Joint Panel Sizing and Appointment Scheduling in Outpatient Care

Christos Zacharias
Department of Management Science, School of Business Administration, University of Miami, Coral Gables, FL 33146.
czacharias@bus.miami.edu

Mor Armony
Department of Information, Operations & Management Sciences, Stern School of Business, New York University, New York, NY 10012.
marmony@stern.nyu.edu

Patients nationwide experience difficulties in accessing medical care in a timely manner due to long backlogs of appointments. Medical practices aim to utilize their valuable resources efficiently, provide timely access to care, and at the same time they strive to provide short waits for patients present at the medical facility. We address the joint problem of determining the panel size of a medical practice and the number of offered appointment slots per day, so that patients do not face long backlogs and the clinic is not overcrowded. We explicitly model the two time scales involved in accessing medical care: appointment delay (order of days, weeks) and clinic delay (order of minutes, hours). Closed-form expressions are derived for the performance measures of interest based on diffusion approximations. Our model captures many features of the complex reality of outpatient care, including patients’ no-shows, balking behavior, and random service times. Our analysis provides theoretical and numerical support for the optimality of an “Open Access” policy in outpatient scheduling, and demonstrates the importance of considering panel sizing and scheduling decisions in a joint framework.

Key words: patient flow management; panel-size; appointment scheduling; balking; diffusion approximations; open access

History: last update on Sep 26 2015

1. Introduction

Patients nationwide experience difficulties in accessing medical appointments in a timely manner due to long backlogs. Merritt (2014), by surveying 1399 medical offices in 15 US metropolitan areas throughout the year of 2013, found that the waiting time to get an appointment averages 18.5 days and depends on the specialty: 16.8 days for cardiology, 28.8 days for dermatology, 17.3 days for obstetrics/gynecology, 9.9 days for orthopedic surgery, 19.5 days for family physician.

More recently, in May 2014, major problems at the United States Department of Veterans Affairs (VA) with scheduling timely access to medical care became public. As reported in Kesling (2014), VA had an official goal that no more than 14 days would pass between a patient’s desired date for an appointment and the actual appointment date. Actual patient wait times were reported to average 115 days, and 84% of patients had to wait longer than the 14-day target.

The healthcare delivery system also suffers from patient no-show behavior. No-shows in medical care have been well documented, with no-show rates reaching up to 60%, depending on the
medical practice's characteristics (Cayirli et al. (2006)). Defife et al. (2010) report a 21% no-show rate in psychotherapy appointments, Dreher et al. (2008) report a 30% proportion of nonattendance in an outpatient obstetrics and gynecology clinic, Rust et al. (1995) report a 31% appointment failure rate in pediatric resident continuity clinics nationally. Unattended appointments result in clinic under-utilization, and limit the access for other patients who could have filled the missed slots.

Medical practices aim to utilize their valuable resources efficiently, provide timely access to care, and at the same time they strive to provide short waits for patients present at the medical facility. Appointment overbooking is one operational strategy employed by healthcare providers that addresses both the issues of long appointment backlogs and patient no-shows. On the other hand, overbooking potentially results in an overcrowded clinic, with increased patients’ waits and physicians’ overtime. As argued in Krueger (2009): “Patient time is an important input in the healthcare system. Failing to take account of patient time leads us to exaggerate the productivity of the health care sector, and to understate the cost of healthcare”. On the flip side, LaGanga and Lawrence (2007) show that a sensible practice of appointment overbooking can significantly improve a clinic’s performance by increasing patients’ access and improving physicians’ productivity.

Many papers have appeared in the literature on appointment scheduling. Cayirli and Veral (2003), Gupta and Denton (2008) provide literature surveys and overviews of the research challenges. One way to classify the existing literature is with respect to the type of delay addressed: appointment delay versus clinic delay. Appointment delay is defined as the time gap between the appointment request and the offered appointment. Clinic delay is the physical waiting time experienced by the patients once they arrive at the medical facility. Very few studies consider the appointment delay, and to the best of our knowledge, ours is the first to consider both, and to provide closed-form expressions for the performance measures of interest.

There are two main levers that healthcare providers can use to manage their patient flow, and consequently their productivity and the two aforementioned types of delay. The first one comes by controlling the demand side through their “Panel Size”, i.e., the size of the population of patients who receive their care from the practice on some regular basis. Choosing a proper panel size is a strategic decision and has a direct effect on the demand volume. Having an appropriately sized patient panel defines workload, permits prediction of patient demand, and improves outcomes of care (Murray et al. (2007)). The second lever is the appointment availability, i.e., the number of offered appointment slots per day, which is a tactical decision. It is important to consider both levers in a joint framework. A clinic’s panel size should not only depend on the desired demand volume, but also on the capacity of the system to serve patients. Conversely, appointment availability affects the system’s competence to provide timely access to the desired panel of patients.
It is intuitive that an increased panel size comes along with a more congested clinic and longer appointment delays. Apropos of the number of appointments to offer per day, as demonstrated later on in this paper, it is unclear as to how it affects the appointment delay. Intuitively, one would expect the appointment backlog to be decreasing in the daily rate that the clinic offers appointment slots. But note that there is a second order effect induced by the patients’ balking behavior: the more the appointment availability, the more likely it is that a patient will have access to a slot of her preference and actually book an appointment.

We address the joint problem of determining the panel size of a medical practice and the number of offered appointments per day, so that patients do not face long backlogs and the medical facility is not overcrowded. By optimizing over both levers in a joint framework, we are able to demonstrate that an “Open Access” policy is optimal in outpatient scheduling when both types of delay matter. Open Access (Murray and Tantau (2000)) is a scheduling paradigm in which the clinic attempts to “meet today’s demand today” by offering to each patient the opportunity for a same-day appointment with high probability.

The rest of the paper is structured as follows. In §2 we discuss the related literature. In §3 we introduce the modeling framework and define the optimization problem under study. In §4 and §5 we analyze the two queueing systems associated with each time scale respectively, and we provide analytical expressions for the performance measures of interest based on diffusion approximations. In §6 we optimize the daily net benefit of the medical facility from providing care to patients, based on the expressions derived in §4 and §5, with respect to the panel size and the appointment availability. Finally, in §7, we conclude and present future research directions. All the proofs appear in Appendix A.

2. Related Literature

There is an extensive literature on appointment scheduling, mostly motivated by healthcare applications. The vast majority of the literature focuses on the clinic delay and studies the trade-offs between the benefits of efficient resource utilization and the costs of patients’ waiting time and physician’s overtime. Hall (2012) provides a comprehensive review of models and methods used for scheduling the delivery of patient care for all parts of the healthcare system. The analysis may be based on anyone from a variety of approaches, including stochastic programming, queueing models, and simulation. Kaandorp and Koole (2007), Hassin and Mendel (2008), Klassen and Yoogalingam (2009), Robinson and Chen (2010), LaGanga and Lawrence (2012), Zacharias and Pinedo (2014, 2015) are some recent works that address the question of how to optimally allocate the offered appointment slots throughout the working day, taking into account patients’ no-show behavior. There are no analytical expressions for the patients’ waits and physicians’ overtime. Only
under certain assumptions on the service times’ distribution (deterministic or exponential), and
by assuming punctual patients, recursive expressions can be derived. In contrast to the aforemen-
tioned stream of literature, we do not consider the optimal intraday scheduling per se. By assuming
that patients arrive at the clinic according to a renewal process and that the service times are
iid random variables, we derive closed-form expressions for the performance measures of interest
based on diffusion approximations. We make use of these expressions, along with the analysis of
the appointment delay, in taking joint panel sizing and scheduling decisions.

Very few works take under consideration the appointment delay and the panel sizing problem.
Green et al. (2007) propose a probabilistic model to study the timeliness of care, while considering
the constraints on physicians working hours. They find that supply of medical appointments must
be substantially higher than the demand, in order to sustain Open Access scheduling. Green and
Savin (2008) and Liu and Ziya (2014), the most closely related papers to this study, focus also on
the appointment delay by modeling the appointment book as a single server queue in steady state
with state dependent no-shows.

Green and Savin (2008) develop expressions for the steady state distribution of the appointment
backlog by considering both $M/M/1/K$ and $M/D/1/K$ queueing models with delay dependent
no-show probabilities. They identify proper panel sizes for medical practices with a given capacity
that aim to implement an Open Access policy. In particular, their models help determine panel
sizes that would allow same day appointments with a certain high probability. In contrast, in
our study, rather than assuming a priori that Open Access is desirable, Open Access arises as
the optimal operational regime. In particular, we do not target a desired appointment delay; we
let appointment delay (among other performance measures) to be the outcome of a net benefit
maximization problem.

Liu and Ziya (2014) model the appointment queue as a single-server system with state dependent
no-shows. They address the problem of taking joint panel sizing and overbooking decisions, so as
to maximize a net reward function subject to the appointment delay not exceeding a certain level.
While they do not model the clinic delay explicitly, they do consider a penalty function (strictly
increasing and convex) for overbooking. We take a different approach in modeling patients’ access
to care by developing a two-stage queueing model, comprised by an appointment book (single
server queue with state-dependent balking) that provides input to the clinic, and which facilitates
the analysis of the trade-offs between appointment delay and clinic delay. A detailed comparison
between our study and the two aforementioned ones appears in §6.3.

Liu et al. (2015) show empirically that, among other operational attributes, both types of delay -
clinic and appointment - affect patients choices in accessing medical care and should be taken under
consideration when designing appointment systems. The importance of considering the separate
time scales involved in the context of inpatient care has been discussed in Armony et al. (2015) and Dai and Shi (2015). To the best of our knowledge, our work is the first to explicitly model the two time scales involved in accessing outpatient medical care: appointment delay (order of days, weeks) and clinic delay (order of minutes, hours), and to provide closed-form expressions for the performance measures of interest based on diffusion approximations. Luo et al. (2015) study performance analysis of a similar model. They do not, however, consider the question of how to optimize system performance with respect to the available levers. Two additional distinctive characteristics of our study are the state-dependent balking behavior of the patients who face long appointment backlogs, and the transient-state analysis of the in-clinic queue, which bear unique technical challenges.

Empirical studies in the literature consider the appointment delay as well. Balasubramanian et al. (2010, 2012) analyze the trade-offs between timely access and continuity of care. In the former study they propose a redesign of physicians’ panel-composition, based on data derived from a large group practice. In the latter work they investigate the value of flexibility in medical practices by addressing the problem of how to optimally allocate the available physicians’ slots among pre-scheduled and Open Access appointments.

Finally, we point out that the appointment scheduling problem, and specifically the number of offered appointments per day component, has natural connection with the control of perishable goods inventory (Bulinskaya (1964), Abad (1996)). Consider the multi-period inventory control problem of a product that perishes after one period, random demand, and state-dependent back-ordering. Positive inventory at the end of the period corresponds to unused time slots, which perish and cannot be used in future periods, and negative inventory corresponds to a backlog of appointments.

3. Problem Formulation

In this section we introduce the modeling framework and we formulate the optimization problem under study. Figure 1 depicts a schematic representation of the model.

Consider a clinic with a panel of size $N$ patients, which triggers a demand for scheduled appointments via an appointment book. For example, according to data collected from a primary care practice in Murray et al. (2007), a panel of 1806 patients triggers on average a demand for 24 appointments per day. The patients’ decision whether to join the backlog depends on its state. Let $A_a(t)$ be the cumulative number of patients that have booked an appointment in $[0, t]$, $A_b(t)$ be the cumulative number of balking patients in $[0, t]$, and $W_a(t)$ be the workload (or equivalently the offered waiting time) of the appointment book at time $t$. No-shows are treated as follows. A patient who does not show up in the clinic occupies a position in the appointment book until the
time of her scheduled appointment, and the backlog dynamics are unaffected (that is, we assume that no-show patients do not notify the clinic in advance).

Each day, depending on the random evolution of the appointment book, a number of patients is scheduled to arrive at the clinic. In contrast with the appointment book, the clinic starts empty at the beginning of each period. The length of a regular working day is $T$ hours, during which the scheduled appointments are allocated. The server continues to work overtime as well, beyond $T$, until the queue empties. Let $A_c(t)$ be the cumulative number of actual clinic arrivals in $[0,t]$, not including no-shows, and $W_c(t)$ be the workload of the clinic at time $t$.

Figure 1  Schematic representation of the model.

![Schematic representation of the model]

The objective function has five components. There is a reward $r$ generated per patient served. For the appointment book we consider a holding cost $c_a$ per time unit that each patient who joins the backlog has to wait for her scheduled appointment, and a balking cost $c_b$ per patient who encounters a long backlog and decides to not book an appointment. There are two types of costs associated with the in-clinic queue: a waiting cost $c_w$ per time unit that each patient has to wait to see the physician, and an overtime cost $c_o$ per time unit that the clinic has to operate overtime.

We are interested in optimizing the long run average daily net benefit from providing care to patients with respect to the panel size, $N$, and the number of offered appointment slots per working day, $s$. We consider static policies, where the appointment availability $s$ is the same for every day. In the concluding section we discuss possible extensions in a dynamic setting. Under the above notation, the optimization problem under study is

$$
\max_{N,s} \lim_{t \to \infty} \frac{1}{t} \mathbb{E} \left[ rA_c(t) - c_b A_b(t) - c_a \int_0^t W_a(\tau^-)dA_a(\tau) - c_w \int_0^t W_c(\tau^-)dA_c(\tau) - c_o \int_0^t 1\{W_c(\tau) > 0, \tau \text{ overtime}\}d\tau \right]. \quad (1)
$$

We would like to clarify at this point that the control of the daily appointment availability $s$ does not affect the duration of the in-clinic service times, nor the regular daily capacity to serve patients. A large value of $s$ corresponds to a model where a clinic potentially overbooks capacity.

In §4 and §5 we analyze the two queueing systems associated with the two time scales involved in accessing medical care and, based on diffusion approximations, we provide a tractable expression for the objective function in (1).
4. The Appointment Book

First, we are interested in characterizing the evolution of the appointment backlog as a function of the panel size and the number of offered slots per working day, disregarding the clinic intra-day dynamics. In order to obtain a tractable expression for the appointment delay, we develop a heavy traffic diffusion approximation.

Requests for appointments arrive at a rate $\lambda$ per day, with $\lambda$ being an increasing function of the panel size. For instance, as pointed out in Robinson and Chen (2010), if the panel size of a clinic is $N$, and each patient independently requests an appointment for any given period with a small probability $q$, then the arrival rate of appointment requests is $\lambda = Nq$ per day, and the number of patients requesting appointments will follow the binomial distribution with parameters $N$ and $q$, which converges to a Poisson distribution with mean $Nq$ as $N \to \infty$. Green et al. (2007) suggest how to calculate the daily visit rate per patient $q$ based on the characteristics of a medical practice and past data, and provide an example of $q = 0.008$ for a general/family practitioner.

The supply of medical appointments is $s$ slots per day. We assume that the offered appointments are allocated evenly throughout the working day and we consider that one time unit corresponds to one working day. For example, the offered time slots for day 1 is the discrete set of points in time $\{0, \frac{1}{s}, \frac{2}{s}, \ldots, \frac{s-1}{s}\}$. Each patient who books an appointment occupies one time slot in the queue until the time of her scheduled appointment.

We model the appointment book as a $GI/D/1$ queue with state dependent balking. The single server represents the single appointment book, which deterministically offers appointment slots at a rate $s$ slots per day. Patients’ decision of whether to book an appointment or not depends on the state of the appointment book: the less congested the latter, the more likely it is that a patient will have access to a slot of her preference. In order to quantify the waiting time for an appointment, we assume that the service discipline of the appointment book is FIFO. That is, appointments are booked in the same order of the arrivals of the requests.

In our analysis we adopt the modeling framework of Ward and Glynn (2005), where they develop a heavy traffic diffusion limit for a $GI/GI/1$ queue with either balking or reneging. Let $\{u_i : i \geq 1\}$ and $\{w_i : i \geq 1\}$ be two independent sequences of mean 1 iid random variables defined on a common probability space $(\Omega, \mathcal{F}, P)$, with $u_i \sim G$ and $w_i \sim F$ for all $i \geq 1$. For a given arrival rate $\lambda > 0$, let $\lambda^{-1}u_i$ be the inter-arrival time between the $i-1^{th}$ and $i^{th}$ appointment requests. For a given average patience $m > 0$, let $mw_i$ be the “balking threshold” of patient $i$: patient $i$ will join the appointment book only if $mw_i$ exceeds the offered waiting time upon her arrival. The arrival times for appointment requests constitute a random walk $t_i := \sum_{j=1}^{i} \lambda^{-1}u_j$, with associated renewal (counting) process $A(t) = \max\{i \geq 0 : t_i \leq t\}$. Note that $A(t)$ represents the cumulative number of appointment requests in $[0,t]$, including the balking patients.
Let $Q(t)$ denote the appointment backlog at time $t$, which can be expressed recursively as

$$Q(t) = \sum_{i=1}^{A(t)} 1\{\frac{Q(t-i)}{s} < m w_i\} - D(t)$$

$$= \sum_{i=1}^{A(t)} 1\{\frac{Q(t-i)}{s} < m w_i\} - \lfloor st \rfloor + L(t), \quad (2)$$

where $D(t) := \sum_{i=0}^{\lfloor st \rfloor} 1\{Q(\frac{t}{s}) > 0\}$ is the cumulative number of scheduled appointments up to time $t$, and $L(t) := \lfloor st \rfloor - D(t)$ is the cumulative number of unassigned appointment slots up to time $t$.

### 4.1. Diffusion Limit

For our asymptotic framework we consider a sequence of systems indexed by $n$, in which the arrival rate is $\lambda_n$, the mean patience is $m_n$, the number of offered appointments is $s_n = s$ (constant), and the following heavy traffic and regularity conditions hold:

**Assumption 1.**

(a) $\sqrt{n}(\lambda_n - s) \to \eta \in \mathbb{R}$ as $n \to \infty$.

(b) $m_n s = n$.

(c) $\theta^2 := \text{Var}(u) < \infty$.

(d) $F$ is differentiable about 0 and $0 < \xi := F'(0) < \infty$.

Assumption 1 is consistent with Ward and Glynn (2005), and is quite typical in deriving conventional heavy traffic diffusion limits. Any process associated with system $n$ is subscripted by $n$. Then, we define the *fluid-scaled* process

$$\tilde{A}_n(t) := \frac{A_n(nt)}{n}, \quad (3)$$

and the *diffusion-scaled* processes

$$\tilde{A}_n(t) := \frac{A_n(nt) - n \lambda_n t}{\sqrt{n}},$$

$$\tilde{Q}_n(t) := \frac{Q_n(nt)}{\sqrt{n}},$$

and

$$\tilde{L}_n(t) := \frac{L_n(nt)}{\sqrt{n}}. \quad (4)$$

In preparation for our results, we find it necessary to state the following technicalities. Let $\mathcal{D}$ be the space of real valued functions on $[0, \infty)$ that are right continuous with left limits (RCCL), endowed with the Skorokhod $J_1$ topology, and let $\mathcal{B}$ be the Borel field of $\mathcal{D}$. For a complete definition of the $J_1$ metric on $\mathcal{D}$ see Whitt (2002), Chapter 3. All stochastic processes are measurable functions from a probability space $(\Omega, \mathcal{F}, P)$ into $(\mathcal{D}, \mathcal{B})$. We denote *almost sure convergence* as $X_n \to X$ and *weak convergence* as $X_n \Rightarrow X$. Finally, the expressions "$\approx$" and "$\approx$" denote equality and approximate equality in distribution respectively.
Let $Z = \{Z(t) : t \geq 0\}$ be the reflected diffusion process with state space $[0, \infty)$ that satisfies the stochastic integral equation

$$Z(t) = Z(0) + \alpha t - \gamma \int_0^t Z(\tau) d\tau + \sigma B(t) + I(t),$$

(5)

where $B(t)$ is a standard Brownian motion, $\alpha$, $\gamma$ and $\sigma$ are constants, and $\{I(t) : t \geq 0\}$ is the minimal non-decreasing process such that $Z(t) \geq 0$ $\forall$ $t \geq 0$, and $\int_0^\infty 1\{Z(t) > 0\} dI(t) = 0$. The process $Z$ is referred to as a regulated Ornstein-Uhlenbeck (ROU) process with infinitesimal drift $(\alpha - \gamma z)$, and infinitesimal variance $\sigma^2$. The pair $(Z(t), I(t))$ is uniquely determined as the image of the linearly generalized regulator mapping $(\Phi, \Psi)$, $Z(t) = (\Phi, \Psi)(\alpha t + \sigma B(t))$. The stochastic process $Z(t)$ has analytically tractable transient and steady state behavior. For more details about ROU processes and the linearly generalized regulator mapping $(\Phi, \Psi)$ the reader is referred to Ward and Glynn (2003, 2005).

**Theorem 1.** If $\tilde{Q}_n(0) \Rightarrow Z(0)$ as $n \to \infty$, then $\tilde{Q}_n(t) \Rightarrow Z(t)$ as $n \to \infty$, where $Z(t)$ is a regulated Ornstein-Uhlenbeck process with initial position $Z(0)$, infinitesimal drift $(\eta - s\xi z)$, and infinitesimal variance $s\theta^2$.

The heavy traffic diffusion limit of the appointment backlog is an ROU process, a mean-reverting process with a reflecting boundary at zero. In §4.3 we demonstrate the accuracy of an ROU process, with the proper drift and variance, in approximating the appointment book in steady state.

Let $Z(t)$ be an ROU process with infinitesimal drift $(\alpha - \gamma z)$ and infinitesimal variance $\sigma^2$. Then, as in Browne and Whitt (1995), $Z(t) \Rightarrow Z_\infty$ as $t \to \infty$, where $Z_\infty$ has the distribution of a $\text{Normal}(\frac{\alpha}{\gamma}, \frac{\sigma^2}{\gamma^2})$ restricted to the interval $[0, \infty)$, i.e. $Z_\infty$ has density

$$f_{Z_\infty}(x) = \sqrt{\frac{2}{\pi \sigma^2}} \phi\left(\sqrt{\frac{2}{\pi \sigma^2}}(x - \frac{\alpha}{\gamma})\right), x \geq 0,$$

(6)

where $\phi(y)$ and $\Phi(y)$ are the probability density function (pdf) and the cumulative density function (cdf) of a standard Normal random variable respectively. The mean of $Z_\infty$ is

$$E[Z_\infty] = \frac{\alpha}{\gamma} + \sqrt{\frac{\sigma^2}{\gamma^2}}h\left(-\alpha \sqrt{\frac{2}{\gamma \sigma^2}}\right),$$

(7)

where $h(y) = \frac{\phi(y)}{1-\Phi(y)}$ is the standard Normal hazard rate.

**4.2. Proposed Approximation**

We propose an approximation for the appointment backlog based on the diffusion limit in Theorem 1. The appointment backlog is approximated with an ROU $\tilde{Q}(t)$ having infinitesimal drift $\lambda - s$ -
\(\frac{\lambda_s}{ms}q\), and infinitesimal variance \(\lambda \theta^2\). To understand the logic behind the proposed approximation, note that for large \(n\) and from Theorem 1,

\[
Q_n(t) \approx \sqrt{n}Z(t_n),
\]

(8)

where, by Assumption 1,

\[
\sqrt{n}Z(t_n) = \sqrt{n}Z(0) + \sqrt{n}\eta t_n - \lambda n\xi \int_0^n \sqrt{n}Z(\tau) d\tau + \sqrt{n}B(t_n) + \sqrt{n}I(t_n).
\]

(9)

Therefore, from (8) and (9),

\[
Q(t) \approx Q(0) + (\lambda - s)t - \frac{\lambda s}{\tau n} \int_0^t Q(\tau) d\tau + \sqrt{\lambda}B(t) + L(t).
\]

(10)

Suppose that \(\lambda, s, m > 0\). In steady state, the backlog of appointments is approximated as the random variable \(\tilde{Q}_a\) with pdf

\[
f_{\tilde{Q}_a}(x) = \sqrt{\frac{2}{ms^2 \pi}} \left( \frac{2}{ms^2 \pi} \left( x - \frac{m(\lambda - s)}{\lambda} \right) \right) \phi \left( \frac{x - \frac{m(\lambda - s)}{\lambda}}{\lambda \sqrt{2ms^2 \pi}} \right), x \geq 0,
\]

(11)

and mean

\[
E[\tilde{Q}_a] = \frac{m(\lambda - s)}{\lambda} + \sqrt{\frac{2\pi s^2}{\lambda}} h \left( \frac{s}{\lambda} \sqrt{\frac{2\pi s^2}{\lambda}} \right).
\]

(12)

When either \(\lambda = 0, s = 0\), or \(m = 0\), then \(\tilde{Q}_a = 0\) with probability one.

The following lemma reveals some properties of the expected appointment backlog as a function of the various parameters, which are useful for optimization purposes in §6.

**Lemma 1.** (a) \(E[\tilde{Q}_a]\) is strictly increasing in \(\lambda\) on \((0, \infty)\) \(\forall s, m, \xi > 0\).

(b) \(E[\tilde{Q}_a]\) is unimodal in \(s\) on \((0, \infty)\) \(\forall \lambda, m, \xi > 0\).

(c) \(E[\tilde{Q}_a]\) is strictly increasing in \(m\) on \((0, \infty)\) \(\forall \lambda, s, \xi > 0\).

(d) \(E[\tilde{Q}_a]\) is strictly decreasing in \(\xi\) on \((0, \infty)\) \(\forall \lambda, s, m > 0\).

(e) \(E[\tilde{Q}_a] \to 0\) as \(s \to \infty\) \(\forall \lambda, m, \xi > 0\).

(f) \(E[\tilde{Q}_a]\) is continuously differentiable in \(\lambda\) at \(\lambda = 0\) \(\forall s, m, \xi > 0\).

Aligned with our intuition, the average appointment backlog is increasing in the daily demand for medical appointments, \(\lambda\). Interestingly, \(E[\tilde{Q}_a]\) is not necessarily monotone in the number of offered appointments \(s\). One would expect the appointment backlog to be decreasing in \(s\), the daily rate at which the clinic offers appointment slots. But note that on the other hand, the larger the
value of \( s \), the more likely it is that a patient will not balk and actually join the appointment book. Figure 2 demonstrates the approximated expected appointment backlog \( E[\tilde{Q}_a] \) as a function of \( \lambda \) and \( s \), for a wide range of traffic intensities and for different average patience levels \( m \). Besides the monotonicity properties from Lemma 1, we also observe that \( E[\tilde{Q}_a] \) is neither convex nor concave in either \( \lambda \) or \( s \).

**Figure 2** \( E[\tilde{Q}_a] \) is increasing in \( \lambda \) and \( m \), and unimodal in \( s \).

(a) \( 0 \leq \lambda \leq 40, \ s = 20 \).

(b) \( \lambda = 20, \ 0 \leq s \leq 40 \).

\[ E[\tilde{Q}_a] \]

\[ \text{Note.} \ The \ arrival \ process \ is \ Poisson, \ and \ the \ balking \ threshold \ is \ uniformly \ distributed \ with \ mean \ m. \]

Based on our approximated pdf, we can further approximate the probability of balking as

\[
\Pr(\text{Balking}) := \int_{0}^{\infty} \Pr(\text{Balking} | \tilde{Q}_a = x) f_{\tilde{Q}_a}(x) dx = \int_{0}^{\infty} F\left( \frac{x}{ms} \right) f_{\tilde{Q}_a}(x) dx, \tag{13}
\]

and the effective booking rate as

\[
\lambda_{\text{book}} := \lambda [1 - \Pr(\text{Balking})]. \tag{14}
\]

Little’s Law provides an approximation for the expected (effective) appointment delay, \( E[\tilde{W}_a] := \frac{E[\tilde{Q}_a]}{\lambda_{\text{book}}} \). Finally, another metric of interest is the ability of the medical practice to offer same-day appointments (within 24-hours). This metric is defined as the probability that an arriving request for appointment sees a queue of less than \( s \) patients, and therefore it is approximated as

\[
\Pr(\text{Same-day appointment}) := \Pr(\tilde{Q}_a < s). \tag{15}
\]

Our next lemma provides a lower bound on that probability, a key performance measure of interest for Open Access scheduling.
Lemma 2. Suppose that \(0 \leq \lambda \leq s\). Then

(a) \(\Pr(\text{Same-day appointment}) \geq 2\Phi\left(\sqrt{\frac{2\xi s}{m^2}}\right) - 1\).

(b) \(\lim_{s \to \infty} \Pr(\text{Same-day appointment}) = 1\).

4.3. Simulation Experiments

While diffusion models are typically less accurate than simulation, they have the significant advantages of speed and analytical tractability. In order to assess the accuracy of the diffusion approximation, we compare a simulated backlog of appointments with the approximated one. For our simulation experiments we assume Poisson arrivals. For the balking threshold we consider the following three distributions: Uniform, Truncated Normal and Exponential.

Our diffusion approximation for the appointment backlog depends on the density of the balking threshold at zero, disregarding the rest of the distribution, as it is the case in Ward and Glynn (2005) (as well as in other studies such as Zeltyn and Mandelbaum (2005), Dai and He (2010) and Mandelbaum and Momcilovic (2012)). In contrast, Reed and Ward (2008) and Reed and Tezcan (2012) obtain via a more refined scaling heavy traffic diffusion limits for single server and multiserver queues with reneging, respectively, in which the entire patience distribution plays a role through its hazard rate.

In our first experiment we compare the average queue length of a simulated system with the one provided by \(E[\tilde{Q}_a]\) in (12), for a wide range of traffic intensities and different balking threshold distributions. As shown in Figure 3, the diffusion approximation is very accurate for all traffic intensities when the balking threshold is Uniform. Figure 4 demonstrates that in fact the whole distribution of the appointment backlog is very accurately approximated by (11) for Uniform balking threshold, when compared to the relative frequency histogram of a simulated appointment book.

In the case of a Uniform balking threshold, the density is constant on the whole support of the distribution. We examine other distributions as well. We also consider a Truncated Normal, i.e.,

\[ F(x) = \Pr(\text{Normal}(\mu, \sigma) \leq x|\text{Normal}(\mu, \sigma) \geq 0) \]

with density

\[ F'(x) = \frac{1}{\sigma} \phi\left(\frac{x - \mu}{\sigma}\right) \Phi\left(\frac{\mu}{\sigma}\right), \]

for some \(\mu\) and \(\sigma\) such that \(E(w) = 1\) (recall that the model primitives have mean 1 and are scaled with the average patience \(m\)) and for different values of CoV := \(\sqrt{\text{Var}(w)}\). Regardless of the value of CoV, as shown in Figure 3, the approximation is quite accurate in both heavy and light traffic (when \(\frac{\lambda}{s} < 105\%\)). In an overloaded system (\(\frac{\lambda}{s} > 105\%\)) the accuracy of the approximation
Figure 3  Average Queue Length

Note. Requests for appointment are Poisson with rate $\lambda = 30$ per day.
Figure 4  Queue Length Distribution

Note. Requests for appointment are Poisson and balking threshold is Uniform.
highly depends on CoV. As demonstrated in Figure 5, depending on CoV, the density at the origin $F'(0)$ takes a wide range of different values and the whole distribution has various shapes. Moreover, from Lemma 1, $E[\tilde{Q}_a]$ is decreasing in $F'(0)$, while the simulated system is not quite sensitive to $F'(0)$. In an overloaded system, our approximation is accurate for $\text{CoV} \approx 0.7$, while it overestimates (underestimates) the expected queue length when $\text{CoV} < 0.7$ ($\text{CoV} > 0.7$). When the balking threshold is Exponential, i.e., when $F(x) = 1 - e^{-x}$, the density at zero is equal to 1. The approximation is accurate in both heavy and light traffic (when $\frac{1}{s} < 105\%$), while in an overloaded system ($\frac{1}{s} > 105\%$) it consistently underestimates the expected backlog. For the rest of the paper we consider in our numerical studies a Uniform balking threshold, where the density is constant on the whole support of the distribution, and the proposed approximation is more accurate. Nevertheless, as demonstrated in Section §6, in optimality our regime of interest is light/heavy traffic, where all three distributions approximate the expected backlog well.

Figure 5  Truncated Normal distribution with mean 1 for different values of CoV.

Another important metric, beyond the average backlog of appointments, is the ability of a medical practice to offer same-day appointments. Figure 6 illustrates the probability of a patient having access to a same-day appointment (within 24 hours) as a function of the appointment availability $s$, and the corresponding lower bound from Lemma 2 for $s \geq \lambda$. For small values of $s$ the backlog becomes unmanageable and this probability is close to 0. As $s$ approaches the daily demand rate $\lambda$ and beyond, the same-day appointment probability increases steeply to the point where almost all patients are offered a same-day appointment.

We are further interested in the booking rate $\lambda_{\text{book}}$, the effective rate at which patients book appointments daily. In Figure 7 we compare $\lambda_{\text{book}}$ from (14) with the actual daily booking rate of a simulated appointment book. The corresponding balking probabilities are displayed in Figure
Figure 6  
Same Day Appointment Probability

Note. Requests for appointment are Poisson and balking threshold is Uniform.

Figure 7  
Effective Arrival Rate

Note. Requests for appointment are Poisson and balking threshold is Uniform.

Figure 8  
Balking Probability

Note. Requests for appointment are Poisson and balking threshold is Uniform.

8. Our results suggest that $\lambda_{\text{book}}$ approximates the booking rate well, and for moderately large patience,

$$
\lambda_{\text{book}} \approx \min(\lambda, s).
$$

(16)
Figure 9  Distribution of Daily Appointment Book Output

\textbf{Note.} Requests for appointment are Poisson and balking threshold is Uniform.
Finally, we focus on the steady state behavior of the output process $D(t)$ from the appointment book. In particular, we are concerned with the steady state distribution of the daily number of scheduled appointments

$$\lim_{t \to \infty} [D(t+1) - D(t)].$$

(17)

In general, the task of characterizing the output process of non-markovian queues is difficult. We approximate the distribution of the random output in (17) based on the work of Pack (1975), by substituting their arrival rate with our effective arrival rate, as it appears in (14). Figure 9 demonstrates the accuracy of the proposed approximation compared to the relative frequency histogram of the daily output from a simulated appointment book. We also observe that

$$\Pr(\lim_{t \to \infty} [D(t+1) - D(t)] = s) \approx 1$$

in an overloaded system where demand exceeds supply ($\lambda > s$).

5. The In-Clinic Queue

In this section we study the in-clinic queue and derive closed-form expressions for the expected patients’ waiting time and clinic overtime based on a diffusion approximation. The evolution of the queue depends on how the patients are scheduled to arrive throughout the working day, on patient’s no-show behavior and punctuality, and on the service times distribution. As discussed in §2, there are no closed-form expressions for the patients’ waits and physicians’ overtime in the literature. Recursive expressions can be derived only under certain assumptions on the service times distribution (exponential or deterministic) and by assuming that patients who show up are punctual, see for example Hassin and Mendel (2008), Robinson and Chen (2010), Zacharias and Pinedo (2014, 2015). Typically transient analysis of rich queueing systems is addressed either via simulation or approximations, Bandi and Bertsimas (2012).

Honnappa et al. (2015) also argue that transient queues (e.g. outpatient care) are very difficult to analyze via classical queueing techniques. By assuming that the scheduled appointments are evenly spread throughout the working day, and that all the patients arrive on their scheduled time plus a uniformly distributed noise, they show that the diffusion limit (as the number of scheduled patients goes to infinity) of the queue length process is a function of a Brownian motion (BM) and a Brownian Bridge. Araman and Glynn (2012) demonstrate that a Fractional Brownian motion is the corresponding diffusion limit when the noise is heavy-tailed. For the sake of tractability, and in order to account for no-shows as well, we assume that the arrival process is Renewal, in which case the resulting diffusion limit (see Iglehart and Whitt (1970)) is a Regulated Brownian motion (RBM). Patients’ waiting time is approximated in terms of the area under the RBM, and clinic overtime as the position of the RBM at the end of the working day. A comparison between our approximation and a system with non-renewal arrivals is provided in §5.2 via simulation.
5.1. Diffusion Approximation

We model the outpatient clinic as a $GI/GI/1$ queue. The length of a regular working day is $T$ hours, during which the offered appointments are allocated. The queue starts empty at the beginning of the working day and the server continues to work during overtime as well, beyond $T$, until the queue empties.

Patients show up for their scheduled appointment with probability $p \in (0, 1]$. On a particular day, if $x$ patients are scheduled to see a physician, then the arrival rate to the clinic is $\lambda_c = \frac{px}{T}$ patients per hour. We would like to emphasize that $x$ is a realization of the random daily output from the appointment book in (17). The analysis in this section is based on a fixed realization $x$, whereas $\lambda_c$ is treated as a random variable in §6.

For the arrival process we consider a sequence of iid random variables $\{u_{c,i} : i \geq 0\}$, with associated renewal processes $N_c(t) = \max\{k \geq 0 : \sum_{i=1}^{k} u_{c,i} \leq t\}$. The random variable $u_{c,i}$ denotes the inter-arrival time between the $i-1$th and $i$th patients, has finite mean $\lambda_c^{-1}$ and finite squared coefficient of variation $c^2_{u}$. By convention, $u_{c,0} = 0$.

For the service times we consider an independent sequence of iid random variables $\{v_i : i \geq 0\}$, where $v_i$ corresponds to the service time of the $i$th arrival, has finite mean $\mu^{-1}$ and finite squared coefficient of variation $c^2_{v}$.

The workload right before the $i$th arrival, denoted as $W_i$, can be expressed by the classical Lindley recursion as

$$W_i = \max\{W_{i-1} + v_{i-1} - u_{c,i}, 0\}, \quad \text{for } i \geq 2,$$

and $W_1 = 0$.

As in Chen and Yao (2001) Chapter 6, the workload at time $t$, denoted as $W(t)$, can be approximated as $W(t) \approx \frac{1}{p} Y(t)$, where $Y(t)$ is an RBM with initial position $Y(0) = 0$, drift

$$\alpha := \lambda_c - \mu,$$

and infinitesimal variance

$$\beta^2 := \lambda_c c^2_u + \min(\lambda_c, \mu) c^2_v,$$

reflected at zero. One representation of $Y(t)$ is

$$Y(t) = X(t) + U(t),$$

where $U(t) = \sup_{0 \leq s \leq t} \{-X(t)\}$ and $X(t) = \alpha t + \beta B(t)$,
Lemma 4. The first two moments of $t$ (2001) or p. 184 of Whitt (2002), when the drift is positive and for large drift. As argued in the literature on reflected diffusions (see for example p. 144 of Chen and Yao through Laplace transforms. The same methodology does not go through in the case of positive drift.

Assume that

$$E[Y(t)] = \begin{cases} \frac{\alpha T}{\sqrt{\beta}} & \text{if } \beta \neq 0, \alpha \leq 0 \\ 0 & \text{if } \beta \neq 0, \alpha > 0 \end{cases}$$

$$E[Y^2(t)] = \begin{cases} 0 & \text{if } \beta \neq 0, \alpha \leq 0 \\ \frac{\alpha^2 T^2}{\beta^2} + \frac{\beta^2}{\alpha^2} \Phi \left( \frac{\alpha T}{\beta} \right) + \frac{\beta^2}{\alpha^2} \Phi \left( \frac{\alpha T}{\beta} \right) - \frac{\beta^2}{\alpha^2} & \text{if } \beta > 0, \alpha \neq 0 \end{cases}$$

The aggregate in-clinic waiting time in the $GI/GI/1$ queue for the time interval $[0, T]$ is $\sum_{i=1}^{N_c(T)} W_i$, and can be approximated in terms of the area under an RBM as

$$\sum_{i=1}^{N_c(T)} W_i = \int_0^T W(t^-)dN_c(t)$$

$$= \int_0^T nW ((nt)^-) \frac{dN_c(nt)}{n}$$

$$\approx \int_0^T nW ((nt)^-) d\lambda_c t \quad \text{(for large } n \text{ and from FSLLN)}$$

$$= \lambda_c \int_0^T W(t^-) dt$$

$$\approx \frac{\alpha n}{\mu} \int_0^T Y(t^-) dt =: \tilde{W}_c. \quad (23)$$

The physician’s overtime is equal to $W(T)$, the workload at the end of the regular working day, and is approximated as

$$\tilde{O}_c := \frac{1}{\mu} Y(T). \quad (24)$$

In what follows, we derive closed-form expressions for $E[\tilde{W}_c]$ and $E[\tilde{O}_c]$.

Lemma 3. Assume that $\beta^2 > 0$. Then $E[\int_0^T Y(t) dt] = \begin{cases} \frac{E[Y^2(T)] - \beta^2 T}{2\alpha} & \text{if } \alpha \neq 0 \\ \frac{E[Y^2(T)]}{3\beta^2} & \text{if } \alpha = 0. \end{cases}$

The first two moments of an RBM with negative drift are derived in Abate and Whitt (1987) through Laplace transforms. The same methodology does not go through in the case of positive drift. As argued in the literature on reflected diffusions (see for example p. 144 of Chen and Yao (2001) or p. 184 of Whitt (2002)), when the drift is positive and for large $t$, the effect of reflection becomes negligible, and therefore an RBM is approximated with a BM. However, since clinics typically offer appointments over a finite time interval $[0, T]$ in the order of 8-12 hours, we need to include the reflection in our analysis, even for the case of positive drift ($\lambda_c > \mu$).

Lemma 4. The first two moments of $Y(t)$ are as follows:

$$E[Y(t)] = \begin{cases} 0 & \text{if } \beta = 0, \alpha \leq 0 \\ \frac{\alpha T}{\sqrt{\beta}} & \text{if } \beta = 0, \alpha > 0 \end{cases}$$

$$E[Y^2(t)] = \begin{cases} 0 & \text{if } \beta = 0, \alpha \leq 0 \\ \frac{\alpha^2 T^2}{\beta^2} + \frac{\beta^2}{\alpha^2} \Phi \left( \frac{\alpha T}{\beta} \right) + \frac{\beta^2}{\alpha^2} \Phi \left( \frac{\alpha T}{\beta} \right) - \frac{\beta^2}{\alpha^2} & \text{if } \beta > 0, \alpha \neq 0 \end{cases}$$
Lemma 5. \( E[Y(t)] \) and \( E[Y^2(t)] \) are continuous in \( \alpha \).

We note that our proof of Lemma 4 provides an alternative proof for Theorem 1.1 of Abate and Whitt (1987), for the special case where \( \alpha = -1 \) and \( \beta = 1 \). Now we can express the performance measures of interest as follows.

Theorem 2.

\[
E[\tilde{O}_c] = \begin{cases} 
0 & \text{if } \beta = 0, \alpha \leq 0 \\
\frac{\alpha T}{\beta} \phi\left(\frac{\alpha \sqrt{T}}{\beta}\right) + (\alpha^2 T + \frac{\beta^2}{\alpha}) \Phi\left(\frac{\alpha \sqrt{T}}{\beta}\right) - \frac{\beta^2}{\alpha} & \text{if } \beta > 0, \alpha \neq 0 \\
\frac{\alpha}{\beta} \sqrt{T} \phi\left(\frac{\alpha \sqrt{T}}{\beta}\right) + (2\beta^2 \alpha^2 T + \alpha^4 T^2 - \beta^4) \Phi\left(\frac{\alpha \sqrt{T}}{\beta}\right) + \frac{\beta^2}{2} - \beta^2 \alpha^2 T & \text{if } \beta > 0, \alpha = 0.
\end{cases}
\]

Lemma 6. \( E[\tilde{W}_c] \) and \( E[\tilde{O}_c] \) are continuous in \( \alpha \).

The RBM approximation for the in-clinic queue in transient state is tractable, and provides closed-form expressions for the performance measures of interest. It also captures variability in the arrival process (for positive \( c_u^2 \)), and random service times (for positive \( c_v^2 \)).

5.2. Simulation Experiments

Empirical studies suggest that the service times for certain medical practices have a Lognormal distribution. Cayirli et al. (2006) analyze data collected from a primary health care clinic in a New York metropolitan hospital that provides service to about 300,000 outpatients a year. They find that a lognormal distribution with mean \( \mu = 15.5 \) minutes and a coefficient of variation \( c_v = 0.325 \) is the best fit for the service times of return patients.

The Brownian approximation turns out to be very accurate for a Poisson arrival process, under both light and heavy traffic. In Figure 10 we compare the expressions for \( E[\tilde{W}_c] \) and \( E[\tilde{O}_c] \) in Theorem 2 with the corresponding performance measures of a simulated clinic for different traffic intensities and for different levels of service time variability. For low values of \( c_v \) (less than 0.5) our approximation performs very well, and, as variability increases, we tend to slightly overestimate the aggregate waiting time and physician’s overtime.

We are also interested in evaluating our approximation when the arrival process is not renewal. In our next experiment we adjust the setting of Honnappa et al. (2015) to account for no-shows as well. The scheduled appointments are evenly spread throughout the working day, and each patient shows up with probability \( p \). Those patients who show up arrive on their scheduled time plus a uniformly distributed noise \( \epsilon \sim \text{Unif}[-\delta, \delta] \). In Figure 11 we illustrate the results of a simulation
Figure 10  Aggregate waiting time and overtime: a comparison between the diffusion approximation and a simulated clinic.

(a) \( \mu \in \{2, 3, 4\}, 0 \leq \lambda_c \leq 3, c_v = 0.325 \).

(b) \( \mu \in \{2, 3, 4\}, 0 \leq \lambda_c \leq 3, c_v = 0.325 \).

(c) \( \mu = 4, \lambda_c \in \{3, 4, 5\}, 0 \leq c_v \leq 1 \).

(d) \( \mu = 4, \lambda_c \in \{3, 4, 5\}, 0 \leq c_v \leq 1 \).

Note. The arrival process is Poisson, the service times are Lognormal with mean \( \mu^{-1} \) hours and standard deviation \( c_v \times \mu^{-1} \), \( T = 8 \) hours.

experiment. When \( \delta = 0 \) (in which case the inter-arrival times follow a Geometric distribution and the arrival process is, in fact, renewal) we very accurately approximate the aggregate waiting time both in light and heavy traffic. For positive values of \( \delta \) (where the arrival process is no-longer renewal) the waiting times are overestimated.

6. Optimal Panel Sizing and Appointment Scheduling

Having developed the necessary tools to approximate the appointment backlog and the in-clinic queue, we aim to maximize the long run average daily net benefit of the medical facility from providing care to patients with respect to the panel size, \( N \), and the number of offered appointment slots per working day, \( s \). We assume that \( \lambda \) is strictly increasing in \( N \) (Green et al. (2007) suggest
Figure 11  Clinic delay for non-renewal arrival process: a comparison between the diffusion approximation and a simulated clinic.

(a) $\delta = 0$ minutes

(b) $\delta = 10$ minutes

(c) $\delta = 20$ minutes

Note. No-show rate is 0.1. Patients who show up are unpunctual: they arrive on their scheduled time plus a uniformly distributed noise $\epsilon \sim \text{Unif} [-\delta, \delta]$. The service times are Lognormal with mean $\mu^{-1} = 20$ minutes and coefficient of variation $c_v = 0.325$, $T = 8$ hours. The number of scheduled patients ranges between 1 and 40, and the corresponding appointment intervals range between 8 hours and 12 minutes respectively. The approximated aggregate clinic waiting time is computed as in Theorem 2 with the first two moments of the inter-arrival times being empirically estimated from simulated arrivals.

As a reminder, we consider a reward $r > 0$ generated per patient served, a balking cost $c_b > 0$ per patient who encounters a long backlog and decides to not book an appointment, a holding cost $c_a > 0$ for each day that each patient who joins the backlog has to wait for her scheduled appointment, a waiting cost $c_w > 0$ per hour that each patient has to wait in the clinic to see the physician, and an overtime cost $c_o > 0$ per hour beyond the end of the regular working day $T$.

Recall that the analysis of the in clinic queue in §5 is based on a fixed arrival rate for a given day. For an arbitrary day, $\lambda_c$ is a random variable governed by the distribution of the appointment book’s daily output in (17), and depends on the no-show probability.

We approximate the average daily net benefit of the medical facility from providing care to patients in (1) as

$$R(\lambda, s) := rp\lambda_{\text{book}} - c_b(\lambda - \lambda_{\text{book}}) - c_a\lambda_{\text{book}}E[\tilde{W}_a] - c_wE_{\lambda_c}[E[\tilde{W}_c|\lambda_c]] - c_oE_{\lambda_c}[E[\tilde{O}_c|\lambda_c]]$$

$$= rp\lambda_{\text{book}} - c_b(\lambda - \lambda_{\text{book}}) - c_aE[\tilde{Q}_a] - c_wE_{\lambda_c}[E[\tilde{W}_c|\lambda_c]] - c_oE_{\lambda_c}[E[\tilde{O}_c|\lambda_c]],$$

(25)
where $E[\tilde{Q}_a]$ is as in (12), and $E[\tilde{W}_c|\lambda_c]$ and $E[\tilde{O}_c|\lambda_c]$ are as in Theorem 2 with $\alpha = \lambda_c - \mu$ and $\beta^2 = \lambda_c c_u^2 + \min(\lambda_c, \mu)c_v^2$. We consider the optimization problem

$$
\max_{\lambda, s} R(\lambda, s)
\text{s.t. } \lambda, s \leq M \tag{P}
$$

where $M$ is a constant, and the constraint $\lambda, s \leq M$ ensures that demand and supply of appointments cannot be arbitrarily large (finite demand population and clinic capacity). Lemmas 1(f) and 6, and Weierstrass’ extreme value theorem guarantee the existence of an optimal solution to (P).

### 6.1. Characterization of the Optimal Solution

The optimization problem (P) is analytically intractable; the objective function involves the pdf, cdf, and hazard rate of a standard Normal random variable. However, if we make one simplifying assumption, motivated by (16) and for the sake of tractability, we are able to provide a neat characterization of the optimal solution.

**Theorem 3.** Assume that $\lambda_c = \frac{\min(\lambda, s)}{T}$ with probability one, and let $(\lambda^*, s^*)$ be an optimal solution to (P). Then $\lambda^* \leq s^*$ and one of the following holds:

(a) $\lambda^* = 0$.

(b) $0 < \lambda^* \leq s^* = M$.

(c) $0 < \lambda^* = s^* \leq \frac{\xi \theta^2 \pi}{4m(\pi - 2)^2}$.

Under the assumption of Theorem 3, supply for medical appointments should be at least as high as the demand in outpatient care, and furthermore, (P) reduces to a single variable optimization problem. The optimal solution to the panel sizing and scheduling problem lies within one of the following three regimes:

(a) The objective coefficients are such that it is not cost effective to operate the clinic (under any configuration of $\lambda$ and $s$).

(b) The clinic offers as many appointment slots as possible, and the optimal panel size depends on the clinic’s characteristics.

(c) Supply and demand are perfectly matched and are both very small.

We point out that regime (c) does appear in theory as a separate strategy (following from the unimodality of the appointment backlog in $s$), but in practice it is nearly identical to regime (a). For all practical configurations, the denominator of the fraction $\frac{\xi \theta^2 \pi}{4m(\pi - 2)^2}$ is larger than the numerator. For instance, in the extreme case that the balking threshold has an average of only 2 days (uniformly distributed) and that requests for appointment arrive according to a Poisson
process, then $\lambda^* = s^* < \frac{\theta^2 - \pi}{4m(\pi - 2)} \approx 0.15$, which practically results to a solution close to regime (a). On that account, Theorem 3 suggests that a positive optimal solution lies within regime (b), which we refer to as the Open Access policy. According to Lemma 2, if we consider that the balking threshold is uniformly distributed with an average of 10 days, $M = 30$, and that requests for appointment arrive according to a Poisson process, then $\Pr(\text{Same-day appointment}) \geq 0.92$.

In order to characterize the optimal solution in Theorem 3 we assumed that the arrival rate to the clinic from scheduled appointments is equal to the “fluid” approximation $\rho \times \min(\lambda, s)$ patients per day. We drop this assumption for the rest of the analysis, and $\lambda_c$ is treated as a random variable governed by the distribution of the daily output from the appointment book in (17). In our extensive numerical experiments the optimal pair $(\lambda^*, s^*)$ almost always lies within the Open Access regime. As illustrated in §6.2, a solution arises outside the Open Access regime only when appointment delays are considerably less costly than clinic delays and the revenue generated per patient served, i.e., when $c_a, c_b < c_w, c_o, r$.

We conclude this section with numerical experiments regarding the behavior of the objective function $R(\lambda, s)$ in (25) and the optimal solution. In our analysis the objective is normalized with respect to $c_w$, i.e., $c_w = 1$. As in Robinson and Chen (2010), we consider an overtime cost coefficient which is fifteen times as much as the patients’ waiting cost coefficient, i.e., $c_o = 15$. Meaningful values for $r$ are those such that the average reward generated per hour is of the same order as the overtime cost per hour, i.e., $r \mu \sim c_o$. Since this is the first study to explicitly capture both types of delay, we are not aware of any guidelines in the literature regarding the right ratio $c_a/c_w$, which we expect to be context dependent. First, in order to capture the impact of both time scales, we consider that the cost of waiting one day for an appointment is of the same order as the cost of waiting one hour in the clinic, i.e., $c_a \sim c_w$. In our simulation study we also investigate the role of the ratio $c_a/c_w$ for values between 0 (i.e., appointment delay is not important at all) and 2. Finally, the balking cost per patient is of the same order as the cost of a patient waiting for $m$ days, i.e., $c_b \sim mc_a$.

As shown in Figure 12(a), the objective function appears to be concave in $\lambda$ on $[0, s]$, though its individual components are not necessarily concave (for example, $-E(Q_a)$ is not, as discussed in §4.2). Figure 12(b) demonstrates the optimal solution as given by the MATLAB R2015a optimization toolbox. The no-show probability has a significant effect on the optimal solution and should be taken under consideration. Further, as $r$ increases with values greater than $c_a \mu^{-1}$, the optimal arrival rate $\lambda^*$ increases concavely towards heavy traffic, i.e., $\lambda^* \rightarrow s^*$. For reasonable values of $r$ (around $c_a \mu^{-1} = 5$), the system operates in light traffic, and hence a policy of “satisfying today’s demand today” (Open Access) with high probability is optimal, see Figure 14.
Figure 12  Approximated objective function and optimal solution.

(a) Objective as a function of $\lambda$; $p = 0.9$, $s^* = 30$.

(b) $\lambda^*$ as a function of $r$; $s^* = 30$, $p \in \{0.7, 0.8, 0.9, 1.0\}$.

Note. The balking threshold is uniformly distributed with mean $m = 10$ days, the arrivals to the appointment book are Poisson, $T = 8$ hours, $\frac{1}{m} c_b = c_a = c_w = \frac{1}{15} c_o = 1$, $\mu = 3$ patients per hour (24 patients per day), $c_v = 0.325$, $c_a = 1$, $M = 30$.

Figure 13  Decomposed objective function.

(a) Performance measures.

(b) Weighted perf. meas., $\frac{1}{m} c_b = c_a = c_w = \frac{1}{15} c_o = \frac{1}{5} r = 1$.

Note. The balking threshold is uniformly distributed with mean $m = 10$ days, the arrivals to the appointment book are Poisson, $T = 8$ hours, $\mu = 3$ patients per hour, $\lambda = \mu T = 24$ requests per day, $c_v = 0.325$, $c_a = 1$, $p = 1$. 
Here is what drives the optimality of an Open Access policy. Both decision variables, $\lambda$ and $s$, balance the trade-offs between accumulating reward from serving patients and providing timely access to care (under both time scales). An increasing panel size comes along with higher rewards, and at the same time longer appointment delays and a more congested clinic. On the other hand, an increasing appointment availability $s$ results also in higher rewards and a more congested clinic, but potentially in shorter appointment delays. By optimizing over the two levers simultaneously, we can set a large $s$ that provides short appointment delays, and control the clinic throughput by setting a proper panel size. To make this point more clear, take for example a look at the decomposed objective function in Figure 13, where demand and regular capacity are perfectly matched, i.e., $\lambda = \mu T = 24$ patients per day, and only the control $s$ varies. In Figure 13(a) we observe that the daily appointment delay is decreasing more steeply around $\lambda$ than the clinic waits and clinic overtime increase: as $s$ approaches $\lambda = 24$ and beyond, the appointment backlog is transitioning from an overloaded system to heavy traffic and eventually light traffic, while the in-clinic queue remains steadily in heavy traffic. To explain the latter, note that $\lambda_{\text{book}}$ is not very sensitive to changes in $s$ for values beyond $\lambda$ (look for instance at Figure 7). When we scale the corresponding quantities with their weights in Figure 13(b), we observe that indeed, when $s$ increases around and beyond $\lambda = 24$, the benefits from providing timely access to appointments outweigh the costs from potentially overbooking capacity. As a result, $R(24, s)$ is maximized at the boundary $s^* = 30$. As demonstrated below, a solution outside the Open Access regime only arises when the weight of the appointment delay $c_a$ is an order of magnitude smaller than clinic delay weight $c_w$.

6.2. A Simulation Experiment

In this section we demonstrate via simulation the accuracy of the proposed approximation in solving the joint panel sizing and scheduling problem, and also the optimality of an Open Access policy. Our experimental setup is as follows:

- **Appointment book**: Requests for an appointment arrive according to a Poisson process at a rate $\lambda$ per day, and $s$ appointment slots are offered per day. There is an upper bound $M = 30$ for both $\lambda$ and $s$. The balking threshold is uniformly distributed with an average patience of $m = 10$ days.

- **Clinic**: The length of the working day is $T = 8$ hours. Depending on the random evolution of the appointment book, $s_k$ patients are scheduled to arrive throughout day $k$, $1 \leq k \leq 10000$. There is a no-show probability of $1 - p = 0.1$. We consider that the arrival process to the clinic is Poisson at a rate $ps_k/T$ patients per hour. Finally, service times are Lognormal with mean 20 minutes (i.e. regular capacity of serving 24 patients per day with $\mu = 3$ patients per hour) and coefficient of variation $c_v = 0.325$. 
Figure 14 Effect of the reward $r$.

Note. All the random variables are as described in §6.2. The cost coefficients are $\frac{1}{m} c_a = c_a = c_w = \frac{1}{15} c_o = 1$ and $r$ varies.

Figure 15 Effect of the ratio $c_a/c_w$.

Note. All the random variables are as described in §6.2. The objective coefficients are $c_w = \frac{1}{15} c_o = 1$, $r = \frac{1}{5} c_o = 5$ and $\frac{1}{m} c_b = c_a$ varies.

Our simulation experiments confirm that indeed an Open Access policy is optimal when we optimize over both $\lambda$ and $s$ simultaneously and when we take into account both types of delay, i.e., when $c_a \sim c_w$. In Figure 14 we illustrate the effect of the reward $r$ on the optimal solution. For all configurations, it is optimal to offer as many appointment slots as possible, $s^* = M = 30$. The optimal arrival rate $\lambda^*$ is increasing in $r$ and is such that the probability of a patient having access to a same day appointment is at least 97.5% and the balking probability is very small.

We are also interested to see how the optimal solution is affected by the ratio $c_a/c_w$, the ratio that captures the relative importance between appointment delay and clinic delay. As illustrated in Figure 15, when $c_a$ is close to zero we obtain a solution where $\lambda^* > s^*$; a medical practice can afford having a long backlog of appointments with a very small probability of a same day appointment, while the input to the clinic is controlled via $s^*$. As the ratio $c_a/c_w$ increases to
0.2 and beyond, Open Access quickly becomes the optimal regime and the same-day appointment probability approaches 100%.

The simulated optimal solution was derived by exhaustively searching the whole solution space, which was discretized in increments of 0.05 for $\lambda$, and increments of 1 for $s$.


Our model suggests that an Open Access policy is optimal when we control jointly the practice’s panel size and appointment availability. It is further of interest to put our model in a realistic setting and to compare our panel sizing recommendations with those from Green and Savin (2008) and Liu and Ziya (2014).

Green and Savin (2008) address the question of what is the proper panel size for a medical practice that aims to implement an Open Access policy. By analyzing an $M/D/1/K$ and an $M/M/1/K$ queue with state dependent no-shows (and more elaborate simulation models) they propose panel sizes that achieve a certain (high) probability of a same-day appointment, while the appointment availability remains fixed. In one of the models developed in Liu and Ziya (2014) the appointment availability also remains fixed, and the objective is to find a panel size that maximizes the throughput of an $M/M/1/K$ queue with state dependent no-shows.

In order to make a meaningful comparison, as in Liu and Ziya (2014), we also use the data from the Columbia MRI facility presented in Green and Savin (2008). Appointment availability is $s = 20$ slots per day. When the panel size is $N$, the daily demand is a Poisson process with rate $\lambda = 0.008N$. In Green and Savin (2008) the buffer size of the appointment book is $K = 400$ patients. To make the proper mapping to our model with state-dependent balking, we consider that patients’ willingness to wait is uniformly distributed on $[0, 40]$ with an average $m = K/s = 20$ days. There is no data available regarding the operations of the in-clinic queue. According to Wessman et al. (2014), there is variability in the duration of an MRI scan, and appointments are scheduled in blocks of 45 minutes. Accordingly, we will consider Lognormal service times with mean $\mu^{-1} = 0.75$ hours, a coefficient of variation 0.325, and that the length of a working day is $T = s\mu^{-1} = 15$ hours. Finally, we consider an average show-up rate of $p = 98.7\%$ and a Poisson arrival process. This show-up rate was recovered based on Proposition 2 of Green and Savin (2008) and their empirical estimates.

<table>
<thead>
<tr>
<th>Panel Size</th>
<th>Green and Savin (2008)</th>
<th>Liu and Ziya (2014)</th>
<th>our model with $r = 11$</th>
<th>our model with $r = 12$</th>
<th>our model with $r = 13$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr(Same Day Appointment)</td>
<td>90%</td>
<td>27.6%</td>
<td>99.7%</td>
<td>99.4%</td>
<td>99.0%</td>
</tr>
</tbody>
</table>
Our objective function has 5 components, capturing the various trade-offs discussed throughout the paper. Following the discussion in §6.1, we consider that \( \frac{1}{m}c_b = c_a = c_w = \frac{1}{15}c_o = 1 \). We optimize \( R(\lambda, 20) \) (i.e., \( s = 20 \) is fixed) as it appears in (25) and report the corresponding panel size in Table 1 for different values of \( r \sim c_o\mu^{-1} = 11.25 \), along with the analogous recommendations from Green and Savin (2008) and Liu and Ziya (2014). Our model suggests moderately lower panel sizes compared with the recommendation from Green and Savin (2008). This is not surprising, since our objective function penalizes also for having a congested clinic and patients who balk. Our suggested panel sizes also provide a considerably higher probability for a same day appointment. Under the same setting, Liu and Ziya (2014) propose a significantly larger panel size, which maximizes the daily throughput (captured by one of the five components of our objective). Their probability for a same day appointment is 27.6%, while the probability of an appointment within two days is 47.6%. Our proposed panel size approaches the one suggested by Liu and Ziya (2014) as \( r \) goes beyond 12; higher values of \( r \) promote a higher daily throughput.

7. Conclusion

In this study we address the joint problem of determining the panel size of a medical practice and the number of offered appointments per day. We explicitly model the two separate time scales involved in accessing medical care: appointment delay (order of days, weeks) and clinic delay (order of minutes, hours). Two additional distinctive characteristics of this study are the state-dependent balking behavior of the patients who face long appointment backlogs, and the transient-state analysis of the in-clinic queue, which make the study of a system with high traffic intensity meaningful and bear unique technical challenges.

7.1. Managerial Implications

It is important to consider panel sizing and scheduling decisions in a joint framework. A clinic’s panel size should not only depend on the desired demand volume, but also on the capacity of the system to serve patients. Conversely, appointment availability affects the system’s competence to provide timely access to the desired panel of patients. Furthermore, if these decisions are not taken jointly, undesirable outcomes may occur. For example, we demonstrate that increasing appointment availability does not necessarily guarantee a reduced appointment backlog, as a result of the latent effects of patient’s balking behavior. Therefore, an uninformed approach that assumes that a larger appointment availability secures shorter appointment delays may be false.

By optimizing over the two levers simultaneously, we demonstrate that an Open Access policy is optimal in outpatient scheduling when we account for both clinic and appointment delay. The
benefits from providing timely access to appointments outweigh the costs from potentially overbooking capacity. As a result, under our Open Access regime, the clinic offers a large number of appointment slots per day, reflecting the maximum (potentially overbooked) workload that can be handled. The clinic throughput is controlled by setting a proper panel size which is such that patients can have access to a same-day appointment with high probability and the clinic is not overcrowded.

7.2. Limitations - Future Research

There are a few research directions that we further intend to explore. In this study the no-show rate is treated as constant and does not depend on the appointment backlog. Empirical studies though suggest that the probability of a patient not showing up depends on the appointment delay (Gallucci et al. (2005), Dreher et al. (2008), Norris et al. (2014)). Let $\gamma(k)$ denote the probability of a patient who faces an appointment backlog of $k$ patients being a no-show, increasing in $k$. One possible expression for $\gamma(k)$ is given in Green and Savin (2008): $\gamma(k) = \gamma_{\text{max}} - (\gamma_{\text{max}} - \gamma_{\text{min}}) e^{-\frac{k}{C}}$, where $\gamma_{\text{max}}$ and $\gamma_{\text{min}}$ are the maximum and minimum observed no-show rates respectively, and $C$ is a constant that captures the characteristics of the medical practice. This could be a starting point in incorporating state dependent no-shows in our model.

We have assumed that patients arrive at the clinic according to a renewal process. The literature of appointment scheduling under no-shows suggests that a front-loaded schedule is optimal, i.e., more patients are scheduled to arrive towards the beginning of the working day (see for example Robinson and Chen (2010), LaGanga and Lawrence (2012), Zacharias and Pinedo (2014)). Such a clinic with front-loaded schedules may be approximated as follows: the working day is partitioned into two time intervals $[0,T_1] \cup (T_1,T] = [0,T]$, with the first interval having a higher arrival rate than the second one. Then, the workload can be approximated accordingly by a regulated diffusion process with a piecewise constant drift, and a piecewise constant infinitesimal variance. Further, it is of interest to study rigorously a system with a more refined arrival process, where patients who show up arrive at the time of their scheduled appointment plus a stochastic noise, as in Araman and Glynn (2012) and Honnappa et al. (2015). For the sake of analytical tractability, and in order to account for no-shows, we only incorporate this more realistic setting in our simulation studies that complement our theory.

Though there is a balking penalty in our objective function, our model does not capture the full potential effect of balking on a clinic’s operations, and the demand for medical appointments remains a renewal process. A more elaborate model with patient retrials and/or patients who permanently abandon the panel is a future direction of interest and remains an open problem.
A careful treatment of emergency walk-ins is of interest as well. As demonstrated in Appendix B, the analysis of the in-clinic queue can be readily adjusted to capture emergency walk-ins, given that we know their arrival rate. To take this analysis further, a relationship should be established first between the walk-in rate, the state of the appointment backlog and the clinic’s panel size.

Finally, our closed-form expressions for the expected patients’ in-clinic waiting time and physician’s overtime can be used effectively to make scheduling decisions dynamically. In other words, the tactical decision $s$ can be adjusted dynamically based on the state of the appointment book, and in anticipation of future demand. Such a dynamic scheduling setting can also capture a seasonal effect (e.g., flu) on the demand for medical care. For example, our expressions can be used to represent the cost function from Truong (2015) for accommodating a certain number of patients in one day.

Appendix A: Proofs

Proof of Theorem 1 As in Ward and Glynn (2005), we represent the appointment backlog in terms of the martingale $\{M(i, F_i) : i \geq 0\}$, where $F_i = \sigma((u_1, w_1), ..., (u_i, w_i), u_{i+1}) \subset F$, and

$$M(i) := \sum_{j=1}^{i} \left[ \mathbf{1}\left\{ \frac{Q(t)}{s} \geq mw_j \right\} - E\left( \mathbf{1}\left\{ \frac{Q(t)}{s} \geq mw_j \right\} | F_{j-1} \right) \right]$$

$$= \sum_{j=1}^{i} \left[ \mathbf{1}\left\{ \frac{Q(t)}{s} \geq mw_j \right\} - F\left( \frac{Q(t)}{sm} \right) \right].$$

(26)

We can now write the evolution equation for the backlog as a stochastic integral

$$Q(t) + \int_0^t F\left( \frac{Q(\tau)}{ms} \right) dA(\tau) = A(t) - M(A(t)) - \lfloor st \rfloor + L(t).$$

(27)

Consider further the diffusion scaled process $\tilde{M}_n(t) := \frac{M_n(\lfloor nt \rfloor)}{\sqrt{n}}$. From the pathwise equation for the backlog of appointments in (27), Assumption 1, the scaling in (3) and (4), and some algebra, we obtain

$$\tilde{Q}_n(t) + s\xi \int_0^t \tilde{Q}_n(\tau)d\tau = \tilde{X}_n(t) + \tilde{L}_n(t),$$

(28)

where

$$\tilde{X}_n(t) = \tilde{A}_n(t) + \frac{1}{\sqrt{n}}(n\lambda_n t - \lfloor nst \rfloor) - \tilde{M}_n(\tilde{A}_n(t)) + s\xi \int_0^t \tilde{Q}_n(\tau)d\tau - \int_0^t \sqrt{n}F\left( \frac{\tilde{Q}_n(\tau)}{\sqrt{n}} \right)d\tilde{A}_n(\tau).$$

(29)

To provide an intuition for the representation of the workload in (28), note that from L’Hôpital’s rule $\lim_{y \to \infty} yF(\frac{x}{y}) = \xi x$, so that

$$\frac{1}{\sqrt{n}} \int_0^t F\left( \frac{Q(\tau)}{ms} \right) dA(\tau) = \int_0^t \sqrt{n}F\left( \frac{\tilde{Q}_n(\tau)}{\sqrt{n}} \right)d\tilde{A}_n(\tau) \overset{d}{=} \int_0^t s\xi \tilde{Q}_n(\tau)d\tau$$

(from (31)).
Note that $\tilde{L}_n(0) = 0$, $\tilde{L}_n$ is non-decreasing, $\tilde{L}_n$ increases only when $\tilde{Q}_n = 0$. Therefore for $\gamma = s\xi$ we have $(\tilde{Q}_n, \tilde{L}_n) = (\Phi_\gamma, \Psi_\gamma)(\tilde{X}_n)$.

Next, we wish to derive a diffusion limit for the process $\tilde{X}_n(t)$. Under Assumption 1, and from the Functional Central Limit Theorem and Functional Strong Law of Large Numbers,

$$\tilde{A}_n(t) \Rightarrow \sqrt{s\theta}B(t),$$

$$\tilde{A}_n(t) \to st,$$

and $\frac{1}{\sqrt{n}}(n\lambda_n t - \lfloor snt \rfloor) \to \eta t,$

where $B(t)$ is a standard Brownian motion. Ward and Glynn (2005) proved (in their Theorem 1) that

$$\tilde{M}_n(\tilde{A}_n(t)) \Rightarrow 0$$

and

$$\int_0^t s\xi \tilde{Q}_n(\tau) d\tau - \int_0^t \sqrt{n}F\left(\frac{\tilde{Q}_n(\tau)}{\sqrt{n}}\right) d\bar{A}_n(\tau) \Rightarrow 0.$$

Combining (29), (30), (31), (32), (33), (34) we get the desired weak convergence for $\tilde{X}_n(t)$

$$\tilde{X}_n(t) \Rightarrow \sqrt{s\theta}B(t) + \eta t.$$

From the continuity of the Linearly Generalized Regulator Mapping, and from the Continuous Mapping Theorem, we finally conclude that

$$(\tilde{Q}_n(t), \tilde{L}_n(t)) \Rightarrow (\Phi_\gamma, \Psi_\gamma)(\sqrt{s\theta}B(t) + \eta t),$$

for $\gamma = s\xi$. □

Proof of Lemma 1  (a) Let’s write the expected backlog as $E[\tilde{Q}_a] = \sqrt{\frac{ms^2}{2\pi}} [h(y) - y]$, where $y = \sqrt{\frac{2ms^2 (s-\lambda)}{\lambda \xi^2}}$. Therefore

$$\frac{\partial E[\tilde{Q}_a]}{\partial \lambda} = \sqrt{\frac{ms^2}{2\pi}} \frac{\partial}{\partial \lambda} [h'(y) - 1]$$

$$= -\sqrt{\frac{ms^2}{2\pi}} \sqrt{\frac{2ms}{\xi^2}} \frac{s}{\lambda^2} [h'(y) - 1]$$

$$= \frac{ms^2}{\lambda^2} [1 - h'(y)].$$

It is well known that $0 < h'(x) < 1$ for all $x \in \mathbb{R}$ (see for example Barrow and Cohen (1954)), concluding that $\frac{\partial E[\tilde{Q}_a]}{\partial \lambda} > 0$ for $\lambda > 0$.

(b) It suffices to show that every critical point of $E[\tilde{Q}_a]$ (with respect to $s$) is a local maximum, i.e., $\frac{\partial^2 E[\tilde{Q}_a]}{\partial s^2} < 0$ whenever $\frac{\partial E[\tilde{Q}_a]}{\partial s} = 0$. Then, since $E[\tilde{Q}_a]$ is continuous (and differentiable) in $s$, there can be at most one local maximum, concluding that $E[\tilde{Q}_a]$ is unimodal in $s$. Firstly, we prove the following intermediate result:
Recall that Barrow and Cohen (1954) proved that inequality (36) holds.

\[ \frac{h''(x)|h(x) - x|}{|h'(x)|^2} < 2 \text{ for all } x \in \mathbb{R}. \]

Proof of Lemma 7

\[ \frac{h''(x)|h(x) - x|}{|h'(x)|^2} < 2 \]
\[ \iff \frac{h(x)[h(x) - x]^3 + h^2(x)[h(x) - x]^2 - h(x)[h(x) - x]}{1 - 2h(x)[h(x) - x] + h^2(x)[h(x) - x]^2} < 2 \]
\[ \iff 0 < -h(x)[h(x) - x]^3 + h^2(x)[h(x) - x]^2 - 3h(x)[h(x) - x] + 2, \]

where (35) follows from the fact that

\[ h'(x) = h(x)[h(x) - x], \quad (37) \]
and \[ h''(x) = h(x)[h(x) - x]^2 + h^2(x)[h(x) - x] - h(x). \quad (38) \]

Barrow and Cohen (1954) proved that inequality (36) holds. □

Recall that

\[ E[\hat{Q}_a] = \sqrt{\frac{m\theta^2}{2\pi}} \sqrt{s}[h(y) - y], \]
where \( y = \sqrt{\frac{2m}{\xi^2}} \sqrt{s \lambda}. \quad (39) \)

Then

\[ \frac{\partial y}{\partial s} = \sqrt{\frac{2m}{\xi^2}} \left( \frac{3\sqrt{\pi}}{2\sqrt{2\pi}} - \frac{1}{2\sqrt{2\pi}} \right) = \frac{3\sqrt{2\pi}}{2\sqrt{2\pi}} + \frac{1}{\sqrt{2\pi}} \sqrt{\frac{2m}{\xi^2}}, \]
\[ \frac{\partial^2 y}{\partial s^2} = \frac{3}{2s} \frac{\partial y}{\partial s} - \frac{3y}{2s^2} - \frac{1}{2s\sqrt{2\pi}} \sqrt{\frac{2m}{\xi^2}}, \]
and \[ \frac{\partial E[\hat{Q}_a]}{\partial s} = \sqrt{\frac{m\theta^2}{2\pi}} \left[ \frac{1}{2\sqrt{2\pi}} [h(y) - y] - \sqrt{s}[1 - h'(y)] \frac{\partial y}{\partial s} \right]. \quad (40) \]

From (43), whenever \( \frac{\partial E[\hat{Q}_a]}{\partial s} = 0 \) then

\[ \frac{\partial y}{\partial s} = \frac{h(y) - y}{2\sqrt{2\pi}[1 - h'(y)]}, \quad (41) \]
and \[ \sqrt{\frac{2m}{\xi^2}} \frac{[h(y) - y] - 3y[1 - h'(y)]}{2\sqrt{s}[1 - h'(y)]} = \frac{[h(y) - y]}{2\sqrt{s}[1 - h'(y)]} - \frac{3y}{2\sqrt{s}}. \quad (42) \]

From (42), (44) and (45), whenever \( \frac{\partial E[\hat{Q}_a]}{\partial s} = 0 \) then

\[ \frac{\partial^2 y}{\partial s^2} = \frac{3[h(y) - y]}{4\sqrt{2\pi}[1 - h'(y)]} - \frac{3y}{2s^2} - \frac{|h(y) - y|}{4\sqrt{2\pi}[1 - h'(x)]} + \frac{3y}{2s^2} \]
\[ = \frac{2|h(y) - y| - 3y[1 - h'(y)]}{4\sqrt{2\pi}[1 - h'(y)]}. \quad (43) \]
From (43), (44), and (46), whenever $\frac{\partial E[\hat{Q}_0]}{\partial s} = 0$ then
\[
\frac{\partial^2 E[\hat{Q}_0]}{\partial s^2} = \sqrt{\frac{\alpha_0}{2\xi}} \left[ -\frac{1}{4\sqrt{\pi}} [h(y) - y] - \frac{1}{2\sqrt{\pi}} (1 - h'(y)) \frac{h(y) - y}{2s [1 - h'(y)]} - \frac{1}{2\sqrt{\pi}} (1 - h'(y)) \frac{h(y) - y}{2s [1 - h'(y)]} \right]
+ \sqrt{\frac{\alpha_0}{2\xi}} \left[ \frac{h(y) - y}{2s [1 - h'(y)]} - \sqrt{\pi} (1 - h'(y)) \frac{2h(y) - y - 3y [1 - h'(y)]}{4s^2 [1 - h'(y)]} \right]
= \sqrt{\frac{\alpha_0}{2\xi}} \left[ -5 [h(y) - y] + 3y [1 - h'(y)] + \frac{h''(y) [h(y) - y]^2}{[1 - h'(y)]^2} \right]
< \sqrt{\frac{\alpha_0}{2\xi}} \left[ -3 [h(y) - y] + 3y [1 - h'(y)] \right] + 3 \left( \frac{h(y) - y}{2s} \right)^2
- 2y + yh^2(y) - y^2h(y) < 0.
\]

Inequality (47) follows from Lemma 7 and the fact that $h(x) - x > 0$ for all $x \in \mathbb{R}$. Equality (48) follows from the relationship $h'(x) = h(x)[h(x) - x]$. Inequality (49) requires a little extra work: From Theorem 2.5 of Baricz (2008), the function $x \mapsto xh(x) \left[ \frac{1}{h(x)} \right]'$ is strictly decreasing on $(0, \infty)$, implying that
\[
\left[ xh(x) \left[ \frac{1}{h(x)} \right]' \right]' < 0
\iff
\left[ xh(x) \left[ \frac{h'(x)}{h^2(x)} \right]' \right] < 0
\iff
\left[ \frac{xh'(x)}{h(x)} \right]' > 0
\iff
\left[ \frac{xh''(x) + h'(x)h(x) - x[h'(x)]^2}{h^3(x)} \right] > 0
\iff
\left[ \frac{xh^2(x)[h(x) - x]^2 + xh^3(x)[h(x) - x] - xh^2(x) + h^2(x)[h(x) - x] - xh^2(x) + h(x) - x^2h(x)}{h^3(x)} \right] > 0
\iff
h(x) - 2x + xh^2(x) - x^2h(x) > 0,
\]
where (50) follows from the fact the $h(x) > 0$ for all $x \in \mathbb{R}$, and (51) follows from (37) and (38). Therefore, $\frac{\partial^2 E[\hat{Q}_0]}{\partial s^2} < 0$ whenever $\frac{\partial E[\hat{Q}_0]}{\partial s} = 0$.

(c) Recall that $E[\hat{Q}_0] = \sqrt{\frac{\alpha_0}{\xi r^2}} [h(y) - y]$, where $y = \sqrt{\frac{2\alpha_0}{\xi r^2} \frac{(\xi - \lambda)}{\lambda}}$. Therefore
\[
\frac{\partial E[\hat{Q}_0]}{\partial r} = \frac{1}{2\sqrt{\pi}} \left[ h(y) - y \right] + \sqrt{\frac{\alpha_0}{2\xi}} \frac{\partial y}{\partial r} \left[ h'(y) - 1 \right].
\]
From (53) and some algebra, the condition $\frac{\partial E[\hat{Q}_0]}{\partial r} > 0$ is true if and only if $h(y) - 2y + yh^2(y) - y^2(h(y)) > 0$, which is indeed true from (52).

(d) Recall that $E[\hat{Q}_0] = \sqrt{\frac{\alpha_0}{\xi r^2}} [h(y) - y]$, where $y = \sqrt{\frac{2\alpha_0}{\xi r^2} \frac{(\xi - \lambda)}{\lambda}}$. Therefore
\[
\frac{\partial E[\hat{Q}_0]}{\partial \xi} = -\frac{1}{2\xi} \left[ \frac{3}{2} \sqrt{\frac{\alpha_0}{\xi r^2}} [h(y) - y] + \sqrt{\frac{\alpha_0}{2\xi}} \frac{\partial y}{\partial \xi} \left[ h'(y) - 1 \right] \right].
\]
From (54) and some algebra, the condition \( \frac{\partial E[\tilde{Q}_a]}{\partial s} < 0 \) is true if and only if \( h(y) - 2y + yh^2(y) - y^2h(y) > 0 \), which is indeed true from (52).

(e) From Baricz (2008), \( 0 < h(x) - x < \frac{1}{x} \) for all \( x > 0 \). Therefore, for \( s > \lambda \),

\[
0 < E[\tilde{Q}_a] = \sqrt{\frac{ms^2}{2\lambda}} \sqrt{h(y) - y} < \sqrt{\frac{ms^2}{2\lambda}} \frac{1}{y}
\]

\[
\Rightarrow 0 < E[\tilde{Q}_a] < \sqrt{\frac{ms^2}{2\lambda}} \sqrt{\frac{\sqrt{\lambda}}{2\pi} \sqrt{s(\lambda - s)}} = \frac{\sqrt{s}}{2(s - \lambda)}
\]

\[
\Rightarrow 0 \leq \lim_{s \to \infty} E[\tilde{Q}_a] \leq \lim_{s \to \infty} \frac{\sqrt{s}}{2(s - \lambda)} = 0
\]

(f) Firstly we show that \( \lim_{\lambda \to 0} E[\tilde{Q}_a] = 0 \).

\[
\lim_{\lambda \to 0} E[\tilde{Q}_a] = \lim_{\lambda \to 0} \left[ \frac{ms(\lambda-s)}{\lambda^2} + \sqrt{\frac{ms^2}{2\lambda}} \frac{h(\frac{s-\lambda}{\lambda} \frac{2ms}{\sqrt{\xi \theta^2}})}{\lambda^2} \right]
\]

\[
= \lim_{\lambda \to 0} \left[ \frac{ms(\lambda-s)}{\lambda^2} + \frac{ms^2}{2\xi} \frac{\phi(h)}{\lambda} \frac{2ms}{\sqrt{\xi \theta^2}} \right]
\]

\[
= \lim_{\lambda \to 0} \left[ \frac{ms(\lambda-s)}{\lambda^2} - \frac{ms^2}{2\xi} \frac{s}{\lambda^2} \frac{\lambda}{\sqrt{\xi \theta^2}} \right]
\]

\[
= 0,
\]

where (58) comes from the fact that \( \lim_{\lambda \to 0} \phi(h) = \lim_{\lambda \to 0} \Phi(h) = 0 \) and L'Hôpital's rule.

Then we show that \( \lim_{\lambda \to 0} \frac{\partial E[\tilde{Q}_a]}{\partial \lambda} = 0 \). It is well known that \( \lim_{x \to \infty} h'(x) = 1 \) (see for example Barrow and Cohen (1954)), and therefore

\[
\lim_{\lambda \to 0} \frac{\partial E[\tilde{Q}_a]}{\partial \lambda} = \lim_{\lambda \to 0} \left[ \frac{ms}{\sqrt{\xi \theta^2}} - \sqrt{\frac{ms^2}{2\lambda}} \sqrt{\frac{2ms}{\xi \theta^2}} \phi(h) \right]
\]

\[
= \lim_{\lambda \to 0} \left[ \frac{ms}{\sqrt{\xi \theta^2}} - \frac{ms}{\sqrt{\xi \theta^2}} \right]
\]

\[
= 0.
\]

\[ \square \]

Proof of Lemma 2 (a) Let \( \tilde{Q}_a(\lambda, s) \) be the steady state queue length as a function of \( \lambda \) and \( s \).

Then \( \tilde{Q}_a(\lambda, s) \leq \tilde{Q}_a(s, s) \), for all \( \lambda \leq s \). Therefore,

\[
\Pr(\text{Same-day appointment}) = \Pr(\tilde{Q}_a(\lambda, s) < s)
\]

\[
\geq \Pr(\tilde{Q}_a(s, s) < s)
\]

\[
= \int_0^s 2\sqrt{\frac{2s}{ms^2}} \phi\left(\sqrt{\frac{2s}{ms^2}}\right) dx
\]

\[
= 2\Phi\left(\sqrt{\frac{2s}{ms^2}}\right) - 1, \text{ for all } \lambda \leq s.
\]
From (59) we get that
\[ 1 \geq \Pr(\text{Same-day appointment}) \geq 2\Phi\left( \sqrt{\frac{2\mu}{\sigma^2}} \right) - 1. \]

\[ \Rightarrow 1 \geq \lim_{s \to \infty} \Pr(\text{Same-day appointment}) \geq \lim_{s \to \infty} 2\Phi\left( \sqrt{\frac{2\mu}{\sigma^2}} \right) - 1 = 1, \]

concluding that \( \lim_{s \to \infty} \Pr(\text{Same-day appointment}) = 1. \) □

**Proof of Lemma 3** By Ito formula (see Theorem 3.3 of Karatzas and Shreve (1991), p.149), we get
\[
\begin{align*}
  f(Y(T)) &= f(Y(0)) + \int_0^T f'(Y(t))dX(t) + \int_0^T f'(Y(t))dL(t) + \frac{1}{2} \int_0^T f''(Y(t))d<X>_t \\
  &= f(0) + \alpha \int_0^T f'(Y(t))dt + \beta \int_0^T f'(Y(t))dB(t) + \int_0^T f'(Y(t))dB(t) + \frac{\sigma^2}{2} \int_0^T f''(Y(t))dt,
\end{align*}
\]

where \( f \) is a twice continuously differentiable function. Firstly assume that \( \alpha \neq 0 \). For \( f(x) := x^2 \) we get
\[
Y^2(T) = 2\alpha \int_0^T Y(t)dt + 2\beta \int_0^T Y(t)dB(t) + 2 \int_0^T Y(t)dB(t) + \alpha^2 \int_0^T dt
\]
\[
= 2\alpha \int_0^T Y(t)dt + 2\beta \int_0^T Y(t)dB(t) + 2\alpha^2 T. \tag{59}
\]

From (59) we get that \( Y^2(T) - 2\alpha \int_0^T Y(t)dt - 2\beta^2 T = 2\beta \int_0^T Y(t)dB(t) \), a martingale, concluding that \( E\left[ \int_0^T Y(t)dt \right] = E[Y^2(T)] - \sigma^2 T. \)

Next, assume that \( \alpha = 0 \). For \( f(x) := x^3 \) we get
\[
Y^3(T) = 3\beta \int_0^T Y^2(t)dB(t) + 3 \int_0^T Y^2(t)dB(t) + 3\beta^2 \int_0^T Y(t)dt
\]
\[
= 3\beta \int_0^T Y^2(t)dB(t) + 3\beta^2 \int_0^T Y(t)dt. \tag{60}
\]

From (60) we get that \( Y^3(T) - 3\beta^2 \int_0^T Y(t)dt = 3\beta \int_0^T Y^2(t)dB(t) \), a martingale, concluding that \( E\left[ \int_0^T Y(t)dt \right] = E[Y^3(T)]. \) □

**Proof of Lemma 4** When \( \beta = 0 \) there is no variability, and therefore \( Y(t) = \max(\alpha, 0)t \) with probability one.

Then consider the case where \( \beta > 0 \). First, suppose that \( \alpha \neq 0 \) and let \( g(x) := \frac{2\alpha}{\beta^2} e^{-\frac{2\alpha^2}{\beta^2}} \Phi\left( \frac{-\alpha t}{\beta \sqrt{t}} \right) \).

From the probability density of \( Y(t) \) in (22) we get
\[
\int_0^\infty g(x)dx = \int_0^\infty \frac{2}{\sqrt{\pi}} \phi\left( \frac{x - \alpha t}{\beta \sqrt{t}} \right)dx - 1
\]
\[
= 2\Phi\left( \frac{2\sqrt{t}}{\beta} \right) - 1. \tag{61}
\]
By definition,

$$E[Y(t)] = \int_0^\infty xf_Y(t)dx = \int_0^\infty \frac{2x}{\beta \sqrt{t}} \phi(t \frac{x-\alpha t}{\beta \sqrt{t}})dx - \int_0^\infty xg(x)dx.$$  \hspace{1cm} \text{(62)}$$

For the first integral in (62) we use the transformation $u = \frac{x-\alpha t}{\beta \sqrt{t}}$ to get

$$\int_0^\infty \frac{2x}{\beta \sqrt{t}} \phi(t \frac{x-\alpha t}{\beta \sqrt{t}})dx = 2\beta \sqrt{t} \phi(\frac{\alpha \sqrt{t}}{\beta}) + 2\alpha t \Phi(\frac{\alpha \sqrt{t}}{\beta}).$$  \hspace{1cm} \text{(63)}$$

Then, we note that

$$g(x) = \frac{\beta^2}{2\alpha} g'(x) + \frac{1}{\alpha \sqrt{t}} \phi(\frac{x-\alpha t}{\beta \sqrt{t}}),$$  \hspace{1cm} \text{(64)}$$

and therefore

$$\int_0^\infty xg(x)dx = \int_0^\infty \frac{\beta^2}{2\alpha} g'(x)dx + \int_0^\infty \frac{1}{\alpha \sqrt{t}} \phi(\frac{x-\alpha t}{\beta \sqrt{t}})dx.$$  \hspace{1cm} \text{(65)}$$

The second integral in (65) is the same as the one in (63), divided by 2. For the first integral in (65) we apply integration by parts to get

$$\int_0^\infty \frac{\beta^2}{2\alpha} g'(x)dx = \lim_{x \to \infty} \frac{\beta^2}{2\alpha} g(x) - \int_0^\infty \frac{\beta^2}{2\alpha} g(x)dx$$

$$= -\frac{\beta^2}{2\alpha}(2\Phi(\frac{\alpha \sqrt{t}}{\beta}) - 1).$$  \hspace{1cm} \text{(66)}$$

The equality in (67) comes from the fact that $\lim_{x \to \infty} \frac{\beta^2}{2\alpha} g(x) = 0$ and from (61). If we combine (62), (63), (65), (67) we get

$$E[Y(t)] = \beta \sqrt{t} \phi(\frac{\alpha \sqrt{t}}{\beta}) + 2\beta \sqrt{t} + 2\alpha t \Phi(\frac{\alpha \sqrt{t}}{\beta}) - \frac{\beta^2}{2\alpha}.$$  

We apply similar methods to compute the second moment

$$E[Y^2(t)] = \int_0^\infty x^2 f_Y(t)dx = \int_0^\infty \frac{2x^2}{\beta \sqrt{t}} \phi(t \frac{x-\alpha t}{\beta \sqrt{t}})dx - \int_0^\infty x^2g(x)dx.$$  \hspace{1cm} \text{(68)}$$

For the first integral in (68) we use the transformation $u = \frac{x-\alpha t}{\beta \sqrt{t}}$ and integration by parts to get

$$\int_0^\infty \frac{2x^2}{\beta \sqrt{t}} \phi(t \frac{x-\alpha t}{\beta \sqrt{t}})dx = 2\alpha \beta t \sqrt{t} \phi(\frac{\alpha \sqrt{t}}{\beta}) + 2(\beta^2 t + \alpha^2 t^2) \Phi(\frac{\alpha \sqrt{t}}{\beta}).$$  \hspace{1cm} \text{(69)}$$

For the second integral in (68), we make use of (64) to get

$$\int_0^\infty x^2g(x)dx = \int_0^\infty \frac{\beta^2}{2\alpha} g'(x)dx + \int_0^\infty \frac{\beta^2}{2\alpha} \phi(\frac{x-\alpha t}{\beta \sqrt{t}})dx.$$  \hspace{1cm} \text{(70)}$$

The second integral in (70) is the same as the one in (69), divided by 2. For the first integral in (70) we apply integration by parts to get

$$\int_0^\infty \frac{\beta^2}{2\alpha} g'(x)dx = \lim_{x \to \infty} \frac{\beta^2}{2\alpha} g(x) - \int_0^\infty \frac{\beta^2}{2\alpha} g(x)dx$$

$$= \frac{\beta^2}{2\alpha^2} \left(2\Phi(\frac{\alpha \sqrt{t}}{\beta}) - 1\right) - \frac{\beta^2}{2\alpha} y(\frac{\alpha \sqrt{t}}{\beta}) - \beta^2 t \Phi(\frac{\alpha \sqrt{t}}{\beta}).$$  \hspace{1cm} \text{(71)}$$
The equality in (72) comes from the fact that \( \lim_{x \to \infty} \frac{\beta^2 x^2}{2a} g(x) = 0 \) (applying L'Hôpital's rule) and from (65). If we combine (68), (69), (70), (72) we get

\[
E[Y^2(t)] = \beta \sqrt{t} \left( \alpha t + \frac{\beta^2}{\alpha} \phi \left( \frac{\alpha \sqrt{t}}{\beta} \right) \right) + (2\beta^2 t + \alpha^2 t^2 - \frac{\beta^4}{\alpha^2}) \Phi \left( \frac{\alpha \sqrt{t}}{\beta} \right) + \frac{\beta^4}{\alpha^2}.
\]

Finally, consider the case where \( \alpha = 0 \). \( E[Y(t)] = \int_0^\infty \frac{2}{\beta \sqrt{t}} \phi \left( \frac{x}{\beta \sqrt{t}} \right) dx = \frac{2\beta \sqrt{t}}{\sqrt{\pi}} \). Using integration by parts one can show that \( E[Y^2(t)] = \frac{2\beta \sqrt{t}}{\sqrt{\pi}} \). □

**Proof of Theorem 2**  
Consider the case where \( \beta > 0 \) and \( \alpha = 0 \). We apply integration by parts to get

\[
E[Y^2(t)] = \int_0^\infty \frac{2}{\beta \sqrt{t}} \phi \left( \frac{x}{\beta \sqrt{t}} \right) dx = \frac{2\beta \sqrt{t}}{\sqrt{\pi}}.
\]

Equations (73) and (74) follow from the fact that \( \lim_{\alpha \to 0} \Phi \left( \frac{\alpha \sqrt{t}}{\beta} \right) - \frac{1}{2} = 0 \) and L'Hôpital's rule. □

**Proof of Theorem 2**  
Consider the case where \( \beta > 0 \) and \( \alpha = 0 \). We apply integration by parts to get

\[
E[Y^3(t)] = \int_0^\infty \frac{2}{\beta \sqrt{t}} \phi \left( \frac{x}{\beta \sqrt{t}} \right) \left( (2\beta^2 t + \alpha^2 t^2 - \frac{\beta^4}{\alpha^2}) \Phi \left( \frac{\alpha \sqrt{t}}{\beta} \right) + \frac{\beta^4}{\alpha^2} \right) dx
\]

The rest of the proof follows from Lemmas 3 and 4 and equation (23). □

**Proof of Lemma 6**  
Since \( \phi(\cdot) \) and \( \Phi(\cdot) \) are continuous functions, we only need to show that \( E[\hat{W}_c] \) and \( E[\hat{O}_c] \) are continuous in \( \alpha \) at \( \alpha = 0 \). We only consider the case where \( \beta^2 > 0 \), the case where \( \beta^2 = 0 \) is trivial and is omitted. It suffices to show that \( \lim_{\alpha \to 0} E[\hat{W}_c] = \frac{4\beta T \sqrt{T}}{3 \sqrt{2\pi}} \) and that

\[
\lim_{\alpha \to 0} E[\hat{O}_c] = \frac{2\beta \sqrt{T}}{\sqrt{2\pi}}.
\]

\[
\lim_{\alpha \to 0} E[\hat{W}_c] = \lim_{\alpha \to 0} \left[ \frac{1}{2\alpha^2} \left[ \beta^3 \alpha \sqrt{T} T + \beta^3 \alpha \sqrt{T} \phi \left( \frac{\alpha \sqrt{T}}{\beta} \right) \right] + (2\beta^2 \alpha^2 T + \alpha^4 T^2 - \beta^4) \Phi \left( \frac{\alpha \sqrt{T}}{\beta} \right) + \frac{\beta^4}{\alpha^2} - \beta^2 \alpha^2 T \right]
\]

\[
= \lim_{\alpha \to 0} \left[ \frac{1}{2\alpha^2} \left[ 4\beta \alpha^2 T \sqrt{T} \phi \left( \frac{\alpha \sqrt{T}}{\beta} \right) + (4\beta^2 \alpha^2 T + 4\alpha^3 T^2) \Phi \left( \frac{2\alpha \sqrt{T}}{\beta} \right) - 2\beta^2 \alpha^2 T \right] \right]
\]

(76)
\[
\lim_{\alpha \to 0} \left[ 4\beta T \sqrt{T} \phi \left( \frac{\alpha \sqrt{T}}{\rho} \right) + 4\beta^2 T \frac{\phi \left( \frac{\alpha \sqrt{T}}{\rho} \right)}{\alpha} \right] = \lim_{\alpha \to 0} \left[ 8\beta T \sqrt{T} \phi \left( \frac{\alpha \sqrt{T}}{\rho} \right) \right] = \frac{2\beta \sqrt{T}}{\rho \sqrt{\pi}}.
\]

\[
\lim_{\alpha \to 0} E[\tilde{O}_\alpha] = \lim_{\alpha \to 0} \frac{1}{\mu} E[Y(T)] = \frac{2\beta \sqrt{T}}{\rho \sqrt{\pi}}.
\]

Equation (76) comes from L’Hôpital’s rule and some algebra. \( \square \)

**Proof of Theorem 3** Suppose that \((\lambda^*, s^*)\) is an optimal solution to \((P)\). We denote the feasible region with \(D := \{ (\lambda, s) : s \leq M, \lambda \leq M, s \geq 0, \lambda \geq 0 \} \subset \mathbb{R}^2\), and let \(D_1 := \{ (\lambda, s) \in D : s \leq \lambda \}\) and \(D_2 := \{ (\lambda, s) \in D : \lambda \leq s \}\). Clearly \(D = D_1 \cup D_2\). We will firstly show that \((\lambda^*, s^*)\) is in \(D_2\).

Consider the optimization problem \(\max_{(\lambda, s) \in D_1} R(\lambda, s)\), i.e.,

\[
\max_{\lambda, s} R(\lambda, s) = r p \lambda o b k - c_b (\lambda - \lambda o b k) - c_a E[\tilde{Q}_a] - c_w E[M_s \lambda_c] - c_o E[\tilde{O}_c | \lambda_c] \\
\text{s.t. } s \leq \lambda \\
\lambda \leq M \\
s \geq 0
\]

\((P_1)\)

Under the assumption that \(\lambda_c = \frac{p \min (\lambda, s)}{T}\) with probability one, the objective function and its partial derivatives can be written as

\[
R = r ps - c_b (\lambda - s) - c_a E[\tilde{Q}_a] - c_w E[M_s | \lambda_c = ps/T] - c_o E[\tilde{O}_c | \lambda_c = ps/T]
\]

\[
\frac{\partial R}{\partial \lambda} = -c_b - c_a \frac{\partial E[\tilde{Q}_a]}{\partial \lambda} \\
\frac{\partial R}{\partial s} = r p - c_u \frac{\partial E[\tilde{Q}_a]}{\partial s} - c_w \frac{\partial E[M_s | \lambda_c = ps/T]}{\partial s} - c_o \frac{\partial E[\tilde{O}_c | \lambda_c = ps/T]}{\partial s}
\]

The first order conditions for optimality are

\[
\frac{\partial R}{\partial \lambda} + \xi_1 - \xi_2 = 0 \\
\frac{\partial R}{\partial s} - \xi_1 + \xi_3 = 0 \\
s - \lambda \leq 0, \lambda - M \leq 0, -s \leq 0 \\
\xi_1 (s - \lambda) = \xi_2 (\lambda - M) = \xi_3 s = 0 \\
\xi_1, \xi_2, \xi_3 \geq 0.
\]

Suppose (for contradiction) that \(s < \lambda\). From (80) we have that \(\xi_1 = 0\), and (77) implies that \(c_b + c_a \frac{\partial E[\tilde{Q}_a]}{\partial \lambda} + \xi_2 = 0\), which is a contradiction, since from Lemma 1(b) we have that \(\frac{\partial E[\tilde{Q}_a]}{\partial \lambda} > 0\).

Therefore, if \((\lambda^*, s^*) \in D_1\) then \(s^* = \lambda^*\), concluding that \((\lambda^*, s^*)\) is in \(D_2\).
Consider now the optimization problem $\max_{(\lambda, s) \in D^2} R(\lambda, s)$, i.e.,

$$\max_{\lambda, s} R(\lambda, s) = rp\lambda_{\text{book}} - c_0(\lambda - \lambda_{\text{book}}) - c_a E[\tilde{Q}_a] - \frac{c_w}{a} E_{\lambda_c} [E[W_c | \lambda_c]] - c_o E_{\lambda_c} [E[\tilde{O}_c | \lambda_c]]$$

s.t. $\lambda \leq s$

$$s \leq M$$

$$\lambda \geq 0.$$

Under the assumption that $\lambda_c = \frac{a \min(\lambda, s)}{T}$ with probability one, the objective function and its partial derivatives can be written as

$$R = rp\lambda - c_a E[\tilde{Q}_a] - \frac{c_w}{a} E[W_c | \lambda_c = p\lambda/T] - c_o E[\tilde{O}_c | \lambda_c = p\lambda/T]$$

$$\frac{\partial R}{\partial \lambda} = rp - c_a \frac{\partial E[\tilde{Q}_a]}{\partial \lambda} - \frac{c_w}{a} \frac{\partial E[W_c | \lambda_c = p\lambda/T]}{\partial \lambda} - c_o \frac{\partial E[\tilde{O}_c | \lambda_c = p\lambda/T]}{\partial \lambda}$$

$$\frac{\partial R}{\partial s} = -c_a \frac{\partial E[\tilde{Q}_a]}{\partial s}.$$

The first order conditions for optimal solution to $(P_2)$ are

$$\frac{\partial R}{\partial \lambda} - \xi_1 + \xi_3 = 0$$  \hspace{1cm} (82)

$$\frac{\partial R}{\partial s} + \xi_1 - \xi_2 = 0$$  \hspace{1cm} (83)

$$\lambda - s \leq 0, s - M \leq 0, -\lambda \leq 0$$  \hspace{1cm} (84)

$$\xi_1(\lambda - s) = \xi_2(s - M) = \xi_3\lambda = 0$$  \hspace{1cm} (85)

$$\xi_1, \xi_2, \xi_3 \geq 0.$$  \hspace{1cm} (86)

Note that a feasible solution can only belong to one of the following (mutually exclusive) regions:

$$(\lambda, s) : 0 < \lambda < s < M, \quad (\lambda, s) : 0 = \lambda \leq s \leq M, \quad (\lambda, s) : 0 < \lambda = s < M, \quad (\lambda, s) : 0 < \lambda \leq s = M.$$

**Case 1:** Suppose that $0 < \lambda^* < s^* < M$. From (85) we have that $\xi_1 = \xi_2 = \xi_3 = 0$ and from (83) we get that $\frac{\partial R}{\partial s} = 0$. From Lemma 1, $E[\tilde{Q}_a]$ is unimodal in $s$, and in particular, the unique critical point is a local maximum. Therefore, if $(\lambda^*, s^*)$ is an optimal solution to $(P_2)$, then either $s^* = \lambda^*$ or $s^* = M$, which is a contradiction.

**Case 2:** Suppose that $\lambda^* = 0$. Then $E[\tilde{Q}_a] = 0$ for all $s$, and from (83) and (85) we get that $\xi_1 = 0$. From (82) we get $\xi_3 = -\frac{\partial R}{\partial \lambda} |_{(0, s^*)}$. Therefore, the conditions (82)-(86) are satisfied only if $\frac{\partial R}{\partial \lambda} |_{(0, s^*)} \leq 0$.

**Case 3:** Suppose that $0 < \lambda^* = s^* < M$. From (85) we have that $\xi_2 = \xi_3 = 0$, and from (83) we get that $\xi_1 = c_a \frac{\partial E[\tilde{Q}_a]}{\partial s} \geq 0$. Therefore, from (41) and (43), the condition $\frac{\partial E[\tilde{Q}_a]}{\partial s} \geq 0$ is satisfied if and only if

$$\frac{1}{2\sqrt{\pi}} [h(0) - 0] - \sqrt{s^*} [1 - h'(0)] \frac{1}{\sqrt{2\pi s^*}} \sqrt{\frac{2n}{\xi s^*}} \geq 0 \iff \frac{1}{\sqrt{2\pi s^*}} - \frac{\sqrt{\pi}}{\sqrt{2\pi s^*}} \geq 0 \iff s^* \leq \frac{\xi s^* \pi}{4m(\pi - 2\pi)}.$$  \hspace{1cm} $\Box$
Appendix B: Extension accounting for walk-ins
The analysis in §5.1 can be extended to account for walk-ins. Consider a $\sum_{i=1}^{2} GI_i/GI/1$ queue, where the two streams of arrivals come from scheduled appointments and emergency walk-ins. Besides the model primitives in §5.1, consider further an independent sequence of iid random variables $\{u_{w,i}: i \geq 0\}$, with associated arrival times $N_w(t) = \max\{k \geq 0: \sum_{i=1}^{k} u_{w,i} \leq t\}$. The random variable $u_{w,i}$ denotes the inter-arrival time between the $(i-1)^{th}$ and $i^{th}$ emergency walk-in patients, has finite mean $\lambda_w^{-1}$ and finite squared coefficient of variation $c_w^2$. As a convention, $u_{w,0} = 0$. The arrival process at the single server queue is a superposition of the two arrival streams $\{N(t) := N_c(t) + N_w(t), t \geq 0\}$, with associated arrival times $t_n := \inf\{t \geq 0 : N(t) \geq n\}$ and inter-arrival times $\tau_n := t_n - t_{n-1}$. The superposition arrival process $\{N(t), t \geq 0\}$ is a renewal process if and only if the processes $\{N_c(t), t \geq 0\}$ and $\{N_w(t), t \geq 0\}$ are Poisson. In Whitt (1982), $\{N(t), t \geq 0\}$ is approximated by a renewal process with the inter-arrival times having mean $(\lambda_c + \lambda_w)^{-1}$ and squared coefficient of variation $\frac{\lambda_c c_c^2 + \lambda_w c_w^2}{(\lambda_c + \lambda_w)^2}$.

Under this setting, the patients’ expected aggregate waiting time and physician’s overtime can be approximated as in Theorem 2, with the drift $\alpha$ being replaced with $\hat{\alpha} := \lambda_c + \lambda_w - \mu$, and the infinitesimal variance $\beta^2$ being replaced with $\hat{\beta}^2 := \lambda_c c_c^2 + \lambda_w c_w^2 + \min(\lambda_c + \lambda_w, \mu) c_v^2$.

References


LaGanga, L.R., S.R. Lawrence. 2007. Clinic overbooking to improve patient access and increase provider productivity. Decision Sciences 38(2) 251–276.


