of people requesting an appointment during the n -th time period (i.e., between times n and $n+1$) equals k with probability

$$\hat{\alpha}_{kn} = \begin{cases} (1 - r\gamma(n - \tau_{1n}) a_{1n}) e^{-\rho}, & k = 0, \\ (1 - r\gamma(n - \tau_{1n}) a_{1n}) \alpha(k) + r\gamma(n - \tau_{1n}) a_{1n} \alpha(k - 1), & k \geq 1. \end{cases} \quad (24)$$

Consider the case of $p_f = 1$, where patients always select the earliest available appointment (the generalization of the description of the state dynamics for the case of $p_f < 1$ is straightforward albeit more cumbersome). At time $n+1$, the state of the system is $\mathbf{s}^{n+1} = (a_{1,n+1}, \tau_{1,n+1}, a_{2,n+1}, \tau_{2,n+1}, \dots, a_{K,n+1}, \tau_{K,n+1})$ where, with probability $\hat{\alpha}_{0n}$,

$$a_{i,n+1} = \begin{cases} a_{i+1,n}, & i = 1, \dots, K-1, \\ 0, & i = K, \end{cases} \quad (25)$$

$$\tau_{i,n+1} = \begin{cases} \tau_{i+1,n}, & i = 1, \dots, K-1, \\ \mathcal{N}, & i = K, \end{cases} \quad (26)$$

and, with probabilities $\hat{\alpha}_{kn}$, $k = 1, \dots, e_n - 1$,

$$a_{i,n+1} = \begin{cases} 1, & \text{if } i+1 \notin E_n, \\ 1, & \text{if } i+1 \in E_n, i = E_n(1), \dots, E_n(k), \\ 0, & \text{if } i+1 \in E_n, i = E_n(k+1), \dots, E_n(e_n), \end{cases} \quad (27)$$

$$\tau_{i,n+1} = \begin{cases} \tau_{i+1,n}, & \text{if } i+1 \notin E_n, \\ n, & \text{if } i+1 \in E_n, i = E_n(1), \dots, E_n(k), \\ \mathcal{N}, & \text{if } i+1 \in E_n, i = E_n(k+1), \dots, E_n(e_n), \end{cases} \quad (28)$$

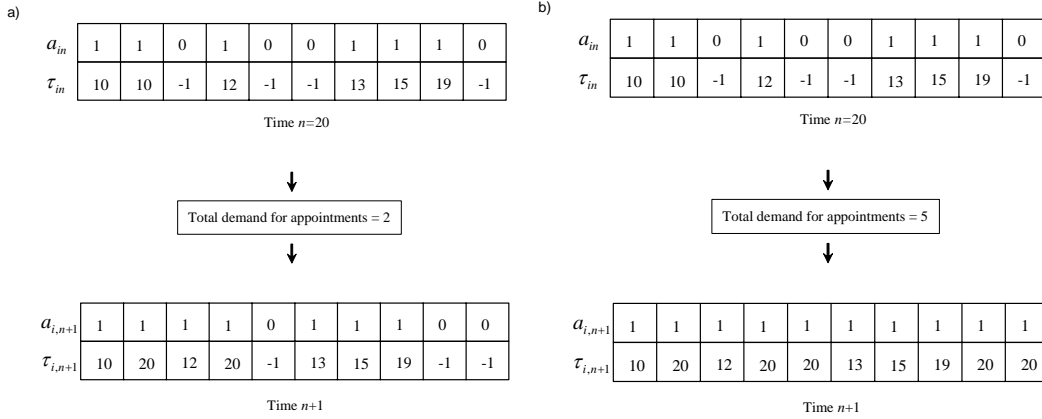


Figure 6: Example of appointment log change between the times $n = 20$ and $n + 1$ when the total demand for appointments is 2 (a) and 5 (b).

and, finally, with probability $\sum_{k=e_n}^{+\infty} \hat{\alpha}_{kn}$,

$$a_{i,n+1} = 1, \quad (29)$$

$$\tau_{i,n+1} = \begin{cases} \tau_{i+1,n}, & \text{if } i+1 \notin E_n, \\ n, & \text{if } i+1 \in E_n. \end{cases} \quad (30)$$

Note that the discrete-time transition dynamics of our simulation model differs from the dynamics of the queueing models introduced in the previous sections in following way: when the appointment system is empty, the next patient requesting service will be immediately admitted into service by the queueing systems, but not by the simulation system. In fact, the simulation system will always start serving the next available patient at discrete time instances n , and not at the time of patient arrival.

Figure 6 illustrates the appointment log transition between times $n = 20$ and $n + 1$ described by (25)-(30) for the example with $K = 10$, $E_n = \{3, 5, 6, 10\}$, $e_n = 4$, for the cases when the total demand for appointments between times $n = 20$ and $n + 1$ is 2 (Figure 6a) and 5 (Figure 6b).

In the next section we present the results of numerical experiments which compare the

performance metrics of the appointment system obtained through simulation with those derived from the $M/D/1/K$ and $M/M/1/K$ queueing approximations.

5 Numerical Results and Observations

We evaluated the expected backlog and the probability of a same-day appointment using the two cancellation functions described in section 3 for the two queueing models and the simulation. We considered two patient appointment behaviors in the simulation model. The first, which we refer to as the “no-preference” model, assumes that all patients take the first available appointment. The second, which we call the “preference” model, assumes that while 75% of patients take the first available appointment, the other 25% request an appointment time which is uniformly distributed across the available slots during the week following the first available appointment. Though we do not have any specific data on patient appointment preferences, this assumption reflects the observation that 25% of patients who are offered same-day appointments opt to see a physician at a later time (Murray and Tantau 2000).

Our results for the expected appointment backlog are illustrated in Figures 7 and 8 for the data from the Columbia MRI facility and the Johns Hopkins Mental Care facility, respectively. The expected backlog curves prompt two observations.

First, the $M/D/1/K$ model offers a remarkably good approximation to the stationary simulation dynamics of the no-preference model. This is not very surprising since there are only two major differences in the assumptions of the no-preference simulation and the $M/D/1/K$ model. First, the simulation calculates the no-show probability for a patient using the state of the appointment system at the time he/she *joins* the appointment queue, while the $M/D/1/K$ model uses the state of the appointment system at the time he/she *leaves* the system. The other difference is in the service dynamics when the appointment

system is empty. At these times, the arrival of a patient immediately triggers the start of the service in the $M/D/1/K$ system, while in the discrete-time simulation such a patient will start its service only at the beginning of the next time interval. Thus, we would anticipate that the expected backlog values produced using the $M/D/1/K$ dynamics would be somewhat lower than those produced with the no-preference simulation, particularly, for smaller patient panel sizes. This is confirmed in Figure 7 which also shows that over a wide range of panel size values the expected backlog calculated using the recursion relations (10) is nearly identical to the values provided by a no-preference simulation.

As is to be expected, for the preference simulation the no-show dynamics of (24) results in longer expected backlogs than those estimated by the no-preference simulation. Therefore, the $M/D/1/K$ system does not do as good a job when we include this assumption. In the MRI setting characterized by moderate and slowly rising no-show rates, the $M/M/1/K$ system provides a better approximation for panel sizes resulting in expected backlogs of one day or less. However, for larger panel sizes, the two queueing models provide lower and upper bounds on the “actual” expected backlog. In the mental care setting, with its higher and quickly rising no-show rates, the $M/M/1/K$ model is not reliable. While the $M/D/1/K$ underestimates the expected backlog for the preference model, it still does a good job in identifying the panel size at which the expected backlog starts growing rapidly.

To get a more practical perspective on the quality of the queueing approximations for use with an advanced access approach, we examine the models’ estimations of panel sizes that assure a specified probability for a patient to get a same-day appointment. These are presented in Tables 2 and 3 for the two sets of data. The data in Table 2, corresponding to the MRI setting, confirm our earlier observation regarding the excellent degree of approximation that the $M/D/1/K$ model provides under the assumption that patients take the first available appointment. When patients exhibit preferences for later

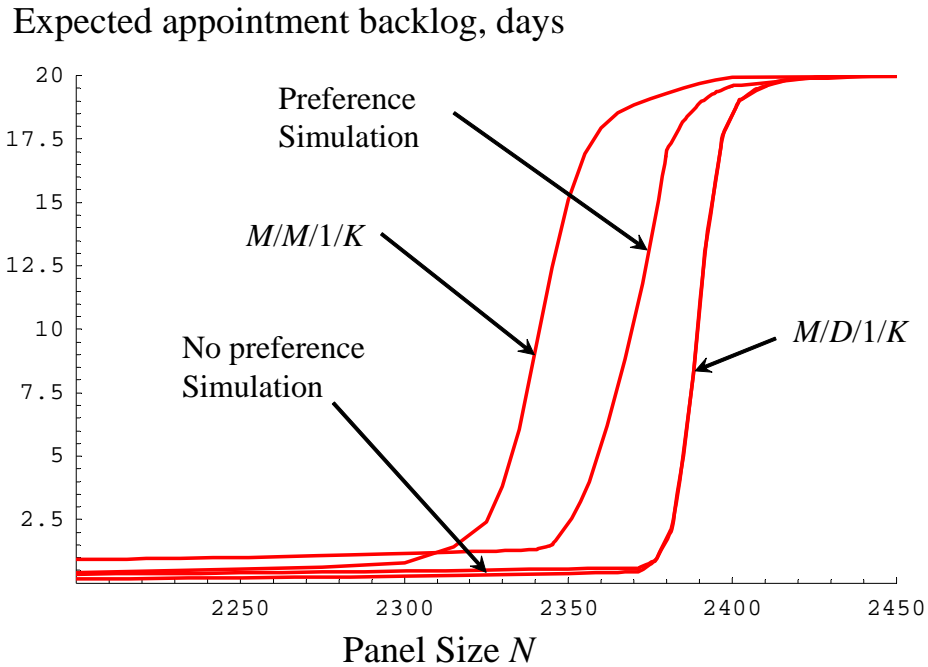


Figure 7: *Expected stationary backlog as a function of the panel size for 4 stochastic models based on the data from the Columbia-Presbyterian MRI facility ($\lambda = 0.008$ per day, $T = 1/20$ days, $\gamma_0 = 0.01$, $C = 50$ days, $\gamma_{\max} = 0.31$). K is set at 400 appointment slots which is equivalent to 20 days. For each panel size, the MSE of the simulation results are less than or equal to 0.1.*

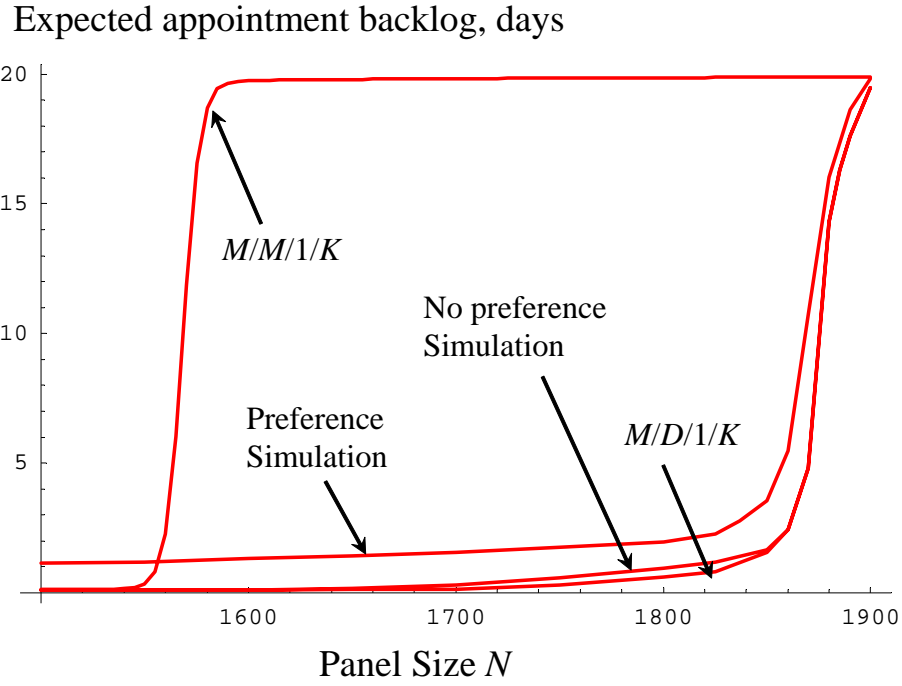


Figure 8: *Expected stationary backlog as a function of the panel size for 4 stochastic models based on the data form the Johns-Hopkins Mental Care Center ($\lambda = 0.008$ per day, $T = 1/20$ days, $\gamma_0 = 0.15$, $C = 9$ days, $\gamma_{\max} = 0.51$). K is set at 400 appointment slots which is equivalent to 20 days. For each panel size, the MSE of the simulation results are less than or equal to 0.1.*

appointments, the $M/D/1/K$ values, while overestimates, are still quite close with relative errors of less than 2.5%. The $M/M/1/K$ model underestimates the "true" panel size as estimated by the preference simulation and so the two models can be used to bound the panel size needed to achieve a given level of same-day availability.

As indicated by the expected backlog graphs, the $M/M/1/K$ model is not reliable for the type of cancellation behavior that appears in the mental care setting and therefore we don't include its estimates in Table 3. Again, we see that the $M/D/1/K$ provides extremely good estimates for the no-preference case. Though its estimates are not as good for the case of patient preference, the relative errors are fairly low for the levels of same-day availability that are most likely to be used in practice. In particular, for the case in which the goal is to offer 75% of patients a same-day appointment, the $M/D/1/K$ model's estimate is less than 2% above the one resulting from the simulation.

P(same-day appt.)	$M/D/1/K$	No-Pref. Simulation	Pref. Simulation	$M/M/1/K$
0.90	2320	2315	2275	2205
0.85	2345	2340	2305	2245
0.80	2357	2355	2330	2275
0.75	2363	2363	2345	2295
0.70	2368	2368	2355	2307

Table 2. Estimated panel size values corresponding to different probabilities of getting a same-day appointment for 4 stochastic models for the data from the Columbia-Presbyterian MRI facility ($\lambda = 0.008$ per day, $T = 1/20$ days, $\gamma_0 = 0.01$, $C = 50$ days, $\gamma_{\max} = 0.31$, $r = 1$, $K = 400$ appointment slots).

P(same-day appt.)	$M/D/1/K$	No-Pref. Simulation	Pref. Simulation
0.90	1785	1775	1650
0.85	1800	1788	1720
0.80	1809	1800	1760
0.75	1812	1808	1777
0.70	1817	1812	1800

Table 3. Estimated panel size values corresponding to different probabilities of getting a same-day appointment for 3 stochastic models for the data from the Johns Hopkins Mental Care facility ($\lambda = 0.008$ per day, $T = 1/20$ days, $\gamma_0 = 0.15$, $C = 9$ days, $\gamma_{\max} = 0.51$, $K = 400$ appointment slots). Note that the corresponding values for the $M/M/1/K$ model are too different from the values of any other model and are not shown here.

6 Discussion

Healthcare practices are increasingly competing not only on cost, but also in quality and patient satisfaction. In this environment, timely access to care has become a more important issue. As a result, physician practices are eager to embrace new approaches to patient appointment scheduling in order to reduce backlogs, increase productivity, and improve patient satisfaction.

Currently, these practices have no guidelines or framework to help identify an appropriate balance between physician capacity and patient panel sizes in order to achieve manageable patient backlogs. In this paper, we presented two queueing models which we believe will be very helpful in this regard. In particular, these are the first models to explicitly incorporate a backlog-dependent cancellation rate which has been widely observed and reported upon in the professional medical literature. As we have demonstrated, this cancellation factor has a significant impact on system performance and on the max-

imum patient panel size that can be reasonably handled by a practice. While no model is a perfect representation of reality, we believe that these are useful for guiding patient panel decisions since they capture the essential dynamics of a patient appointment system. Specifically, our results indicate that the deterministic service model developed here is extremely reliable when patients take the next available appointment. Though we know that this does not always happen in reality, some practices that are trying to implement advanced restrict the number of days in advance that they will book an appointment and encourage their patients to take one of the first few available appointments. More generally, our results from the MRI facility indicate that using both the deterministic and exponential models can be very useful in bounding the maximum panel size that will be compatible with an advanced access approach in primary care settings where overall no-shows do not exceed 30% of all appointments - the level corresponding to the no-show values observed in practice (Hixon *et al.* 1999, Moore et al. 2001, Xakellis and Bennett 2001, Ulmer and Troxler 2006).

And even in what seems to be the extreme case of the mental health facility with its high and rapidly increasing level of cancellations, the $M/D/1/K$ model is useful in providing “ballpark” estimates of appropriate panel sizes which, if reduced by about 5%, will be very reliable.

Though our focus has been on primary care practices which have been struggling to implement advanced access systems, our model can be used to help determine capacity requirements and/or patient panel sizes for any outpatient facility which accepts appointments. In particular, it could be useful in specialty practices such as pediatrics, gynecology, and cardiology.

References

- American Academy of Family Physicians. 2006. Facts about Family Medicine, <http://www.aafp.org/online/en/home/aboutus/specialty/facts/14.html> (checked on May 24, 2007).
- Bailey, N.T.J. 1954. A continuous time treatment of a simple queue using generating functions, *Journal of Royal Statistical Society*, **B 16**, 288-291.
- Barron, W. M. 1980. Failed appointments. Who misses them, why they are missed, and what can be done, *Primary Care*, **7(4)**, 563-574.
- Brahimi, M., and D.J. Worthington. 1991. Queueing models for out-patient appointment systems – A case study, *Journal of the Operational Research Society*, **42(9)**, 773-746.
- Cayirli, T., and E. Veral. 2003. Outpatient scheduling in health care: a review of literature, *Production and Operations Management*, **12(4)**, 519-549.
- Clark, A.B. 1953. The time-dependent waiting line problem. University of Michigan Report M720-1R39.
- Denton, B., and D. Gupta. 2003. Sequential bounding approach for optimal appointment scheduling, *IIE Transactions*, **35**, 1003–1016.
- Galucci, G., W. Swartz, and 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**, 344-346.
- Garcia, J.-M., O. Brun, and D. Gauchard. 2002. Transient Analytical Solution of M/D/1/N Queues, *Journal of Applied Probability*, **39 (4)**, 853-864.
- Gerchak, Y., D. Gupta, and M. Henig. 1996. Reservation planning for elective surgery under uncertain demand for emergency surgery, *Management Science*, **42(3)**, 321–334.
- Hassin, R., and S. Mendel. 2006. Scheduling arrivals to queues; a model with no-shows,

Working Paper, Department of Statistics and Operations Research, Tel-Aviv University.

Hixon, A.L., R.W. Chapman, and J. Nuovo. 1999. Failure to keep clinic appointments: Implications for residency education and productivity, *Family Medicine*, **31(9)**, 627-630.

Institute of Medicine. 2001. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, D.C.: National Academy Press.

Koole, G., and G. Kaandorp. 2006. Optimal outpatient appointment scheduling, Working Paper, Department of Mathematics, Vrije Universiteit Amsterdam, The Netherlands, <http://www.math.vu.nl/~koole/articles/hcms07b/> (checked on May 24, 2007).

Mercer, A. 1960. A Queuing Problem in which Arrival Times of the Customers are Scheduled, *Journal of the Royal Statistical Society, Series B*, **22**, 108-113.

Mercer, A. 1973. Queues with Scheduled Arrivals: A Correction, Simplification and Extension, *Journal of the Royal Statistical Society, Series B*, **35 (1)**, 104-116.

Mondschein, S., and G.Y. Weintraub. 2003. Appointment Policies in Service Operations: A Critical Analysis of the Economic Framework, *Production and Operations Management*, **12 (2)**, 266-286.

Moore, C.G., P. Wilson-Witherspoon, and J.C. Probst. 2001. Time and money: Effects of no-shows at a family practice residency clinic, *Family Medicine*, **33(7)**, 522-527.

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

Murray, M., and D.M. Berwick. 2003. Advance access: reducing waiting and delays in primary care, *JAMA*, **289 (8)**, 1035-1039.

Murray, M. 2005. Answers to Your Questions About Same-Day Scheduling, *Family Practice Management*, **12 (3)**, 59-64.

Hing E., D.K. Cherry, and D.A. Woodwell. 2006. National Ambulatory Medical Care Survey: 2004 Summary. Advance Data from Vital and Health Statistics; No. 374.

Hyattsville, Maryland: National Center for Health Statistics.

Pesata, V., G. Palliga, and A.A. Webb. 1999. A descriptive study of missed appointments: Families' perceptions of barriers to care, *Journal of Pediatric Health Care*, **13**, 178-182.

Preater, J. 2001. A bibliography of queues in health and medicine. *Keele Mathematics Research Report*, **01-1**.

Rust, C. T., N.H. Gallups, W.S. Clark, D.S. Jones, and W.D. Wilcox. 1995. Patient appointment failures in pediatric resident continuity clinics, *Archives of Pediatric and Adolescent Medicine*, **149**, 693-695.

Sharma, O.P. 1997. Markovian Queues, Allied, New Delhi.

Smoller M. 1992. Telephone Calls and Appointment Requests: Predictability in an Unpredictable World, *HMO Practice*, **6(2)**, 25-29.

Specht, E. M. and C.C. Bourget. 1994. Predictors of nonattendance at the first newborn health supervision visit. *Clinical Pediatrics*, **33**, 273-279.

Strunk, B.C. and P.J. Cunningham. 2002. Treading Water: Americans' Access to Needed Medical Care, 1997-2001. Center for Studying Health System Change.

Tuso, P. J., K. Murtishaw, and W. Tadros. 1999. The easy access program: A way to reduce patient no-show rate, decrease add-ons to primary care schedules, and improve patient satisfaction, *The Permanente Journal*, **3(3)**.

Ulmer, T., and C. Troxler. 2006. The Economic Cost of Missed Appointments and the Open Access System, working paper, University of Florida Area Health Education Centers Program, <http://www.flahec.org/ahec/chs/2002/Noshows.html> (checked on May 24, 2007).

Vikander, T., K. Parnicky, R. Demers, K. Frisof, P. Demers, and N. Chase. 1986. New patient no-shows in an urban family practice center: Analysis and intervention, *The*

Journal of Family Practice, **22(3)**, 263-268.

Weingartner, N., D.L. Meer, and J.A. Schneid. 1997. Failed appointments in residency practices: Who misses them and what providers are most affected? *Journal of the American Board of Family Practice*, **10**, 407-411.

Xakellis, G. C., and A. Bennett. 2001. Improving clinic efficiency of a family medicine teaching clinic, *Family Medicine*, **33(7)**, 533-538.

Appendix for “Providing Timely Access to Medical Care: a Queueing Model”

Proof of Proposition 1

Below we extend the proof approach outlined in Garcia et al. (2002) to include the state-dependent no-show process. Consider a probability of a service completion in the interval $(t, t + \Delta)$ which leaves $k = 1, \dots, K - 2$ patients in the backlog. Four mutually exclusive scenarios can lead to such an event.

First, it is possible that the appointment system was empty at time $t - T$ and 1) a patient requested service between $t - T$ and $t - T + \Delta t$, 2) the patient either has actually arrived for his/her service or exhibited a no-show, but did not reschedule his/her service for a later time, and 3) during the time interval allocated for this patient’s service there were k new appointment requests. The probability of this scenario is $p(0, t - T)\lambda N\Delta t(1 - r\gamma(k))\alpha(k)$.

Second, it is possible that the appointment system was empty at time $t - T$ and 1) a patient requested service some time between $t - T$ and $t - T + \Delta t$, 2) the patient exhibited a no-show *and* rescheduled his/her service for a later time, and 3) during the time interval allocated for this patient’s service there were $k - 1$ new appointment requests. The probability of this scenario is $p(0, t - T)\lambda N\Delta t r\gamma(k - 1)\alpha(k - 1)$.

Third, it is possible that 1) a departure that left $i = 1, \dots, k + 1$ patients behind occurred between $t - T$ and $t - T + \Delta t$, 2) the first of the patients left behind either shows up or is a no-show who does not reschedule service, and 3) during the period of time allocated for the service of this patient, there are $k + 1 - i$ appointment requests. The probability associated with this scenario is $D(i, t - T, t - T + \Delta t)(1 - r\gamma(k))\alpha(k + 1 - i)$.

Finally, it is also possible that 1) a departure that left $i = 1, \dots, k$ patients behind

occurred between $t - T$ and $t - T + \Delta t$, 2) the first of the patients left behind exhibits a no-show *and* reschedules service, and 3) during the period of time allocated for the service of this patient, there are $k - i$ appointment requests. The probability associated with this scenario is $D(i, t - T, t - T + \Delta t)r\gamma(k - 1)\alpha(k - i)$.

Combining these scenarios, we get

$$\begin{aligned}
D(k, t, t + \Delta t) &= p(0, t - T)\lambda N\Delta t((1 - r\gamma(k))\alpha(k) + r\gamma(k - 1)\alpha(k - 1)) \\
&\quad + (1 - r\gamma(k))\alpha(0)D(k + 1, t - T, t - T + \Delta t) \\
&\quad + \sum_{i=1}^k ((1 - r\gamma(k))\alpha(k + 1 - i) + r\gamma(k - 1)\alpha(k - i)) \\
&\quad \times D(i, t - T, t - T + \Delta t), \tag{A1}
\end{aligned}$$

or, equivalently,

$$\begin{aligned}
d(k, t) &= p(0, t - T)\lambda N((1 - r\gamma(k))\alpha(k) + r\gamma(k - 1)\alpha(k - 1)) \\
&\quad + (1 - r\gamma(k))\alpha(0)d(k + 1, t - T) \\
&\quad + \sum_{i=1}^k ((1 - r\gamma(k))\alpha(k + 1 - i) + r\gamma(k - 1)\alpha(k - i)) \\
&\quad \times d(i, t - T). \tag{A2}
\end{aligned}$$

Note that for $k = 0$ we get, following the same argument,

$$d(0, t) = p(0, t - T)\lambda N(1 - r\gamma(0))\alpha(0) + (1 - r\gamma(0))\alpha(0)d(1, t - T). \tag{A3}$$

On the other hand, for $k = K - 1$ the term $D(k + 1, t - T, t - T + \Delta t)$ in (A1) disappears since there can be no departure which leaves behind K customers. Also, if at $t - T$ the system was empty, and there was an arrival of a patient between $t - T$ and $t - T + \Delta t$, any number of subsequent arrivals during the service duration above $K - 2$ will result in the same future state trajectory. Similarly, if between $t - T$ and $t - T + \Delta t$ there was a departure which left $i = 1, \dots, K - 1$ patients behind, then any number of subsequent arrivals during the service duration above $K - i - 1$ will result in the same future state

trajectory. Thus,

$$\begin{aligned}
d(K-1, t) &= p(0, t-T) \lambda N \left(\left(1 - \sum_{i=0}^{K-2} \alpha(i) \right) (1 - r\gamma(K-1)) + r\gamma(K-2) \alpha(K-2) \right) \\
&\quad + \sum_{i=1}^{K-1} d(i, t-T) \\
&\quad \times \left(\left(1 - \sum_{j=0}^{K-1-i} \alpha(j) \right) (1 - r\gamma(K-1)) + r\gamma(K-2) \alpha(K-1-i) \right). \quad (\text{A4})
\end{aligned}$$

Finally, (7)-(9) are identical to the results of Proposition 2 in Garcia et al. (2002).

Proof of Proposition 2

The stationary solution to (7)-(9) satisfy

$$\pi(k) = \frac{d^*(k)}{\lambda N}, \quad k = 0, \dots, K-1. \quad (\text{A5})$$

where

$$\begin{aligned}
\pi(k) &= \lim_{t \rightarrow \infty} p(k, t), \\
d^*(k) &= \lim_{t \rightarrow \infty} d(k, t), \quad k = 0, \dots, K-1. \quad (\text{A6})
\end{aligned}$$

Note that

$$\pi(K) = 1 - \sum_{k=0}^{K-1} \pi(k). \quad (\text{A7})$$

From (6) it follows that

$$d^*(0) = d^*(0)(1 - r\gamma(0))\alpha(0) + (1 - r\gamma(0))\alpha(0)d^*(1), \quad (\text{A8})$$

$$\begin{aligned}
d^*(k) &= d^*(0) \left((1 - r\gamma(k))\alpha(k) + r\gamma(k-1)\alpha(k-1) \right) + (1 - r\gamma(k))\alpha(0)d^*(k+1) \\
&\quad + \sum_{i=1}^k \left((1 - r\gamma(k))\alpha(k+1-i) + r\gamma(k-1)\alpha(k-i) \right) d^*(i),
\end{aligned}$$

$$k = 1, \dots, K-2,$$

$$\begin{aligned}
d^*(K-1) &= d^*(0) \left(\left(1 - \sum_{i=0}^{K-2} \alpha(i) \right) (1 - r\gamma(K-1)) + r\gamma(K-2)\alpha(K-2) \right) \\
&\quad + \sum_{i=1}^{K-1} \left(\left(1 - \sum_{j=0}^{K-1-i} \alpha(j) \right) (1 - r\gamma(K-1)) + r\gamma(K-2)\alpha(K-1-i) \right) d^*(i). \quad (\text{A9})
\end{aligned}$$

Further, defining

$$f(k) = \frac{d^*(k)}{d^*(0)}, \quad k = 0, \dots, K-1, \quad (\text{A10})$$

we get

$$\begin{aligned} f(0) &= 1, \\ f(1) &= \frac{e^\rho}{1 - r\gamma(0)} - 1, \end{aligned} \quad (\text{A11})$$

and, using (A9),

$$\begin{aligned} f(k+1) &= \frac{e^\rho}{(1 - r\gamma(k))} (f(k) - (1 - r\gamma(k))\alpha(k) - r\gamma(k-1)\alpha(k-1)) \\ &\quad - \frac{e^\rho}{(1 - r\gamma(k))} \left(\sum_{i=1}^k ((1 - r\gamma(k))\alpha(k+1-i) + r\gamma(k-1)\alpha(k-i)) f(i) \right), \\ k &= 1, \dots, K-2. \end{aligned} \quad (\text{A12})$$

Note that (A10) is equivalent to

$$\pi(k) = f(k)\pi(0), \quad k = 0, 1, \dots, K-1, \quad (\text{A13})$$

and that, in steady state, the overall arrival rate of patients to the system must be equal to the overall departure rate of patients from the system:

$$\lambda N \left(\sum_{k=0}^{K-1} \pi(k) \right) = \frac{1}{T} \left(\sum_{k=1}^K (1 - r\gamma(k)) \pi(k) \right). \quad (\text{A14})$$

Note that (A14) is equivalent to

$$\begin{aligned} \rho\pi(0) \left(\sum_{k=0}^{K-1} f(k) \right) &= \left(\sum_{k=0}^K (1 - r\gamma(k)) \pi(k) \right) - (1 - r\gamma(0)) \pi(0) \\ &= 1 - (1 - r\gamma(0)) \pi(0) \\ &\quad - r \left(\pi(0) \sum_{k=0}^{K-1} \gamma(k) f(k) + \gamma(K) \left(1 - \pi(0) \left(\sum_{k=0}^{K-1} f(k) \right) \right) \right) \end{aligned} \quad (\text{A15})$$

so that

$$\pi(0) \left(\rho \left(\sum_{k=0}^{K-1} f(k) \right) + (1 - r\gamma(0)) + r \left(\sum_{k=0}^{K-1} (\gamma(k) - \gamma(K)) f(k) \right) \right) = 1 - r\gamma(K), \quad (\text{A16})$$

or

$$\pi(0) = \frac{1 - r\gamma(K)}{1 - r\gamma(K) + \rho \left(\sum_{i=0}^{K-1} f(i) \right) - r \sum_{i=1}^{K-1} (\gamma(K) - \gamma(i))f(i)} \quad (\text{A17})$$

Combining (A7) with (A13) and (A17), we get

$$\begin{aligned} \pi(0) &= \frac{1 - r\gamma(K)}{1 - r\gamma(K) + \rho \left(\sum_{i=0}^{K-1} f(i) \right) - r \sum_{i=1}^{K-1} (\gamma(K) - \gamma(i))f(i)}, \\ \pi(k) &= \frac{(1 - r\gamma(K))f(k)}{1 - r\gamma(K) + \rho \left(\sum_{i=0}^{K-1} f(i) \right) - r \sum_{i=1}^{K-1} (\gamma(K) - \gamma(i))f(i)}, \quad k = 1, \dots, K-1, \\ \pi(K) &= 1 - \frac{(1 - r\gamma(K)) \left(\sum_{i=0}^{K-1} f(i) \right)}{1 - r\gamma(K) + \rho \left(\sum_{i=0}^{K-1} f(i) \right) - r \sum_{i=1}^{K-1} (\gamma(K) - \gamma(i))f(i)}. \end{aligned} \quad (\text{A18})$$

Proof of Proposition 3

Using (17), we get for $k = 1$:

$$\pi(1) = \frac{\rho}{(1 - r\gamma(0))} \pi(0). \quad (\text{A19})$$

Further, for $k = 2$ we obtain

$$\left(\lambda N + T^{-1} (1 - r\gamma(0)) \right) \pi(1) = \pi(0) \lambda N + \pi(2) T^{-1} (1 - r\gamma(1)), \quad (\text{A20})$$

so that

$$\begin{aligned} \pi(2) &= \left(\frac{\rho + (1 - r\gamma(0))}{(1 - r\gamma(1))} \right) \pi(1) - \frac{\rho}{(1 - r\gamma(1))} \pi(0) \\ &= \frac{\rho}{(1 - r\gamma(1))} \pi(1) = \left(\frac{\lambda N T}{(1 - r\gamma(1))} \right)^2 \pi(0). \end{aligned} \quad (\text{A21})$$

Assuming that for all $k = 1, \dots, M < K - 1$

$$\pi(k) = \left(\frac{\rho}{1 - r\gamma(k)} \right)^k \pi(0), \quad (\text{A22})$$

we obtain for $k = M + 1$

$$(\rho + (1 - r\gamma(M - 1))) \pi(M) = \pi(M - 1) \rho + \pi(M + 1) (1 - r\gamma(M)), \quad (\text{A23})$$

so that

$$\begin{aligned}
\pi(M+1) &= \frac{1}{(1-r\gamma(M))} ((\rho + (1-r\gamma(M-1)))\pi(M) - \pi(M-1)\rho) \\
&= \frac{\rho\pi(M)}{(1-r\gamma(M))} = \left(\frac{\rho}{1-r\bar{\gamma}(M+1)} \right)^{M+1} \pi(0).
\end{aligned} \tag{A24}$$

Further, for $k = K$ we get

$$\pi(K) = \frac{\rho}{(1-r\gamma(K-1))} \pi(K-1) = \left(\frac{\rho}{1-r\bar{\gamma}(K)} \right)^K \pi(0), \tag{A25}$$

Finally, the value of $\pi(0)$ is obtained using the normalization condition (18):

$$\pi(0) \left(1 + \frac{\rho}{(1-\bar{\gamma}(1))} + \dots + \left(\frac{\rho}{1-\bar{\gamma}(K)} \right)^K \right) = 1 \Rightarrow \pi(0) = \left(1 + \sum_{k=1}^K \left(\frac{\rho}{1-\bar{\gamma}(k)} \right)^k \right)^{-1}. \tag{A26}$$