

# Reserving Capacity for Urgent Patients in Primary Care

Gregory Dobson

Simon School of Business, University of Rochester, Rochester, New York 14627, USA, greg.dobson@simon.rochester.edu

Sameer Hasija

INSEAD, 1 Ayer Rajah Avenue, Singapore 138676, Singapore, sameer.hasija@insead.edu

Edieal J. Pinker

Simon School of Business, University of Rochester, Rochester, New York 14627, USA, pinker@simon.rochester.edu

This paper examines the effect of the common practice of reserving slots for urgent patients in a primary health care practice on two service quality measures: the average number of urgent patients that are not handled during normal hours (either handled as overtime, referred to other physicians, or referred to the emergency room) and the average queue of non-urgent or routine patients. We formulate a stochastic model of appointment scheduling in a primary care practice. We conduct numerical experiments to optimize the performance of this system accounting for revenue and these two service quality measures as a function of the number of reserved slots for urgent patients. We compare traditional methods with the advanced-access system advocated by some physicians, in which urgent slots are not reserved, and evaluate the conditions under which alternative appointment scheduling mechanisms are optimal. Finally, we demonstrate the importance of patient arrival dynamics to their relative performance finding that encouraging routine patients to call for same-day appointments is a key ingredient for the success of advanced-access.

*Key words:* advanced-access; primary care; appointment scheduling; urgent patients

*History:* Received: October 2008; Accepted: August 2010, after 2 revisions.

## 1. Introduction

Providing timely service is a key goal of primary health care service providers. In its report, *Crossing the Quality Chasm: A New Health System for the 21st Century*, the Institute of Medicine's Committee on Quality Care in America (2001) recognized timeliness as one of the six key measures for improvement of the quality of health care in the United States. Unfortunately in most primary care practices, patients wait days for the next available appointment. The paper addresses how appointment systems balance long appointment queues with the need to see urgent patients.

Routine and urgent patients are essentially two classes of demand that are competing for the resources of the practice. *Urgent* patients are those that must receive same-day care. The categorization of a patient as urgent or routine is not entirely medical. For some cases, the patient's symptoms would lead the physician to categorize him as urgent. In other cases, it is more the patient's own sense of urgency that drives the categorization. *Routine* patients are those that can wait. When fewer resources are available to routine patients the result is a long appointment queue that delays care for these patients. The ill effects of long delays for routine appointments are delay in care, lost patients, and no-shows. Because urgent patients must, by definition, receive same-day care, reducing the resources available

to them leads to a variety of phenomenon. Urgent patients are often seen in overtime hours, squeezed into the existing schedule by double booking, or referred to other physicians or emergency clinics thus breaking the continuity of care. For the purpose of this paper, we will refer to patients seen outside of normal appointment slots, by one of the above methods, as *urgent overflow* patients. All of the above approaches to handling urgent overflow patients impose costs on the practice, in terms of some combination of pressure on staff, their work hours, quality of care, or in-office patient waiting. Therefore, minimizing the number of urgent overflow patients is also an important performance goal for primary care.

The appointment scheduling system used by a medical practice is the mechanism by which resources are allocated to routine and urgent patients. There are many appointment scheduling systems observed in practice, and each physician may operate in his own idiosyncratic way. However, these scheduling systems can be broadly characterized, see Murray and Berwick (2003), as being one of three models, the traditional model, the carve-out model, and the advanced-access model. The traditional model stratifies appointment demands into two streams: *urgent* (same-day) and *routine*. In this model, a patient contacts a receptionist, who, after consultation with a nurse or the doctor, determines the urgency of the incoming call. Routine patients are scheduled several days or weeks into the

future while urgent patients are often scheduled by double booking an appointment slot because most slots have been scheduled in advance for the routine patients. Murray and Berwick (2003) argue that such a system performs inefficiently as it results in wasted capacity and/or inflated demand. Triaging reduces the time available to see patients and double booking can lead to high in-office waiting time for patients, which in turn may cause physicians to rush patients through the visit, potentially requiring more future visits. The carve-out model operates in a similar manner to the traditional model but reserves some number of appointment slots for urgent care. Murray and Berwick (2003) suggest that physicians often reserve more carve-out time than is strictly needed to meet the urgent demand, thus causing longer queues for routine patients and moreover wasting time triaging and scheduling patients. Murray and Tantau (2000) and Murray and Berwick (2003) advocate the adoption of the advanced-access model in order to reduce the delays in health care. The advanced-access system, as defined by them, differs from the traditional and carve-out models in several important ways. First, advanced-access does not distinguish between routine and urgent patients so no triaging is necessary. Second, all patients are offered appointments on the same day they call, so there will be a relatively small appointment queue. Third, patients are encouraged to call for appointments the day or the day before they want to come. Such a system will lead to reduced waiting times for patients, but it may increase the number of patients that are handled during overtime. The fundamental philosophy of advanced-access can be summarized as “do today’s work today.” To implement advanced-access successfully, a practice must work down the backlog of appointments and make sure that its capacity to supply service is well matched with its patient panel’s demand. The main goal of this paper is to use mathematical models of the above scheduling systems to compare how they manage the competing objectives of small routine patient queues and low urgent overflow.

Our models are based on the common scheduling approach termed *carve-out*, which reserves a fraction of the daily capacity (or slots) for urgent patients. Under *carve-out*, appointment slots are designated either routine or urgent. While urgent patients may be scheduled in routine slots when all urgent slots are filled, routine patients are never assigned to urgent slots. If no slots are available, routine patients are scheduled for later days while urgent overflow patients are assumed to be satisfied in one of the ways described above. By definition, in *carve-out* policies the urgent slots cannot be assigned to routine patients. In this paper, we define zero-carve-out as giving same-day appointments to patients on a first-come-first-serve basis, and if the current day’s

capacity is full, carrying over appointments of routine patients to the next day and satisfying all urgent overflow on the same day in one of the ways described above. This approach is similar to the traditional approach described above. Finally we model advanced-access, as zero-carve-out with the subtle difference that no distinction is made between routine and urgent patients and all are given same-day care. An important element of advanced-access, as defined by Murray and Berwick (2003), is that patients are encouraged to call for appointments as close as possible to the time they want their appointment. Therefore, we also model the relative timing of routine and urgent call arrivals and investigate how it affects the performance of the different scheduling mechanisms. We find that this plays a very important role in facilitating the success of advanced-access and in determining system performance. Another prescription for advanced-access is that supply and demand should be well balanced, therefore we construct two versions of our models, one in which supply and demand are well balanced and one in which they are not. We find that overloaded systems do not work well with advanced-access unless the handling of urgent overflow is relatively inexpensive compared with patient delay. Finally we also investigate, for a two-physician practice, how the appointment scheduling mechanism influences continuity of care. We find evidence that in group practices advanced-access improves continuity of care.

While many doctors believe that the delays are the inevitable result of overloaded practices, others believe (Murray and Berwick 2003) that quite often these waits are due to the inefficient operating procedures of the service provider’s office, and not due to resource scarcity. While there are many cases cited as advanced-access successes (Murray and Berwick 2003), there is also resistance to this approach from physicians (Shuster 2003). Some critics of advanced-access seem to show a lack of understanding of the queueing dynamics that underlie the operation of a primary care practice. This lack of understanding is not unexpected as physicians rarely have training in industrial engineering or related fields. It is exacerbated by the fact that the advocates for advanced-access have relied on heuristic arguments and examples of medical practices to advance their case rather than rigorous system models. One goal of this paper is to put the comparison of these different primary care appointment scheduling systems on a more rigorous footing.

In the next section, we review the relevant literature. In section 3, we formulate a model of patient service with a constant average arrival rate and analyze two special cases of the model. In section 4, we formulate a model with queue dependent arrival rates to capture situations in which supply and demand are

not well matched. We present results of numerical experiments comparing the alternative appointment scheduling systems in section 5 and summarize our findings in section 6.

## 2. Literature Review

Out-patient scheduling has been well studied in the operations research literature (see Bailey 1952, Jackson et al. 1964, Vissers and Wijngaard 1979, Cayirli and Veral 2003, Denton and Gupta 2003, Cayirli et al. 2008). Most of these papers study the application of appointment scheduling rules with patient in-office waiting time as the service criteria. In contrast, the literature in advanced-access focuses on the time between the patient's request and the scheduled appointment time. O'Hare and Corlett (2004) present benefits of the advanced-access system beyond decreasing the backlogs experienced by a medical practice. They suggest that the advanced-access patient scheduling along with matching patients with their primary care physician (PCP) results in better-compensated physicians who provide additional services to patients. According to O'Hare and Corlett (2004), such a scheduling practice leads to more efficient clinic operations because it reduces the number of triage nurses required and the use of urgent-care services or the emergency room. Other studies, O'Connor et al. (2006), Mehrotra et al. (2008), Belardi et al. (2004), Bennett and Baxley (2008), and Phan and Brown (2009) have also attempted to empirically determine the impact of advanced-access on a variety of practice performance measures. All of these studies have shown shorter delays for appointments. O'Connor et al. (2006) show fewer missed appointments. Belardi et al. (2004) have shown improved continuity of care while Phan and Brown (2009) have found it degraded this measure. Interestingly, there was noticeable variability across these studies in how advanced-access was defined and implemented. For example, Bennett and Baxley (2008) used carve-out as part of their advanced-access implementation, which is clearly inconsistent with how it is defined by Murray and Berwick (2003). Therefore, we think there is value to mathematically modeling the key features of advanced-access in this paper to add more rigor to the discussion of its performance.

Robinson and Chen (2009) model appointment scheduling for a single provider to compare traditional and advanced-access systems. Their focus is on the impact of no-shows, which we do not model here but they consider only one class of patients. They do not distinguish between urgent and routine patients. Gerchak et al. (1996) study advanced scheduling of elective surgery when time of elective surgery, arrivals of emergency surgery, and time of emergency surgery are all uncertain. Their system is analogous to ours, in

that the elective and emergency surgeries are analogous to routine and urgent patients, respectively. Their objective is to minimize the time required for all surgeries over regular time and the penalty for not scheduling some elective surgeries for that day. The key difference between our work and Gerchak et al. (1996) is that they assume that the demand for all elective surgeries for the day is known before the schedule is decided. In our paper we assume that patients call one by one (all before the day's work begins) and receive appointments when they call. To provide a patient with an appointment when he calls, the capacity of the day (and future days) needs to be known. Therefore, we look at a static capacity reservation policy, which we believe is most appropriate for our setting. In Gerchak et al. (1996), the schedule for each day is made only at the beginning of the day, therefore they examine a queue dependent acceptance rule for elective surgeries. Further, Gerchak et al. (1996) assume that all routine demand arrives before 'urgent' demand (by definition). We relax this assumption in our paper and we characterize the trade-off between the equilibrium queue of routine patients and the urgent overflow as a function of the urgent reservation level and the arrival pattern of urgent and routine patients.

Gupta and Wang (2008) develop a Markov decision process model for appointment booking in order to maximize revenue by modeling patient choice explicitly. They argue that a patient's perception of urgency is an important factor in determining his overall satisfaction, that is, a patient who perceives his need to be urgent incurs a higher cost of not being served on the same day. Similar to Gerchak et al. (1996), in Gupta and Wang (2008), all routine patients arrive first because they describe all patients that request same-day appointments as urgent patients. Green and Savin (2005) model the process of scheduling patients for a doctor's office as a dynamic program, with a  $T$ -day horizon. They determine daily threshold values that control the number of routine appointments offered for that day in order to maximize the revenue of the practice. Our focus is different than these papers, in that we focus on the effects of reserving capacity for urgent patients and the impact of the arrival dynamics of urgent and routine patients on the queue of routine patients and urgent overflow.

## 3. Model Formulation

Although individual practices differ, our experience observing primary care facilities (Dobson et al. 2009) suggests the following description is typical. A primary care practitioner receives calls from patients seeking appointments throughout the day. However, there is often a high call volume early in the morning, followed by a low call volume throughout the rest of the day. While the time required to provide care at a

physician’s office varies from patient to patient, medical offices schedule appointments using standardized slots,  $c$ , suppressing the details of how long a patient waits within a day. Service time variability will not change the number of routine patients seen during a particular day because routine patients are seen on an appointment basis. Service time variability may potentially induce an in-office wait for patients, but it does not change their expected time between the day of request for an appointment and the day of appointment. The office receives calls from patients with varying levels of urgency. We classify patients into two categories of urgency: *urgent* patients and *routine* patients. A patient calling for a routine appointment will be given the first available slot on or after the day he requests his appointment. If he does not receive a slot on his requested day he is placed in the queue, which we call the *routine queue*. Urgent patients are those patients that request to be seen the same day and the doctor agrees need to be seen on that day.

The appointment process works as follows. A patient calls on day  $T$  and requests an appointment for day  $T+t$ . If the patient is urgent ( $t = 0$ ), he is given an appointment on day  $T$ , if one is available in either an urgent slot or an unreserved routine slot. If one is not available, he becomes an urgent overflow patient and is handled according to the policy for this practice, e.g., seen in overtime on day  $T$ , or sent to another facility or physician. If the patient is routine, then he is given an appointment right away for the next available routine slot on or after  $T+t$ .

In our model we make the following operating assumptions:

- A1. Demand on a particular day is the number of patients who want to be seen by the physician on that day and not the number of patients who physically call on that day
- A2. Patients call one by one with appointment requests
- A3. Patients with same-day appointment requests call at the beginning of the day (before the work begins)
- A4. Patients are offered appointments based on availability of open slots at the point in time when they call
- A5. A routine patient accepts the first available appointment on or after his preferred date
- A6. An urgent patient is scheduled in a routine slot if no urgent slot is available and a routine slot is empty for that day
- A7. There is no renegeing from the routine queue
- A8. There is no balking from the routine queue

#### A9. Arrivals of routine and urgent patients are independent

Assumption A1 is a simple re-arrangement of demand and hence is not limiting. Assumption A2 is a close representation of reality as the probability of receiving two or more patient calls at the same time is negligible. Further, Assumption A2 allows us to sequence the arrival of routine and urgent appointment requests. Assumption A3 limits the state space of our mathematical model because it allows us to avoid tracking the exact timing of appointment requests. Gupta and Wang (2008) similarly assume that urgent patients arrive at the start of the workday and Gupta and Denton (2008) provide an illustrative example of the arrival process of calls to a primary care clinic where they observe that call volumes to the clinic are highest at the start of the day. A limitation of this assumption is that it ignores the possibility that some reserved urgent and unscheduled routine slots may not be utilized, even if there exists demand for the day, due to the late arrivals of some requests. This assumption implies that our model underestimates both the urgent overflow as well as the routine queue length. We do not believe this limitation is significant because a practice can place the urgent slots later in the day. If routine patients are scheduled sequentially from the earliest available slot, such a strategy greatly diminishes the likelihood of having unused slots for the day. Assumption A4 is consistent with how patients obtain appointments from their PCPs in reality. Assumptions A3 and A5 ensure that there are no lost slots because of the timing of requests. Assumption 6 conforms to practice. Assumption A7 is not often seen in reality, however, if routine patients decide to renege from the queue and not inform the physician’s office (i.e., become a no-show, visit another physician, or go to the emergency room), that action does not affect the queue, because it leaves the slot unused. Assumption A8 is relaxed later in the paper when we examine a model that is overloaded. Assumption 9 is standard.

We use the following notation:

- $c$  Capacity of the practice (number of patients that can be seen daily)
- $c_u$  Capacity reserved for urgent patients
- $c_r$  Daily unreserved capacity, i.e.,  $c_r = c - c_u$ , nominally for routine patients
- $q^k$  Length of routine queue at the beginning of day  $k$
- $\bar{q}$  Average length of the routine queue,  $E(q^k)$
- $D_u^k$  Number of urgent patients that request an appointment for day  $k$  (all patients call at the beginning of day  $k$ )
- $D_r^k$  Number of routine patients that request an appointment for day  $k$  (all patients call one by

- one on some day before or at the beginning of day  $k$ ; calls can arrive over different days)
- $p_r(x)$  Probability that demand of routine patients on a particular day is  $x$
- $p_u(x)$  Probability that demand of urgent patients on a particular day is  $x$
- $\lambda_r$  Mean daily demand of routine patients
- $\lambda_u$  Mean daily demand of urgent patients
- $\lambda$  Mean daily demand of all patients,  $\lambda_r + \lambda_u$
- $\alpha^k$  Number of urgent overflow patients for day  $k$
- $\bar{\alpha}$  Average daily number of urgent overflow patients,  $E(\alpha^k)$
- $S^k$  Number of routine slots occupied by urgent patients on day  $k$

Note that the exact definition of  $S^k$  depends on the sequence in which routine and urgent patients arrive for day  $k$  and we make this precise later in this section.

One measure of performance is the length of the queue at the beginning of the day, before any same-day requests, which we denote as  $q^k$ . The length of the queue is the number of patients who requested service for a previous day but remain unserved on day  $k$ . The queue length on day  $k+1$ ,  $q^{k+1}$ , satisfies:

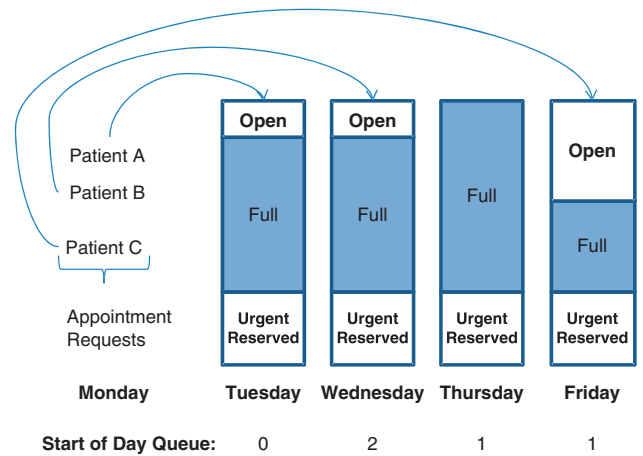
$$q^{k+1} = [q^k + D_r^k + S^k - c_r]^+, \quad (1)$$

where  $[x]^+$  is  $\max(x, 0)$ .

Let us examine the collection of patients represented by the sum  $q^k + D_r^k$ . Those represented by  $q^k$  requested their appointments for day  $k-1$  or earlier and were not served by day  $k-1$ . Those represented by  $D_r^k$  requested an appointment on day  $k$  and will be served on day  $k$  or later. They all received an appointment when they called. It would, of course, be possible to keep a detailed accounting of who was assigned each day, and at each point in time describe the state vector of who is scheduled over the future days. Yet, for our purpose, which is to measure the queue length,  $q^{k+1}$ , on morning  $k+1$ , i.e., those requesting service before  $k+1$  who are not yet served by the beginning of day  $k+1$ , we need to know only the quantities in Equation (1).

Let us consider a small example. First we discuss how patients would actually receive slots if they were placed in the order they called, and then we discuss how our model will assign them to slots and show that in terms of the queue length there is no difference. Assume all the previously booked patients called last week and received an appointment for the day they requested. Call the patients on Tuesday  $T_1, \dots, T_{n-1}$ , the patients on Wednesday,  $W_1, \dots, W_{n-1}$ , the patients on Thursday,  $R_1, \dots, R_n$ , and the patients on Friday,  $F_1$  and  $F_2$ , where  $n = c_r$ . On Monday, three new routine patients, A, B, and C, call (in that order) and request

Figure 1 Example of Patient Arrival and Appointment Scheduling Process



service for Tuesday. The schedules for Tuesday and Wednesday are nearly booked. There is one free slot available for both days. Thursday is completely booked, and Friday has many open slots (see Figure 1). Patient A receives an appointment for Tuesday, patient B receives an appointment for Wednesday, and patient C receives an appointment for Friday. No other patients call until Friday morning. The queue length is the number of patients who requested an appointment for a previous day but were carried over to a later day. The queue lengths at the beginning of each day are 0, 2, 1, and 1 for Tuesday, Wednesday, Thursday, and Friday, respectively. First note that A, B, and C are not in queue on Tuesday morning because they requested an appointment on Tuesday. The fact that they called earlier does not put them in the queue for Tuesday morning.

For our model, patients can be viewed as requesting their appointments on the day they want their appointments and we process first the queue from the previous day and then the requests for that day as suggested by Equation (1). Thus for our example, patients  $T_1, \dots, T_{n-1}$ , A, B, and C all request an appointment for Tuesday and  $T_1, \dots, T_{n-1}$ , and A would be placed on Tuesday. Patients B and C would be queued to the next day and would be in the queue on Wednesday morning. We could alternatively view patients  $W_1, \dots, W_{n-1}$  as requesting appointments for Wednesday. Patients B and C would be scheduled on Wednesday as well as  $W_1, \dots, W_{n-2}$ . Patient  $W_{n-1}$  would be carried over to Thursday morning. Patients  $R_1, \dots, R_n$  then request appointments for Thursday. Patients  $W_{n-1}$  and  $R_1, \dots, R_{n-1}$  would receive slots on Thursday. Patient  $R_n$  would be carried over to Friday morning. It is important to see that, although the sequencing of patients is different, the queue lengths at the beginning of Tuesday through Friday are the same, namely 0, 2, 1, and 1. The average delay incurred by patients is not affected by the

sequence. Thus Equation (1) needs to track only the queue length.

The number of urgent overflow patients for day  $k+1$  is,

$$\alpha^{k+1} = (D_u^{k+1} - c_u)^+ - S^{k+1}. \quad (2)$$

We model the objective function of the doctor's office as a profit maximization problem defined below,

$$\max_{c_u \in \{0, \dots, c\}} R_r \lambda_r + R_u \lambda_u - W\bar{q}(c_u) - H\bar{\alpha}(c_u), \quad (3)$$

where  $R_r$  is the value generated from a routine patient,  $R_u$  is the value earned from an urgent patient,  $W$  is the waiting cost (e.g., loss of good-will) per day of a routine patient, and  $H$  is the cost incurred for an urgent overflow patient (e.g., cost of overtime, loss of revenue due to emergency room referral). To simplify the presentation, without loss of generality, we adjust the units so that  $W = 1$ . Given this adjustment, it can be easily seen that the maximization problem in (3) can be reduced to the following cost minimization problem:

$$\min_{c_u \in \{0, \dots, c\}} \bar{q}(c_u) + H\bar{\alpha}(c_u). \quad (4)$$

In this model, all patients are eventually served; we call this the *base model*. In section 4 we introduce an alternative model for environments with queue dependent arrival rates of routine patients. We call this the *overload model*. In section 5 we investigate the sensitivity of the optimal  $c_u$  to the parameter  $H$  and different patient arrival dynamics. We also compare performance with that of the advanced-access system as defined by Murray and Berwick (2003). Advanced-access is equivalent to the system we have modeled above with zero-carve-out and with all patients being treated as urgent.

Next we present mathematical formulations of two special cases of the model described above in which we specify  $S^k$  in terms of known quantities in the model. In the first we assume that the demand of routine patients "arrives" first, before the demand of urgent patients. In the second special case we assume urgent patients arrive first, before the routine patients. Considering these special cases simplifies the analysis. Clearly reality is somewhere in between but the results from the special cases provide bounds on the general system performance.

### 3.1. Routine Patients Arrive First

For this section we assume that all routine patients arrive before the urgent patients, hereafter referred to as *routine-first*. Routine patients are likely to call before the day they want to see the physician, and because we are defining the demand of a particular day as the number of patients who wanted appointments for that day, it makes sense to view routine patients as "arriving" first. That is, the routine demand already

physically existed in the system but was only accounted for in our model on that day. Urgent patients by definition are those patients who call on a particular day and must be seen on the same day; therefore, their demand physically arrives on the same day. We also assume that both routine and urgent demand follow the Poisson distribution. Based on the assumptions, we obtain  $S^k = \min\{(D_u^k - c_u)^+, (c_r - q^k - D_r^k)^+\}$ . Substituting for  $S^k$  in Equation (1) we obtain the following expression for the queue length:

$$\begin{aligned} q^{k+1} &= [q^k + D_r^k + \min\{(D_u^k - c_u)^+, (c_r - q^k - D_r^k)^+\} - c_r]^+ \\ &= [q^k + D_r^k - c_r]^+. \end{aligned} \quad (5)$$

Let  $\Pr(q|Q)$  be the probability that  $q^{k+1} = q$  given that  $q^k = Q$  then

$$\Pr(q|Q) = \begin{cases} p_r(q - Q + c_r) & \text{if } q > 0, \\ \sum_{i=0}^{c_r - Q} p_r(i) & \text{if } q = 0. \end{cases} \quad (6)$$

LEMMA 1. *The Markov chain defined by (5) with transition probabilities given by (6) has a stationary distribution if and only if the mean demand of routine patients,  $\lambda_r$ , is less than the daily unreserved capacity of the clinic,  $c_r$  ( $\leq c$ ).*

The proofs of all results are contained in the Appendix.

Lemma 1 provides a simple test for determining whether too many slots have been allocated to urgent patients thus making the system unstable. The number of urgent overflow patients on day  $k$  can be expressed as

$$\begin{aligned} \alpha^k &= (D_u^k - c_u)^+ - \min\{(D_u^k - c_u)^+, (c_r - q^k - D_r^k)^+\} \\ &= [(D_u^k - c_u)^+ - (c_r - q^k - D_r^k)^+]^+. \end{aligned} \quad (7)$$

The quantity  $(D_u^k - c_u)^+$  gives us the number of urgent patients in excess of dedicated urgent slots for the day. The quantity  $(c_r - q^k - D_r^k)^+$  is the number of routine slots for the day that are not assigned to routine patients (because routine patients are assumed to arrive before urgent patients). Thus, the average number of urgent overflow patients each day is given by  $\bar{\alpha} = E[\alpha^k]$ .

### 3.2. Urgent Patients Arrive First

In this section, we assume that all urgent patients arrive before the routine patients, hereafter referred to as *urgent-first*. This is an extreme case that we do not expect to see in practice but find to be a mathematically convenient benchmark. First, patients in the queue are assigned to the routine slots and, if any remain, they queue to the next day identically to the routine-first case. Then, urgent patients are allocated to the urgent slots, any remaining routine slots, and overflow in that order. Note, to be consistent with the

routine-first case, we assume that routine patients are not scheduled during slots reserved for urgent patients even if those slots are vacant.

For this special case, it is possible to allow the left-over routine patients that arrive on a particular day to be scheduled in the unused urgent slots. Because all routine patients call for same-day appointments and all urgent patients arrive before the routine patients, there is no negative effect of allocating routine patients to unused urgent slots. Such a policy would make better use of capacity and reduce the routine queue length and in turn reduce the average urgent overflow because a reduced routine queue implies more available routine slots for the excess urgent patients. We do not take this approach here because we model this extreme case for the sole purpose of generating bounds on our two performance measures. If we allow routine patients to be scheduled in unused urgent slots, then the two performance measures obtained in this case will not act as bounds for our general realistic setting where arrivals of routine and urgent patients are mixed. Moreover, we would like this case to be consistent with the definition of carve-out policies for which urgent slots can be used only for urgent patients. Based on these assumptions we obtain  $S^k = \min\{(D_u^k - c_u)^+, (c_r - q^k)^+\}$ . Substituting for  $S^k$  in Equations (1) and (2) yields

$$q^{k+1} = [q^k + \min[(D_u^k - c_u)^+, (c_r - q^k)^+] + D_r^k - c_r]^+, \tag{8}$$

$$\alpha^k = [(D_u^k - c_u)^+ - (c_r - q^k)^+]^+. \tag{9}$$

Let  $\Pr(q|Q, y)$  be the probability that  $q^{k+1} = q$  given that  $q^k = Q$  and  $D_u^k = y$ . Thus we obtain the following expression for  $\Pr(q|Q, y)$ :

$$\Pr(q|Q, y) = \begin{cases} p_r(q - Q + c_r) & Q \geq c_r, \\ p_r(q - Q + c - y) & Q < c_r, y \geq c_u, Q + y - c_u \leq c_r, q > 0, \\ \sum_{i=0}^{c_r - Q - (y - c_u)} p_r(i) & Q < c_r, y \geq c_u, Q + y - c_u \leq c_r, q = 0, \\ p_r(q) & Q < c_r, y \geq c_u, Q + y - c_u > c_r, \\ p_r(q - Q + c_r) & Q < c_r, y < c_u, q > 0, \\ \sum_{i=0}^{c_r - Q} p_r(i) & Q < c_r, y < c_u, q = 0. \end{cases} \tag{10}$$

LEMMA 2. *The Markov chain defined by Equation (8) has a stationary distribution if and only if the mean demand of routine patients,  $\lambda_r$ , is less than the daily unreserved capacity of the clinic,  $c_r \leq c$ .*

### 3.3. Analysis

We want to investigate how the performance of the practice changes with the number of slots in a day that are reserved for urgent patients,  $c_u$ , hereafter referred

to as the *reservation level*. Let  $\bar{q}_r(c_u)$ ,  $\bar{q}_u(c_u)$ , and  $\bar{q}(c_u)$  be the average length of the routine queue, for a given  $c_u$ , for the cases in which routine patients arrive first, urgent patients arrive first, and the arrivals are mixed, respectively. Similarly, let  $\bar{\alpha}_r(c_u)$ ,  $\bar{\alpha}_u(c_u)$ , and  $\bar{\alpha}(c_u)$  be the average urgent overflow, for a given  $c_u$ .

PROPOSITION 1.  $\bar{q}_r(c_u) \leq \bar{q}(c_u) \leq \bar{q}_u(c_u)$ .

PROPOSITION 2.  $\bar{\alpha}_u(c_u) \leq \bar{\alpha}(c_u) \leq \bar{\alpha}_r(c_u)$ .

PROPOSITION 3.  $\bar{q}_r(c_u)$ ,  $\bar{q}_u(c_u)$ , and  $\bar{q}(c_u)$  are increasing in  $c_u$ , the reservation level.

## 4. Model with Queue Dependent Demand Rate (Overload Model)

In the previous section we assumed that no matter how long the queue is, a routine patient will always make an appointment. In reality a routine patient may seek an appointment with another physician when he cannot get an appointment with his physician within his desired time frame, or he may choose not to go to a physician at all. In such a case the demand rate of routine patients will be a decreasing function of the length of the routine queue. The demand rate of urgent patients will still be constant because these patients are always seen on the same day and hence are not affected by queue lengths. For the purpose of this analysis, we assume that the observed routine demand rate when the queue is of length  $q$ ,  $\lambda_r(q)$ , is given by,  $[\lambda_r - \epsilon q]^+$ , where  $\lambda_r$  is the underlying demand rate of routine patients,  $q$  is the length of the routine queue, and  $\epsilon > 0$  is a parameter representing the sensitivity of routine patients to the queue length. We also assume that the physician will not give any appointments beyond a certain queue length,  $M$ . This assumption is not limiting if we consider  $M$  as the panel size of the physician's practice.

LEMMA 3. *The system described above is an irreducible, aperiodic, and positive recurrent Markov chain and thus has a steady-state distribution for both the urgent- and routine-first cases.*

### 4.1. The Optimization Problem with Queue Dependent Arrival Rate

Let  $\bar{\Lambda}_r(c_u)$  be the average throughput of routine patients given a reservation level  $c_u$ . The optimization problem for the doctor's office is

$$\max_{c_u \in \{0, \dots, c\}} R_r \bar{\Lambda}_r(c_u) + R_u \lambda_u - W \bar{q}(c_u) - H \bar{\alpha}(c_u). \tag{11}$$

Again, we adjust the units to make  $W = 1$ , and the optimization problem can be reduced to

$$\max_{c_u \in \{0, \dots, c\}} F(c_u) = R_r \bar{\Lambda}_r(c_u) - \bar{q}(c_u) - H \bar{\alpha}(c_u), \tag{12}$$

because  $\lambda_u$  is constant. It is easy to see that the optimal reservation level is non-decreasing in  $H$  and non-increasing in  $R_r$ .

#### 4.2. Approximation for the Overload Scenario

In this section we present a simple heuristic method to determine the optimal reservation level for an overloaded system with queue dependent arrival rates, i.e., one in which the base demand exceeds the capacity. Owing to the heavy load conditions we can write down the following intuitive approximations:

1.  $\bar{\Lambda}_r(c_u) = c_r$ .
2.  $\bar{q}(c_u) = \frac{\lambda_r - c_r}{\epsilon}$ .
3.  $\bar{\alpha}(c_u) = E[(D_u - c_u)^+]$ .

The first is the result of assuming that, with a heavily loaded system and thus a long queue, the routine patients will capture all the routine slots and thus  $c_r$  will be the throughput of routine patients. The second uses the first and the function for observed demand,  $\lambda_r(q)$ , and solves for the queue length. The third simply says that urgent patients use only the  $c_u$  urgent slots. In numerical experiments we find these approximations to be extremely robust. For example, with a total capacity of 20 slots, a routine arrival rate of 28 or greater leads to differences of <2% in the objective function between the approximation and simulated values. Substituting these approximations into (12) yields

$$\max_{c_u \in \{0, \dots, c\}} F(c_u) = R_r(c - c_u) - \frac{\lambda_r - c + c_u}{\epsilon} - HE[(D_u - c_u)^+]. \quad (13)$$

To solve the maximization problem we compute the marginal value of an additional slot reserved for urgent patients, namely

$$\Delta F(c_u) = F(c_u + 1) - F(c_u) = - \left( R_r + \frac{1}{\epsilon} \right) + H(1 - \Pr(D_u \leq c_u)). \quad (14)$$

It is easy to see that  $\Delta F(c_u)$  is decreasing in  $c_u$  because  $\Pr(D_u \leq c_u)$  is increasing in  $c_u$ , therefore, there exists a unique  $c_u$  that maximizes the profit function. The optimal  $c_u$  is the least value of  $c_u$  such that

$$\Pr(D_u \leq c_u) \geq \frac{H - R_r - 1/\epsilon}{H}. \quad (15)$$

We find that the reservation level determined by this heuristic technique is highly accurate when compared with that produced by our numerical experiments. We can see from the structure of Equation (15) that the optimal reservation level is very sensitive to the relative values of the various revenue and penalty parameters.

## 5. Numerical Experiments

In this section we evaluate the system performance for different reservation levels for both the base and overload models. A day's work for an internal medicine practice typically involves seeing 20–25 patients. For our numerical computations we take  $c = 20$ . Different values of  $c$  give us the same insights as the value we have chosen. We conducted numerical studies for different values of  $\lambda_r$  and  $\lambda_u$ . Again, the insights from all the experiments are similar; hence, we present only a subset of our experiments in this paper. The parameter values for the additional numerical experiments were chosen from the range  $c \in \{20, 25\}$ ,  $\lambda_r \in \{15, 17.5\}$ , and  $\lambda_u \in \{5, 10\}$ .

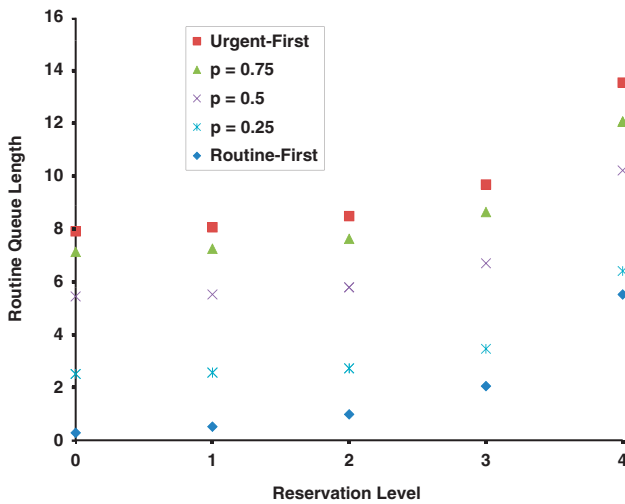
For the special cases of routine-first and urgent-first we perform exact calculations in the numerical studies. To analyze the more realistic situation in which urgent and routine case arrivals are mixed together we use Monte Carlo simulations. In these simulations, the demand of urgent and routine patients for a particular day is each given by the Poisson distribution. First, we schedule any routine patients in queue left over from the previous day. Next, we generate the day's demand for urgent and routine patients for that day, and finally we sequence the arrivals of the later two groups. To do this last step, we generate a uniform  $[0, 1]$  random number for each patient. If the number is less than a parameter  $p$ , which indicates the tendency of urgent patients to arrive before routine patients, then we schedule one patient from the urgent list, otherwise one from the routine list. This procedure is repeated until all the demand for the day has been scheduled. Using these three parameters ( $\lambda_u$ ,  $\lambda_r$ , and  $p$ ) allows us to specify different cases for the arrival dynamics while keeping constant the relative load of the two patient types on the system. The simulation was run for a period of 100,000 iterations with a run-in period of 1000 iterations to determine the two performance measures. The run length and the run-in period were found to be sufficient to ensure convergence to steady state. More details about the simulation are available on request.

The remainder of this section is organized as follows. In section 5.1 we present the results of numerical experiments on our base model. We present graphs that illustrate Propositions 1, 2, and 3. We then show the impact of both  $H$  and  $p$  on the total cost and the corresponding optimal scheduling policy. In section 5.2 we present the results of the overload case and in section 5.3 we analyze a case in which the primary health care practice has two physicians.

### 5.1. Base Model Results

Before determining the optimal reservation level we look at each component of the objective function, Equation (4), independently to better understand the main factors driving the results. In Figure 2, we plot

**Figure 2 Average Length of the Routine Queue vs. Reservation Level**  
 ( $c = 20, \lambda_r = 15, \lambda_u = 5$ )

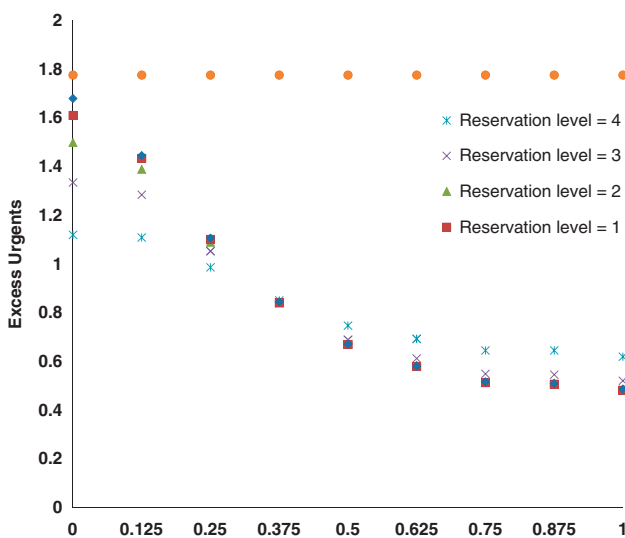


the average length of the routine queue as a function of the reservation level,  $c_u$ . Each series in Figure 2 represents a different demand arrival scenario, urgent-first ( $p = 1$ ), routine-first ( $p = 0$ ), or simulations of the intermediate cases with  $p$  in  $\{0.75, 0.5, 0.25\}$ .

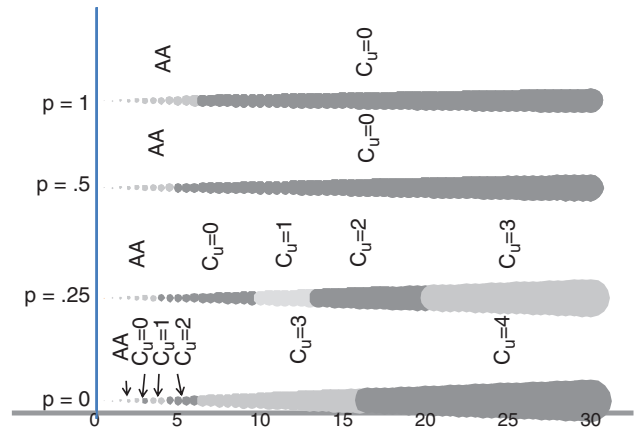
Figure 2 shows that, consistent with Proposition 1, the length of the routine queue for a given  $c_u$  increases with  $p$ . We also see, not surprisingly, that the routine queue increases with  $c_u$  (or with decreasing  $c_r$ ) in all scenarios, and, as  $c_u$  exceeds three, this increase is quite rapid. For advanced-access the routine queue will, by definition, always be zero.

In Figure 3, we plot the daily average number of urgent overflow cases as a function of the arrival dynamics ( $p$ ) for different reservation levels. For advanced-access, this quantity is constant across

**Figure 3 Urgent Overflow vs.  $p$  for Different Reservation Levels**  
 ( $c = 20, \lambda_r = 15, \text{ and } \lambda_u = 5$ )



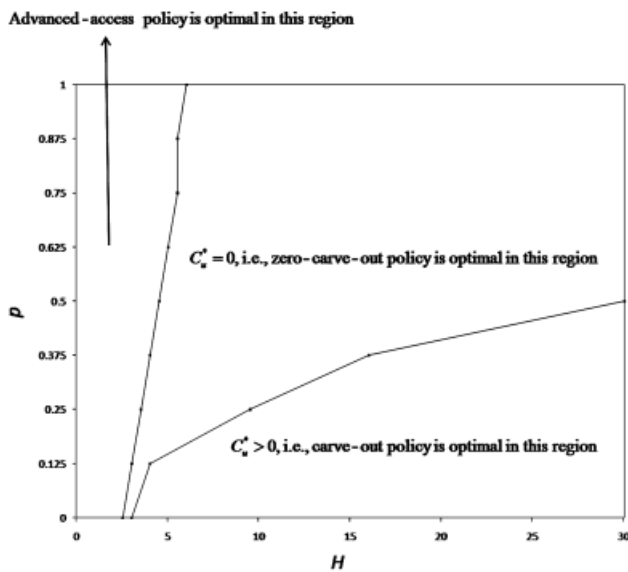
**Figure 4 Optimal Policies and Relative Total Costs Varying  $p$  and  $H$**   
 ( $c = 20, \lambda_r = 15, \text{ and } \lambda_u = 5$ )



$p$  and is shown as a benchmark. The figure shows that the average urgent overflow for a given  $c_u$  decreases with  $p$  (see Proposition 2). We also see that for low values of  $p$  (when routine patients are mostly arriving first), increasing the reservation level decreases the urgent overflow because it prevents the routine patients from completely filling the schedule. But, when  $p$  increases this relationship reverses and increasing the reservation level increases urgent overflow. This increase occurs because an increase in  $c_u$  causes the routine queue to increase, which reduces the room for urgent patients in the routine appointment slots. We also note that except for the lowest values of  $p$ , in absolute terms the amount of urgent overflow is not strongly affected by  $c_u$ . Finally, we see that advanced-access leads to significantly more overflow than the carve-out approaches because in advanced-access no patients are allowed to queue.

Figure 4 shows a bubble plot of the optimal cost and scheduling policy (AA stands for advanced-access) as a function of  $H$  for different arrival patterns for routine and urgent patients. We let  $H$  range from 0.05 to 30. If we assume that it takes a physician 20 minutes to treat an urgent case, then an  $H$  of 30 is similar to saying that we value 20 minutes of physician (plus staff and facility) work during overtime 30 times more than a 1-day delay that a patient experiences in seeing his physician. According to a cost survey of family practices (MGMA 2008), if we average total practice costs across patient encounters the average cost of handling a visit is approximately US\$120, which overstates the marginal cost because it includes fixed costs. Scaling by the conventional 1.5 to reflect overtime gives US\$180 per visit. The range of  $H = 0.05$  to 30 is equivalent to patients valuing 1 day of delay from US\$3600 to US\$6. Therefore, we believe an upper bound of 30 gives a sufficient range for the true value of  $H$ .

**Figure 5** Optimal Policy in the  $H$ - $p$  Space ( $c = 20$ ,  $\lambda_r = 15$ , and  $\lambda_u = 5$ )



In a typical medical practice that allows routine patients to book appointments in advance, we expect  $p$  to be small or zero. However, one tenet of advanced-access is not to book routine patients in advance but rather to have them call the day they want service. Such a policy would have urgent and routine patients mixed together randomly suggesting a  $p$  closer to 0.5. Therefore, in Figure 4, we examine the impact of both  $p$  and  $H$  on total costs and optimal scheduling policies. The size of the bubbles in Figure 4 represents the scaled total cost of the system. Figure 5 summarizes Figure 4 showing which of the three scheduling policies is optimal in the  $H$ - $p$  space. For a given  $H$  and  $p$ , and arrival rates  $\lambda_r$ ,  $\lambda_u$  a primary care practice can choose its optimal policy based on figures like Figures 4 and 5. For example, if a practice has determined that the value of  $H = 15$  and  $p = 0.25$ , then the optimal policy is to reserve two slots for urgent patients. The figures are based on numerical experiments where  $H \in \{0.05, 0.55, \dots, 30.05\}$  and  $p \in \{0, 0.125, \dots, 1\}$ . For clarity we only present a subset of the results in Figure 4. Readers can request the results of the complete set of numerical experiments from the authors.

We see that for a fixed value of  $H$  it is optimal to reserve more capacity for smaller values of  $p$  than for larger values because when routine patients come first they get better access and are less likely to queue while urgent cases are forced into overflow. For  $p$  fixed and low, as  $H$  increases it is optimal to reserve more capacity because you want to reduce urgent overflow. We can also clearly see that for low values of  $p$ , the range of  $H$  values for which it is optimal to reserve no capacity for urgent cases is smaller than for larger  $p$ . For high values of  $H$  ( $H \geq 13.55$  for our range

of parameters), the cost is decreasing in  $p$ . Therefore, getting routine patients to call for appointments closer to the date when they want to come decreases costs and facilitates zero-carve-out scheduling. The critical thing to note is that for our range of parameter values, if  $p \geq 0.5$  it is not optimal to reserve any capacity for urgent patients even if the cost of overflow is very high. For intermediate values of  $H$ ,  $H \in [7.05, 13.55)$  and  $H \in [4.05, 7.05)$ , the cost is minimum at  $p = 0.5$  and  $p = 0.25$ , respectively. Note at both these ranges of  $H$ , the cost is minimized at a value of  $p$ , where zero-carve-out policy is optimal. Further note that for lower values of  $H$ , advanced-access or zero-carve-out policies are optimal irrespective of the value of  $p$ . This illustrates the importance of this element of the capacity reservation operating principles. Modifying patient appointment-making behavior, for intermediate to high values of  $H$ , is a key success factor to implementation of no-reservation systems. As overflow costs increase, however, pure advanced-access becomes less attractive and even while it may be optimal to have no reserved capacity for urgent patients it may be optimal to allow queuing of routine patients so that the practice does not incur too much overflow.

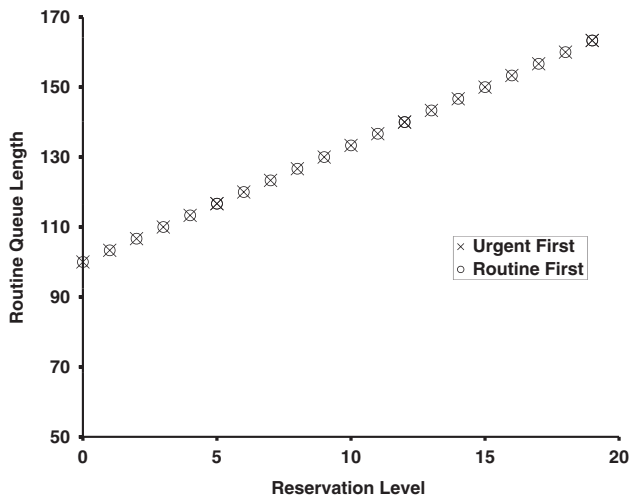
Another benefit of advanced-access, not captured in our analysis above, is that it greatly reduces the triaging and office-to-patient communications needed to set up appointments. In particular, under advanced-access, it is never necessary to determine if a patient is urgent while under more traditional approaches it is. With carve-out scheduling, the scheduler must determine who is allocated to the urgent slots and even with zero-carve-out it is necessary to know which patients to handle as overflow when the current day's appointment schedule is full. As a result under advanced-access the labor required per visit is reduced and therefore cost is potentially reduced. We view the cost of triaging as a fixed cost and therefore one should view Figures 4 and 5 as underestimating the range for which advanced-access would be optimal.

## 5.2. Results for the Overload Model

The results of the numerical experiments from section 5.1 are based on parameter values that ensure stability, i.e., demand and capacity are well matched. In practice, physicians may take on too many patients or may fall behind because of short-term capacity shortages, for example, their own or staff vacations. Our model of queue dependent demand rates is designed to capture these phenomena by allowing demand to exceed capacity for limited periods.

Figures 6 and 7 present the average length of the routine queue and the average urgent overflow for both the routine-first and urgent-first cases with the demand rate of routine patients dependent on the queue length,  $q$ . For our experiments, we assume

**Figure 6 Average Length of the Routine Queue vs. Reservation Level**  
 ( $c = 20, \lambda_r = 50, \lambda_u = 5, \epsilon = 0.3$ )

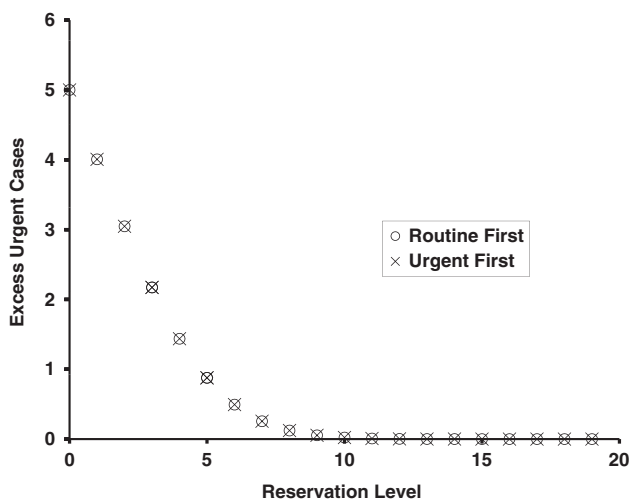


$\lambda_r(q) = [\lambda_0 - \epsilon q]^+, \lambda_0 > c$ , and  $\epsilon > 0$ . In this case the routine queue is almost always greater than the daily capacity, and therefore all the urgent patients that cannot be scheduled in the urgent slots are almost always seen during overtime. Thus in this overload scenario, the routine-first and urgent-first cases perform similarly. The number of appointment slots reserved for urgent patients has a strong impact on both performance measures and it is clear that in an overloaded practice improving one measure is at the expense of the other. The implication is that one cannot effectively implement advanced-access in a practice with insufficient capacity.

**5.3. Model of Practice with Two Physicians**

Many primary care practices are organized as group practices. In typical group practices, individual phy-

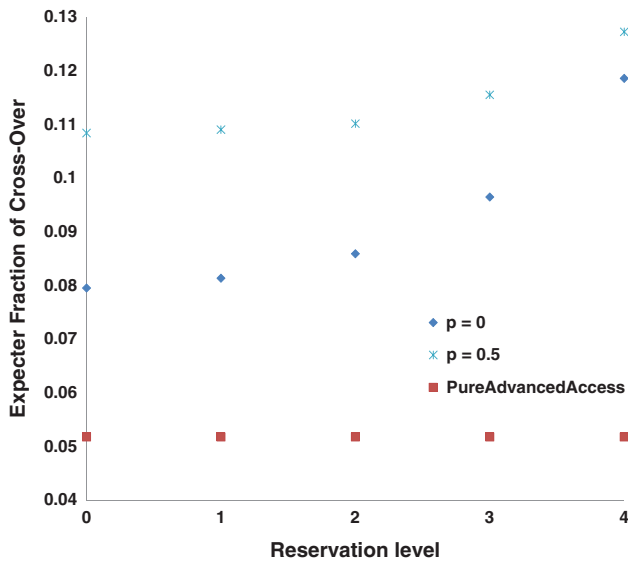
**Figure 7 Urgent Overflow vs. Reservation Level** ( $c = 20, \lambda_0 = 50, \lambda_u = 5, \epsilon = 0.3$ )



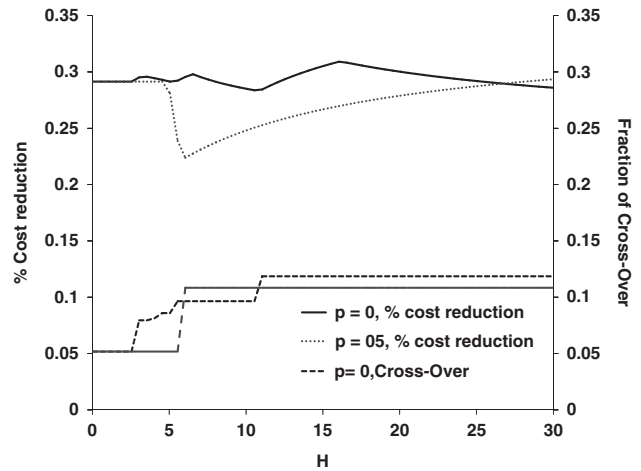
sicians have their own patient panels yet cover for each other to improve patient access to care. If a patient calls for an appointment with his physician, he may be offered an earlier appointment with a different physician in the group who has an available appointment slot. We refer to cases in which a patient sees a physician other than his primary physician as “cross-overs.” While the option of cross-over can improve access by taking advantage of physician resource pooling, it can reduce continuity of care, which may ultimately reduce care quality. There are a variety of protocols a practice can use for offering patients cross-over appointments and, of course, patients choose whether or not to accept them. Our goal here is not to identify the optimal protocol but rather to gain insight into how the appointment scheduling mechanism, i.e., traditional, carve-out, or advanced-access, affects continuity of care in a group practice. In this section we use a simulation model of a simple cross-over protocol in a two-physician practice to investigate how the appointment scheduling mechanism influences the trade-off between access and continuity of care. In our simulation model we assume that there are two physicians A and B and that each arriving patient is a member of panel A or B with equal probability. In a practice with carve-outs for urgent patients, we assume that patients are allocated to physicians each day as follows. If there is a queue of routine patients then they are handled sequentially. An A-patient will be assigned to a routine slot for physician A, and if there are none, a routine slot for physician B, and if there are none, he will be queued to the next day. A new arrival of a routine A-patient is assigned to a routine slot for physician A, and if there are none, a routine slot for physician B, and if there are none, he will be queued to the next day. A new arrival of an urgent A-patient is assigned an urgent slot with physician A. If none is available, then the following sequence is followed: routine slot with physician A, urgent slot with physician B, routine slot with physician B, and finally he becomes an overflow patient for physician A. B-patients are handled analogously for these three cases. In the case of advanced-access, there is no distinction made between urgent and routine patients and no queue is maintained. An arriving patient is assigned to his physician first, and, if there are no slots available, he is seen by the other physician. If there no slots available for either physician then this patient is overflow and each physician sees her own overflow patients in overtime, i.e., there is no cross-over in overtime.

For our numerical experiments we simulate a practice with two doctors using the same parameter values we used in the base model above. Each physician is assumed to have 20 appointment slots available each day. For each physician the arrival rate

**Figure 8** Expected Daily Fraction of Cross-over Patients vs. Urgent Reservation Level



**Figure 9** Cost Improvement and Expected Daily Fraction of Cross-over of Two-Physician Practice over Single Physician with Optimal Scheduling vs. Relative Cost of Overtime to Waiting Time



is 15 per day for routine patients and 5 per day for urgent patients. In addition to computing the system cost using the objective function in Equation (4) we calculate the expected number of cross-over cases per day as a measure of (lack of) continuity of care, i.e., when the number of cross-over cases is low then the level of continuity of care is high.

In Figure 8, we plot the expected daily number of cross-over patients as a function of urgent reservation level for the scenario when routine patients call first ( $p = 0$ ) and when routine and urgent patient calls are interspersed ( $p = 0.5$ ). We also plot the expected daily cross-over for advanced-access as a benchmark. It is clear from the figure that as the reservation level increases, continuity of care is diminished and that continuity of care is better for advanced-access vs. the alternative scheduling mechanisms. Increasing the number of slots reserved for urgent cases potentially wastes capacity and increases the average queue length. The wasted capacity causes more routine patients to cross-over. Some of the routine patients in the queue may also cross-over. In advanced-access systems, there is no wasted capacity and no queue. There may be somewhat more overflow work but we have assumed that this work is done without cross-over. As a result advanced-access has better continuity of care.

In Figure 9, we plot the relative cost improvement of a two-physician practice over a single-physician practice in which the optimal scheduling mechanism is used for different values of the relative cost of overtime to waiting time. The cost reduction fraction is computed as  $(2C_1 - C_2)/(2C_1)$ , where  $C_1$  and  $C_2$  are the optimal costs (routine waiting+urgent overflow)

of a single-physician and two-physician practice, respectively. In the same plot we show the expected daily cross-over. First we observe that the resource pooling of a group practice yields benefits between 20% and 30% over the solo practice. We note that there are some kinks in these graphs because as  $H$  changes the optimal reservation changes for both the single and two-doctor practices, but these policies do not necessarily change at the same points. To achieve these benefits, the practice must take on some degradation in the continuity of care. When  $H$  is low and it is optimal to operate with advanced-access, approximately 5% of visits are cross-over visits. For higher values of  $H$ , it is optimal to push more work into a queue and in these cases the method leads to more cross-over (approximately 10%) as patients take the first available physician. If a practice seeks to maintain better continuity of care, it must sacrifice some of the benefits of group practices. The results presented here overstate the amount of cross-over that occurs because we assume that patients will seek the earliest available appointment without regard to continuity of care, while in reality this may not be true.

#### 5.4. Implementation

To implement this model in practice, as a decision-making tool, requires estimating several parameters, in particular, arrival rates, revenue rates, overflow patient costs, and the cost of waiting. Some of these will have both objective and subjective components. In general we do not believe these to be particularly difficult to derive.

Arrival and revenue rate parameters can be derived from historical data on types of visits and insurance

reimbursements. The cost of an urgent overflow patient can be viewed as the marginal cost of seeing an additional patient scaled by some factor for the inconvenience of working late. This cost will have an objective component that involves the marginal labor rates for overtime of support staff as well as additional variable costs for keeping the practice open later such as energy usage. In a physician-owned practice, there is also the subjective question of the individual physician's indifference point for staying at work later.

The cost of waiting appears to be the most challenging parameter to estimate. However, there are a variety of approaches to take. For example, determining the cost of waiting for surgery in Canada is addressed in a study conducted for the Canadian Medical Association 2008. Alternatively, a practice could query its panel to determine a reservation price for being delayed in receiving an appointment by 1 day, a strategy that marketing professionals do regularly. Thus, while potentially involved the problem of parameter estimation seems solvable.

## 6. Conclusions and Extensions

In this paper, we investigated the carve-out and advanced-access appointment scheduling systems and evaluated the assertions of the advocates of advanced-access appointment scheduling. Our numerical experiments and simulations show that in a practice that is not overloaded (i.e.,  $\lambda_r < c$ ), the optimal policy (advanced-access, zero-carve-out, or  $c_u > 0$ , i.e., carve-out) depends both on the ratio of the cost of handling an urgent overflow patient to the cost of delaying a patient, parameter  $H$ , and on the arrival dynamics of the urgent and routine patients, captured in the parameter  $p$ . For very low values of  $H$ , advanced-access is optimal for all values of  $p$ . For higher values of  $H$ , carve-out is better with the amount of carve-out increasing in  $H$ . For a given  $H$ , the form of the solution is further affected by the arrival dynamics. Higher values of  $p$ , where a larger percentage of urgent cases make appointments first, result in lower costs and fewer reserved urgent slots. This observation supports the assertions of advocates for advanced-access who suggest that physicians encourage patients to call on the day they want to be seen. For overloaded practices, the arrival dynamics do not influence the optimal reservation level. In such a system, the trade-off between the benefit of serving a routine patient and the cost of not serving an urgent patient (or serving the urgent patient during overtime) determines the optimal reservation level.

Although a physician is unlikely to be able to substantially affect the parameter  $p$  for her practice (why tell a patient who has called in to make an appoint-

ment for a few days hence that he should call back on the desired day), the results of this paper suggest the following positive feedback loop. If carve-out is being used, then an increase in  $p$  results in fewer reserved urgent slots to minimize system costs. This change will, in turn, reduce the average queue and then the experience of the patient in terms of the likelihood of getting an appointment when calling on his desired appointment day would be better, i.e., more likely to get the appointment that day, thus making it more likely he would call on the desired appointment day for his next appointment, which implies an increase in  $p$ .

We have presented a simplified view of practice operations that has focused on appointment scheduling and not on the details of the execution of those visits during the day, in order to understand when a policy of zero-carve-out or advanced-access may be better. There are a number of issues that occur in practice that could be studied to enrich our understanding of advanced-access and other scheduling policies further. In the following, we discuss some real-life issues faced by primary health care practices that are not captured in our study. We have assumed that service times are constant for all cases. In practice, there is variability in patient visit durations. We do not think this assumption is very limiting because physicians can modulate their work pace. During a session, if a physician falls behind schedule because one patient visit has taken longer than expected she could shorten the time she spends with others. Also, variability in visit time will affect the variability in the hours worked in a day (and thus potentially affect the number of slots the doctor makes available) and the patient's in-office waiting time but not the two measures we investigated.

Another extension would be to model no-shows and double booking. Many practices experience some rate of no-shows and some use double booking and increasing panel size as a means to maintain high utilization of the physician. There is some empirical evidence to show that reducing queue length can reduce the no-show rate and the need for double booking (Moore et al. 2001). Furthermore, Robinson and Chen (2009) have shown in their models, in which they consider only a single class of patients, that advanced-access policies dominate traditional double-booking policies for a wide range of system parameters. It would be interesting to see whether advanced-access is again a robust solution with both urgent and routine patients. Intuitively, no-shows without double booking create a loss of capacity and thus a loss of revenue for the physician. Double booking may be able to maintain the revenue on average but the variance in workload on a day goes up so  $\alpha$  would be affected. A careful analysis of this interaction would be interesting.

In practice, there are also different classes of appointments: standard visits, extended visits, physicals, etc. The differences in service times across these appointment types are predictable but may affect the appointment queuing dynamics as will the fact that some appointments need to be scheduled farther ahead for clinical reasons. We have also not modeled the fact that a benefit of reducing the queue of routine patients with advanced-access is that the practice avoids turning routine patients into urgent patients because delayed treatment has caused a patient’s health to deteriorate. A potential extension of our model would make the fraction of urgent patients ( $\lambda_u/(\lambda_u+\lambda_r)$ ) endogenous and dependent on the queue length. Our intuition, developed in this paper, suggests that this change would tip the balance further in favor of advanced-access.

Finally we know that real-office practices experience transients that our steady-state model does not capture. First, there will be reductions in capacity when the physician must be out of the office for numerous events: sickness, vacation, professional meetings. Although physicians often form group practices to help with such situations, it is clear that the capacity of even group practices will vary over time for these reasons. Second, there will be predictable patterns of demand (Gupta and Denton 2008), e.g., Monday will be heavier because people become sick over the weekend, or August and September will be heavier as schools restart. Third, there are unpredictable spikes in demand such as a particularly virulent flu which will cause demand to rise for a few weeks. Some standard capacity and demand management techniques can deal with the predictable portion of these variations, yet this implies a deviation from the pure policies that we suggest for the steady state. It would be interesting to understand if dividing the patient requests into a third stream of requests, e.g., physicals, in addition to routine and urgent patients, and matching the remaining demand with the remaining capacity by scheduling the third stream appropriately, leads to any adjustments to our understanding of what policies are best. Lastly, as discussed earlier we do not explicitly model the exact time of the arrival of each appointment request. A more fine grained simulation model is required to capture the real-life dynamics of arrivals of patient calls.

**Acknowledgments**

The authors thank the special-issue editor Stefanos Zenios, the associate editor, and three referees for insightful comments that helped improve the paper.

**Appendix. Proofs**

PROOF OF LEMMA 1: The number of arrivals of routine cases in a day is Poisson and  $c_r > 0$ . Therefore the

Markov chain is irreducible and aperiodic. Assume the stationary probabilities for the Markov chain exist. If  $\pi(q)$  is the steady-state probability that the routine queue length is  $q$ , then

$$\pi(0) = \sum_{Q=0}^{c_r} \pi(Q) \sum_{i=0}^{c_r-Q} p_r(i). \tag{A1}$$

$$\pi(q) = \sum_{Q=0}^{c_r+q} \pi(Q) p_r(q - Q + c_r), \text{ for } q > 0. \tag{A2}$$

Define the generating functions,  $\tilde{\pi}(z), \tilde{p}_r(z)$  for each  $z$  in the open unit disk of the complex plane, as

$$\tilde{\pi}(z) = \sum_{q=0}^{\infty} \pi(q) z^q, \tag{A3}$$

$$\tilde{p}_r(z) = \sum_{i=0}^{\infty} p_r(i) z^i. \tag{A4}$$

Using Equations (A1)–(A4):

$$\tilde{\pi}(z) = \frac{\sum_{q=0}^{c_r-1} \pi(q) \{ \sum_{i=0}^{c_r-q} p_r(i) (z^{c_r} - z^{q+i}) \}}{z^{c_r} - \tilde{p}_r(z)}.$$

Now, since  $\lim_{z \rightarrow 1^-} \tilde{p}_r(z) = 1, \frac{\partial \tilde{p}_r(z)}{\partial z} \Big|_{z=1} = \lambda_r$  using L’Hôpital’s rule we obtain

$$\lim_{z \rightarrow 1^-} \tilde{\pi}(z) = \frac{\sum_{q=0}^{c_r-1} \pi(q) \{ \sum_{i=0}^{c_r-q} p_r(i) (c_r - q - i) \}}{c_r - \lambda_r}.$$

However, since  $\lim_{z \rightarrow 1^-} \tilde{\pi}(z) = \sum_{i=0}^{\infty} \pi(i)$ , this implies that stationary probabilities exist if and only if  $c_r > \lambda_r$ . □

PROOF OF LEMMA 2: The Markov chain is irreducible and aperiodic as the number of routine and urgent arrivals for a day follow Poisson distribution and  $c_r > 0$ . Assume the Markov chain is positive recurrent. This implies that stationary probabilities exist. We use a similar procedure as in proof of Lemma 1 to show that stationary probabilities exist if and only if  $c_r > \lambda_r$ . Let  $\pi(q)$  be the steady-state probability that the queue length of the routine cases is equal to  $q$ ,

$$\pi(q) = \sum_{Q=0}^{\infty} \sum_{y=0}^{\infty} \Pr(q^{k+1} = q | q^k = Q, D_u^k = y) p_u(y) \pi(Q).$$

The steady-state probabilities are

$$\pi(0) = \left\{ \begin{array}{l} p_r(0)\pi(c_r) + \sum_{Q=0}^{c_r-1} \sum_{y=c_u}^{c_r+Q} \left( \sum_{i=0}^{c_r-Q-(y-c_u)} p_r(i) \right) p_u(y)\pi(Q) \\ + \sum_{Q=0}^{c_r-1} \sum_{y=c_r+c_u-Q+1}^{\infty} p_r(0)p_u(y)\pi(Q) \\ + \sum_{Q=0}^{c_r-1} \sum_{y=0}^{c_u-1} \left( \sum_{i=0}^{c_r-Q} p_r(i) \right) p_u(y)\pi(Q) \end{array} \right\}. \tag{A5}$$

and for  $q > 0$ ,

$$\pi(q) = \left\{ \begin{array}{l} \sum_{Q=0}^{c_r-1} \pi(Q) \left( \sum_{y=c_u}^{c_r+c_u-Q} p_r(q+c_r-Q-(y-c_u))p_u(y) \right) \\ + \sum_{y=c_r+c_u-Q+1}^{\infty} p_r(q)p_u(y) + \sum_{y=0}^{c_u-1} p_r(q+c_r-Q)p_u(y) \\ + \sum_{Q=c_r}^{\infty} \pi(Q) \left( \sum_{y=0}^{\infty} p_r(q-Q+c_r)p_u(y) \right) \end{array} \right\}. \tag{A6}$$

Define the generating functions,  $\tilde{\pi}(z), \tilde{p}_r(z)$  for each  $z$  in the open unit disk of the complex plane as

$$\tilde{\pi}(z) = \sum_{q=0}^{\infty} \pi(q)z^q, \tag{A7}$$

$$\tilde{p}_r(z) = \sum_{i=0}^{\infty} p_r(i)z^i. \tag{A8}$$

Using Equations (A5)–(A8) we obtain

$$\tilde{\pi}(z) = \frac{\sum_{q=0}^{c_r-1} \pi(q)[F(z, q) + G(z, q) + H(z, q)]z^{c_r}}{z^{c_r} - \tilde{p}_r(z)},$$

where

$$F(z, q) = \sum_{y=0}^{c_u-1} p_u(y) \left( \sum_{i=0}^{c_r-q} p_r(i) + \tilde{p}_r(z)z^{q-c_r} - z^{q-c_r} \sum_{i=0}^{c_r-q} z^i p_r(i) \right),$$

$$G(z, q) = \sum_{y=c_u}^{c_r+c_u-q} p_u(y) \left( \sum_{i=0}^{c_r-q-(y-c_u)} p_r(i) + (\tilde{p}_r(z) - \sum_{i=0}^{c-q-y} z^i p_r(i))z^{q+y-c} \right),$$

$$H(z, q) = \left( 1 - \sum_{y=0}^{c-q} p_u(y) \right) \tilde{p}_r(z) - \frac{\tilde{p}_r(z)z^q}{z^{c_r}}.$$

Now, since  $\lim_{z \rightarrow 1^-} \tilde{p}_r(z) = 1, \frac{\partial \tilde{p}_r(z)}{\partial z} \Big|_{z=1} = \lambda_r$  using L'Hôpital's rule we obtain

$$\lim_{z \rightarrow 1^-} \tilde{\pi}(z) = \frac{\sum_{q=0}^{c_r-1} \pi(q) \left\{ \sum_{y=0}^{c_u-1} P_u(y) \left( \sum_{i=0}^{c_r-q} (c_r-q-i)p_r(i) \right) + \sum_{y=c_u}^{c_r+c_u-q} p_u(y) \left( \sum_{i=0}^{c_r-q-(y-c_u)} p_r(i)(c-q-y-i) + (y-c_u) \right) \right\} + \left( 1 - \sum_{y=0}^{c-q} p_u(y) \right) (c_r-q)}{c_r - \lambda_r}.$$

Because  $\lim_{z \rightarrow 1^-} \tilde{\pi}(z) = \sum_{i=1}^{\infty} \pi(i) c_r > \lambda_r$ .  $\square$

PROOF OF PROPOSITION 1: In order to prove  $\bar{q}_r(c_u) \leq \bar{q}(c_u) \leq \bar{q}_u(c_u)$ , it is sufficient to show that the following two inequalities hold for a given  $c_u$  (see Shaked and Shanthikumar, 1994):

$$Q_r^0 \leq q^0 \leq q_u^0 \tag{A9}$$

and

$$E[q_r^{k+1} | q_r^k = q_r] \leq E[q^{k+1} | q^k = q] \leq E[q_u^{k+1} | q_u^k = q_u] \text{ whenever } q_r \leq q \leq q_u. \tag{A10}$$

We know that  $q^{k+1} = [q + D_r^k + S^k - c_r]^+, q_r^{k+1} = [q_r + D_r^k - c_r]^+$  and  $q_u^{k+1} = [q_u + \min((D_u^k - c_u)^+, (c_r - q_u)^+) + D_r^k - c_r]^+$ .

As  $S^k$  is the number of routine slots occupied by urgent patients on day  $k, S^k \leq (D_u^k - c_u)^+$  and  $S^k \leq (c_r - q_k)^+$ . Therefore,  $0 \leq S^k \leq \min[(D_u^k - c_u)^+, (c_r - q_k)^+]$ .

Thus, whenever  $q_r \leq q \leq q_u$ ,

$$E[q_r^{k+1} | q_r^k = q_r, D_u^k, D_r^k] \leq E[q^{k+1} | q^k = q, D_u^k, D_r^k] \leq E[q_u^{k+1} | q_u^k = q_u, D_u^k, D_r^k].$$

$$\Rightarrow E[q_r^{k+1} | q_r^k = q_r] \leq E[q^{k+1} | q^k = q] \leq E[q_u^{k+1} | q_u^k = q_u].$$

Also,  $q_r^0 = q^0 = q_u^0 = 0$ .  $\square$

PROOF OF PROPOSITION 2: We first prove that  $\bar{\alpha}_u(c_u) \leq \bar{\alpha}(c_u)$ . For this proof it is sufficient to show that for a given demand realization  $(D_u^1, D_u^2, \dots, D_u^N), (D_r^1, D_r^2, \dots, D_r^N) \forall N \geq 0$  the following two conditions hold:

1.  $\alpha^0 - \alpha_u^0 \geq 0$ .
2.  $\sum_{k=0}^N (\alpha^k - \alpha_u^k) \geq 0$  given  $\sum_{k=0}^{N-1} (\alpha^k - \alpha_u^k) \geq 0$ .

By definition  $q_u^0 = q^0 = 0$ .

Assume that  $q_u^{k-1} \geq q^{k-1}$ . It was shown in the proof of Proposition 1 that  $q_u^k \geq q^k$ .

Thus by induction,  $q_u^k - q^k \geq 0 \forall k \geq 0$ .

Therefore, in order to show that conditions 1 and 2 hold, it is sufficient to show that the following two conditions hold:

$$(A) \alpha^0 - \alpha_u^0 \geq q_u^1 - q^1.$$

$$(B) \sum_{k=0}^N (\alpha^k - \alpha_u^k) \geq q_u^{N+1} - q^{N+1} \text{ given } \sum_{k=0}^{N-1} (\alpha^k - \alpha_u^k) \geq q_u^N - q^N.$$

Now

$$\begin{aligned}\alpha_u^k &= [(D_u^k - c_u)^+ - (c_r - q_u^k)^+]^+ \\ &= (D_u^k - c_u)^+ - \min[(D_u^k - c_u)^+, (c_r - q_u^k)^+].\end{aligned}$$

$$\begin{aligned}\alpha^k &= (D_u^k - c_u)^+ - S^k \\ &= (D_u^k - c_u)^+ - \min[(D_u^k - c_u)^+, (c_r - q^k - \tilde{D}_r^k)^+],\end{aligned}$$

where  $\tilde{D}_r^k$  is the number of routine patients that got a same-day appointment on day  $k$ .

In order to show that condition (A) is satisfied, we need to show that the following holds:

$$\begin{aligned}\min[(D_u^0 - c_u)^+, c_r] - \min[(D_u^0 - c_u)^+, (c_r - \tilde{D}_r^0)] \\ \geq q_u^1 - q^1.\end{aligned}\tag{A11}$$

By definition  $q_u^1 = [D_r^0 + \min[(D_u^0 - c_u)^+, c_r] - c_r]^+$  and  $q^1 = [D_r^0 + \min[(D_u^0 - c_u)^+, (c_r - \tilde{D}_r^0)] - c_r]^+$ .

We examine all possible cases:

Case 1:  $q_u^1, q^1 \neq 0$ .

$$\Rightarrow q_u^1 - q^1 = \min[(D_u^0 - c_u)^+, c_r] - \min[(D_u^0 - c_u)^+, (c_r - \tilde{D}_r^0)],$$

Case 2:  $q_u^1 \neq 0$  and  $q^1 = 0$ .

Equation (A11) can be rewritten as  $-\min[(D_u^0 - c_u)^+, (c_r - \tilde{D}_r^0)] \geq D_r^0 - c_r$ .

As  $q^1 = 0$ , the above inequality holds, therefore (A11) holds.

Case 3:  $q_u^1 = q^1 = 0$ .

We know that  $\min[(D_u^0 - c_u)^+, c_r] - \min[(D_u^0 - c_u)^+, (c_r - \tilde{D}_r^0)] \geq 0$ .

Therefore (A11) as well as condition (A) hold.

Assume that  $\sum_{k=0}^{N-1} (\alpha^k - \alpha_u^k) \geq q_u^N - q^N$ . We now want to show  $\sum_{k=0}^N (\alpha^k - \alpha_u^k) \geq q_u^{N+1} - q^{N+1}$  in order to complete the proof.

$$\sum_{k=0}^N (\alpha^k - \alpha_u^k) \geq q_u^N - q^N + \alpha^N - \alpha_u^N.$$

Therefore it is sufficient to show

$$\alpha^N - \alpha_u^N \geq (q_u^{N+1} - q_u^N) - (q^{N+1} - q^N),\tag{A12}$$

where  $\alpha^N - \alpha_u^N = \min[(D_u^N - c_u)^+, (c_r - q_u^N)^+] - \min[(D_u^N - c_u)^+, (c_r - \tilde{D}_r^N - q^N)^+]$ .

Again we examine all cases.

Case 1: If  $q_u^N, q^N \geq c_r$ , then  $(q_u^{N+1} - q_u^N) - (q^{N+1} - q^N) = 0$  and  $\min[(D_u^N - c_u)^+, (c_r - q_u^N)^+] - \min[(D_u^N - c_u)^+, (c_r - \tilde{D}_r^N - q^N)^+] = 0$ .

Hence, (A12) holds.

Case 2: If  $q_u^N \geq c_r$  and  $q^N < c_r$  then  $\min[(D_u^N - c_u)^+, (c_r - q_u^N)^+] = 0$ .

(i) If  $q^N + D_r^N + \min[(D_u^N - c_u)^+, (c_r - \tilde{D}_r^N - q^N)^+] - c_r \leq 0$ , then  $q^N + D_r^N - c_r \leq -\min[(D_u^N - c_u)^+, (c_r - \tilde{D}_r^N - q^N)^+]$ .

Therefore,  $(q_u^{N+1} - q_u^N) - (q^{N+1} - q^N) = D_r^N - c_r + q^N \leq \alpha^N - \alpha_u^N$ .

Hence, (A12) holds.

(ii) If  $q^N + D_r^N + \min[(D_u^N - c_u)^+, (c_r - \tilde{D}_r^N - q^N)^+] - c_r > 0$ , then  $(q_u^{N+1} - q_u^N) - (q^{N+1} - q^N) = -\min[(D_u^N - c_u)^+, (c_r - \tilde{D}_r^N - q^N)^+]$ .

Hence, (A12) holds.

Case 3: If  $q_u^N, q^N < c_r$ .

(i) If  $q^N + D_r^N + \min[(D_u^N - c_u)^+, (c_r - \tilde{D}_r^N - q^N)^+] - c_r > 0$  and  $q_u^N + D_r^N + \min[(D_u^N - c_u)^+, (c_r - q_u^N)^+] - c_r > 0$ , then

$$\begin{aligned}(q_u^{N+1} - q_u^N) - (q^{N+1} - q^N) &= \min[(D_u^N - c_u)^+, (c_r - q_u^N)^+] \\ &\quad - \min[(D_u^N - c_u)^+, (c_r - \tilde{D}_r^N - q^N)^+].\end{aligned}$$

Hence, (A12) holds.

(ii) If  $q^N + D_r^N + \min[(D_u^N - c_u)^+, (c_r - \tilde{D}_r^N - q^N)^+] - c_r \leq 0$  and  $q_u^N + D_r^N + \min[(D_u^N - c_u)^+, (c_r - q_u^N)^+] - c_r > 0$ .

(A12) holds because:  $(q_u^{N+1} - q_u^N) - (q^{N+1} - q^N) = D_r^N + \min[(D_u^N - c_u)^+, (c_r - q_u^N)^+] - c_r + q^N$ , and

$$\begin{aligned}D_r^N + \min[(D_u^N - c_u)^+, (c_r - q_u^N)^+] - c_r + q^N \\ \leq \min[(D_u^N - c_u)^+, (c_r - q_u^N)^+] \\ - \min[(D_u^N - c_u)^+, (c_r - \tilde{D}_r^N - q^N)^+].\end{aligned}$$

(iii) If  $q^N + D_r^N + \min[(D_u^N - c_u)^+, (c_r - \tilde{D}_r^N - q^N)^+] - c_r \leq 0$  and  $q_u^N + D_r^N + \min[(D_u^N - c_u)^+, (c_r - q_u^N)^+] - c_r \leq 0$ .

(a) If  $\min[(D_u^N - c_u)^+, (c_r - q_u^N)^+] = (c_r - q_u^N)^+$  then

$$\begin{aligned}\min[(D_u^N - c_u)^+, (c_r - q_u^N)^+] - \min[(D_u^N - c_u)^+, \\ (c_r - \tilde{D}_r^N - q^N)^+] \geq (c_r - q_u^N)^+ - (c_r - \tilde{D}_r^N - q^N)^+ \\ \geq q^N - q_u^N.\end{aligned}$$

(b) If  $\min[(D_u^N - c_u)^+, (c_r - q_u^N)^+] = (D_u^N - c_u)^+$  and  $\min[(D_u^N - c_u)^+, (c_r - \tilde{D}_r^N - q^N)^+] = (D_u^N - c_u)^+$ , then

$$\begin{aligned}\min[(D_u^N - c_u)^+, (c_r - q_u^N)^+] - \min[(D_u^N - c_u)^+, \\ (c_r - \tilde{D}_r^N - q^N)^+] = 0 \geq q^N - q_u^N.\end{aligned}$$

(c) If  $\min[(D_u^N - c_u)^+, (c_r - q_u^N)^+] = (D_u^N - c_u)^+$  and  $\min[(D_u^N - c_u)^+, (c_r - \tilde{D}_r^N - q^N)^+] = (c_r - \tilde{D}_r^N - q^N)^+$ , then  $(D_u^N - c_u)^+ - (c_r - \tilde{D}_r^N - q^N)^+ \geq 0$ .

$\Rightarrow \min[(D_u^N - c_u)^+, (c_r - q_u^N)^+] - \min[(D_u^N - c_u)^+, (c_r - \tilde{D}_r^N - q^N)^+] \geq 0 \geq q^N - q_u^N$ .

We have shown for all possible cases that (A12) holds. This completes the proof of the first part.

We now have to prove that  $\bar{\alpha}(c_u) \leq \bar{\alpha}_r(c_u)$ . For this proof it is sufficient to show that for a given demand realization  $(D_u^1, D_u^2, \dots, D_u^N), (D_r^1, D_r^2, \dots, D_r^N) \forall N \geq 0$  the following two conditions hold

1.  $\alpha_r^0 - \alpha^0 \geq 0$ .
2.  $\sum_{k=0}^N (\alpha_r^k - \alpha^k) \geq 0$  given  $\sum_{k=0}^{N-1} (\alpha_r^k - \alpha^k) \geq 0$ .

Using a method similar to the one used in the earlier proof it is easy to show that the above mentioned conditions hold.  $\square$

PROOF OF PROPOSITION 3: It is sufficient to show that  $\bar{q}(c_u)$  is increasing in  $c_u$  because  $\bar{q}_r(c_u)$  and  $\bar{q}_u(c_u)$  are special cases. Assume  $\hat{c}_u \geq c_u$ . Let  $q^k(\hat{c}_u) = \hat{q}^k$  and  $q^k(c_u) = q^k$ . We want to show that for a given demand realization  $(D_u^1, D_u^2, \dots, D_u^{k+1}), (D_r^1, D_r^2, \dots, D_r^{k+1}) \forall k \geq 0$  the following two conditions hold:

1.  $\hat{q}^0 \geq q^0$ .
2.  $\hat{q}^{k+1} \geq q^{k+1}$  if  $\hat{q}^k \geq q^k$ .

Condition 1 holds because  $\hat{q}^0 = q^0 = 0$ .

By definition,  $\hat{q}^{k+1} = [\hat{q}^k + D_r^k + \min((D_u^k - \hat{c}_u)^+, (c - \hat{c}_u - \hat{q}^k - \hat{D}_r^k)^+) - c + \hat{c}_u]^+$  and  $q^{k+1} = [q^k + D_r^k + \min((D_u^k - c_u)^+, (c - c_u - q^k - \tilde{D}_r^k)^+) - c + c_u]^+$ , where  $\tilde{D}_r^k$  and  $\hat{D}_r^k$  are the number of routine patients that got a same-day appointment on day  $k$  with  $c_u$  and  $\hat{c}_u$ , respectively. We examine all cases.

Case 1: If  $q^k \geq c - c_u$  and  $\hat{q}^k \geq c - \hat{c}_u$ :

$\hat{q}^{k+1} = [\hat{q}^k + D_r^k - c + \hat{c}_u]$  and  $q^{k+1} = [q^k + D_r^k - c + c_u]$ .

As  $\hat{q}^k \geq q^k$  and  $\hat{c}_u \geq c_u$ , condition 2 holds.

Case 2 (i): If  $q^k < c - c_u$ ,  $\hat{q}^k \geq c - \hat{c}_u$  and  $q^{k+1} = 0$  then condition 2 holds as  $\hat{q}^{k+1} \geq 0$ .

Case 2 (ii): If  $q^k < c - c_u$ ,  $\hat{q}^k \geq c - \hat{c}_u$  and  $q^{k+1} > 0$ , then

$$\begin{aligned} \hat{q}^{k+1} - q^{k+1} &= \hat{q}^k + \hat{c}_u - q^k \\ &\quad - \min((D_u^k - c_u)^+, (c - c_u - q^k - \tilde{D}_r^k)^+) - c_u \\ &\geq \hat{q}^k + \hat{c}_u - q^k - c + c_u + q^k + \tilde{D}_r^k - c_u \\ &= \hat{q}^k + \hat{c}_u - c + \tilde{D}_r^k \geq 0. \end{aligned}$$

Hence condition 2 holds.

Case 3: If  $q^k < c - c_u$  and  $\hat{q}^k < c - \hat{c}_u$ .

For this case in order to show condition 2 holds, it is sufficient to show

$$\begin{aligned} &(\hat{q}^k - q^k) + (\hat{c}_u - c_u) \\ &\geq \min((D_u^k - c_u)^+, (c - c_u - q^k - \tilde{D}_r^k)^+) \quad (A13) \\ &\quad - \min((D_u^k - \hat{c}_u)^+, (c - \hat{c}_u - \hat{q}^k - \hat{D}_r^k)^+). \end{aligned}$$

(i) If  $\min((D_u^k - \hat{c}_u)^+, (c - \hat{c}_u - \hat{q}^k - \hat{D}_r^k)^+) = (D_u^k - \hat{c}_u)^+$  and  $\min((D_u^k - c_u)^+, (c - c_u - q^k - \tilde{D}_r^k)^+) = \min((D_u^k - \hat{c}_u)^+, (c - \hat{c}_u - \hat{q}^k - \hat{D}_r^k)^+)$  then (A13) holds, which implies that condition 2 holds.

(a) If  $(D_u^k - c_u)^+ = 0$  and  $(D_u^k - \hat{c}_u)^+ = 0$ , then (A13) holds, which implies that condition 2 holds.

(b) If  $(D_u^k - c_u)^+ = D_u^k - c_u$  and  $(D_u^k - \hat{c}_u)^+ = 0$ .

Here  $D_u^k - \hat{c}_u \leq 0$ . This implies that  $D_u^k - c_u \leq \hat{c}_u - c_u$ . Therefore (A13) holds.

(c) If  $(D_u^k - c_u)^+ = D_u^k - c_u$  and  $(D_u^k - \hat{c}_u)^+ = D_u^k - \hat{c}_u$ , then (A13) holds.

(ii) If  $\min((D_u^k - \hat{c}_u)^+, (c - \hat{c}_u - \hat{q}^k - \hat{D}_r^k)^+) = (c - \hat{c}_u - \hat{q}^k - \hat{D}_r^k)$  and  $\min((D_u^k - c_u)^+, (c - c_u - q^k - \tilde{D}_r^k)^+) = \min((D_u^k - \hat{c}_u)^+, (c - \hat{c}_u - \hat{q}^k - \hat{D}_r^k)^+)$  then (A13) holds, which implies that condition 2 holds.

It can be shown that  $(\tilde{D}_r^k - \hat{D}_r^k) \geq 0$ , which implies that (A13) holds.  $\square$

PROOF OF LEMMA 3: The arrivals for both the urgent and routine cases is given by the Poisson distribution. The probability of starting from any state (state space being the routine queue length) and reaching state zero and then to any other state is  $> 0$ . Therefore, the chain is irreducible. Also  $\Pr(q^{k+1} = 0 | q^k = 0) > \sum_{i=0}^{c_r} (p_r(i) p_u(c_r - i)) > 0$  for any  $k \geq 0$ , thus this irreducible chain is aperiodic.

An irreducible aperiodic Markov chain has either all transient states or all null recurrent states in which case there exists no stationary distribution or has all positive recurrent states in which case there exists a unique steady-state distribution (Ross 1992). Thus, in order to complete this proof we need to show that not all states in this chain are transient or null recurrent.

Define  $P_{ij}^k = \Pr(q^{k+m} = j | q^m = i)$ , i.e., the probability that the process goes from state  $i$  to state  $j$  in  $k$  steps. The state space for this chain is  $\{0, 1, \dots, M\}$ . Thus we have

$$\sum_{j=0}^M P_{ij}^k = 1, \forall i \in \{0, 1, \dots, M\}, \forall k \geq 0.$$

If state  $j$  is transient or null recurrent, then  $P_{ij}^k \rightarrow 0$  as  $k \rightarrow \infty$ . If this held for all states, we would obtain the condition  $1 = \lim_{k \rightarrow \infty} \sum_{j=0}^M P_{ij}^k = \sum_{j=0}^M \lim_{k \rightarrow \infty} P_{ij}^k = 0$ , which is a contradiction. Therefore, not all states can be transient or null recurrent.  $\square$

## References

- Bailey, N. T. J. (1952). A study of queues and appointment systems in hospital outpatient departments, with special reference to waiting times. *J. R. Stat. Soc. B* 14: 185–199.
- Belardi, F. G., S. Weir, F. W. Craig. (2004). A controlled trial of an advanced access appointment system in a residency family medicine center. *Fam. Med.* 36: 341–345.
- Bennett, K. J., E. G. Baxley. (2008). The effect of a carve-out advanced access scheduling system on no-show rates. *Fam. Med.* 41(1): 51–56.
- Cayirli, T., E. Veral. (2003). Outpatient scheduling in health care: A review of literature. *Prod. Oper. Manag.* 12(4): 519–549.
- Cayirli, T., E. Veral, H. Rosen. (2008). Assessment of patient classification in appointment system design. *Prod. Oper. Manag.* 17(3): 338–344.
- Canadian Medical Association. (2008). *The Economic Cost of Wait Times in Canada. Prepared for Canadian Medical Association by Centre for Spatial Economics.* Milton, Ontario, Canada.
- Denton, B., D. Gupta. (2003). A sequential bounding approach for optimal appointment scheduling. *IIE Trans.* 35: 1003–1016.
- Dobson, G., E. Pinker, R. L. Van Horn. (2009). Division of Labor in Medical Office Practices. *Manuf. Serv. Oper. Manage.* 11: 525–537.
- Gerchak, Y., D. Gupta, M. Henig. (1996). Reservation planning for elective surgery under uncertain demand for emergency surgery. *Manage. Sci.* 42(3): 321–334.
- Green, L., S. Savin. (2005). Designing appointment systems for outpatient healthcare facilities. Presented at the INFORMS Conference, San Francisco, November 2005.
- Gupta, D., B. Denton. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Trans.* 40: 800–819.

- Gupta, D., L. Wang. (2008). Revenue management for a primary-care clinic in presence of patient choice, working paper. *Oper. Res.* **56**: 576–592.
- Institute of Medicine, Committee on Quality of Health Care in America. (2001). *Crossing the Quality Chasm: A New Health System for the 21st Century*. National Academy Press, Washington, DC.
- Jackson, R. R. P., J. D. Welch, J. Fry. (1964). Appointment systems in hospitals and general practice. *Oper. Res. Q.* **15**: 219–237.
- Mehrotra, A., L. Keehl-Markowitz, J. Z. Ayanian. (2008). Implementing open-access scheduling of visits in primary care practices: A cautionary tale. *Ann. Int. Med.* **148**: 915–922.
- MGMA. (2008). Cost Survey for Single-Specialty Practices—2008 Report Based Upon 2007 Data. Medical Group Management Association.
- Moore, C. G., P. Wilson-Witherspoon, J. C. Probst. (2001). Time and money: Effects of no-shows at a family practice residency clinic. *Fam. Med.* **33**: 522–527.
- Murray, M., D. M. Berwick. (2003). Advance access: Reducing waiting and delays in primary care. *J. Am. Med. Assoc.* **289**: 1035–1040.
- Murray, M., C. Tantau. (2000). Same day appointments: Exploding the access paradigm. *Fam. Pract. Manag.* **7**: 45–50.
- O'Connor, M. E., B. S. Matthews, D. Gao. (2006). Effect of open access scheduling on missed appointments, immunizations, and continuity of care for infant well-child care visits. *Arch. Pediat. Adolesc. Med.* **160**: 889–893.
- O'Hare, C. D., J. Corlett. (2004). The outcomes of open access scheduling. *Fam. Pract. Med.* **11**: 35–38.
- Phan, K., S. R. Brown. (2009). Decreased continuity of in a residency clinic: A consequence of open access scheduling. *Fam. Med.* **41**: 46–50.
- Robinson, L. W., R. R. Chen. (2010). The effects of patient no-shows on appointment scheduling policies. *Manuf. Serv. Oper. Manage.* **12**: 330–346.
- Ross, S. (1992). *Applied Probability Models with Optimization Applications*. Reprint edn. Dover Publications, Mineola, NY.
- Shaked, M., J. G. Shanthikumar. (1994). *Stochastic Orders and Their Application*. Academic Press, New York, NY.
- Shuster, M. (2003). Letter to the editor re: Advanced-access scheduling in primary care. *J. Am. Med. Assoc.* **290**(3): 332–333.
- Vissers, J., J. Wijngaard. (1979). The outpatient appointment system: Design of a simulation study. *Eur. J. Oper. Res.* **13**: 459–463.