Patients nationwide experience difficulties in accessing medical appointments in a timely manner due to long backlogs. Meanwhile, patients do not always show up for their scheduled services, with significant no-show rates. Unattended appointments result in under-utilization of a clinic’s valuable resources, 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. We study 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 separate time scales involved in accessing medical care: appointment delay (order of days, weeks) and clinic delay (order of minutes, hours). We analyze the two queueing systems associated with each type of delay, and provide explicit expressions for the performance measures of interest based on diffusion approximations. In our analysis we capture many features of the complex reality of outpatient care, including patients' non-punctuality, no-shows, balking behavior, and stochastic service times. Two additional distinctive characteristics of this study are the balking behavior of the patients who face long appointment backlogs, and the transient-state analysis of the clinic delay, which allow the study of a system with traffic intensity greater than one.

Concerning the panel sizing and appointment scheduling decisions, our analysis provides theoretical and numerical support that the two-variable optimization problem reduces to a single variable—one, and either an “Open Access” policy is optimal, or supply and demand are perfectly matched and are both very small (“Limited Access” regime). Under our Open Access regime, the clinic offers as many appointment slots as possible per day, and the optimal panel size depends on the clinic’s characteristics. A solution within the Limited Access regime arises when the service times are long, and the patients are very sensitive to the appointment delay.

Key words: healthcare management, patient flow management, panel-size, appointment scheduling, no-shows, balking, diffusion approximations, open access

1. Introduction

Patients nationwide experience difficulties in accessing medical appointments in a timely manner due to long backlogs. Merritt (2011), by surveying 1162 medical offices in 15 US metropolitan areas, found that the waiting time to get an appointment averages 20.5 days and depends on the
specialty: 15.5 for cardiology, 22.1 days for dermatology, 27.5 days for obstetrics/gynecology, 16.8 days for orthopedic surgery, 20.3 days for family practice.

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 clinic’s characteristics (Cayirli et al. (2006)). Defife et al. (2010) report a 21% no-show rate in psychotherapy appointments, Dreier 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 under-utilization of a clinic’s valuable resources, 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 the 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 and 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 jointly, and to provide explicit 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. The second lever is the appointment availability, i.e., the number of offered appointment slots per day. It is clear that an increased panel size comes along with a more congested clinic and longer appointment delays. Concerning 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 as well, induced by the patients’ balking behavior: the more the appointment availability, the more likely it is that a patient will have access to an appointment slot of her preference and not balk.

In this paper, we study 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. Our analysis provides theoretical and numerical support that either an “Open Access” policy is optimal in outpatient scheduling, or supply and demand are perfectly matched and are both very small.

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, 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 the Appendix.

2. Related Literature

There is an extensive literature on appointment scheduling, mostly motivated by healthcare applications. The vast majority of the literature considers the clinic delay and studies the trade-offs between the benefits of efficient physician 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 (2013) 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 aforementioned stream of literature, we do not consider the optimal intra-day 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 provide analytical expressions for the performance measures of interest based on diffusion approximations. The renewal arrival process captures both patients’ non-punctuality, and no-show behavior.

Very few works consider the appointment delay and the panel sizing problem. Green and Savin (2008) and Liu and Ziya (2013) model the appointment system as a single server queue and take under consideration state dependent no-show behavior. Green and Savin (2008) develop expressions for the appointment delays, by considering both $M/M/1/K$ and $M/D/1/K$ queueing models in steady state, and identify proper panel sizes for medical practices that aim to implement an Open Access policy. Liu and Ziya (2013) model the appointment queue as an $M/M/1$ service system in steady state, and address the problem of taking joint optimal panel sizing and overbooking decisions. Finally, 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 sufficiently higher than the demand, in order to sustain Open Access scheduling.

Balasubramanian et al. (2010, 2012) analyze the tradeoffs 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.

In this paper, we study 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. To the best of our knowledge, our work is the first to explicitly model the two separate time scales involved in accessing outpatient medical care: appointment delay (order of days, weeks) and clinic delay (order of minutes, hours), and to provide explicit expressions for the performance measures of interest based on diffusion approximations. The importance of considering the separate time scales involved in different sectors of healthcare delivery has also been discussed in Armony et al. (2011), Dai and Shi (2013), and Luo et al. (2013). 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 bear unique technical challenges, and allow the study of a system with traffic intensity greater than one.

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. 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 \([0, t] \), 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 day of her scheduled appointment, and the backlog dynamics are unaffected.

Each day, depending on the state 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 clinic arrivals in \([0, t] \), subject to no-show behavior, and \( W^c(t) \) be the workload of the clinic at time \( t \).

We are interested in optimizing the long run average daily net benefit of a medical facility from providing care to patients, with respect to the panel size, \( N \), and the number of offered appointment slots per working day, \( s \). 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. 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 a physician, and an overtime cost \( c_o \) per time unit that the clinic has to operate overtime. The optimization problem under study is

\[
\max_{N,s} \lim_{t \to \infty} \frac{1}{t} \left( rA^c(t) - c_a \int_0^t W^a(\tau)dA^a(\tau) - c_w \int_0^t W^c(\tau)dA^c(\tau) - c_o \int_0^t 1\{W^c(\tau) > 0, \tau \text{ overtime}\}d\tau \right). \tag{1}
\]

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 an analytical 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 intra-day dynamics. In order to obtain a tractable expression for the appointment delay, we develop a heavy traffic diffusion approximation.

The supply of medical appointments is $s$ slots per day. 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 $N \to \infty$.

More generally, we assume that appointment requests arrive to the appointment book according to a renewal process. The 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. Each patient requires service of 1 time slot, and $s$ appointment slots are offered each working day. Each day the appointment book provides the clinic with a schedule of up to $s$ patients.

We model the appointment book as a $GI/D/1$ queue with batch services and state dependent balking. The single server represents the single appointment book, which deterministically provides a schedule to the clinic every day. The realized size of the schedule (batch) though, depends on the state of the appointment queue at the beginning of the working day.

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 balking and one at a time services. 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 $w_i \sim F$, $\forall i \geq 1$. 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 $mw_i$ does not exceed 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.

In order to quantify the waiting time for an appointment, we assume that the service discipline of the appointment book is FIFO. We point out though that the evolution of the appointment backlog remains the same if we relax the FIFO assumption, and consider a work conserving queue instead: patients are offered a slot according to their preferences and subject to availability, and if by the beginning of a working day there are at least $s$ patients in the appointment book, then the number of offered appointments for that day will be equal to $s$. In this case, the offered waiting time is a proxy for how congested is the appointment book.

Let $Q(t)$ denote the appointment backlog at time $t$, which can be expressed recursively as

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

$$= \sum_{i=1}^{A(t)} 1\{\frac{Q(t_i)}{s} < mw_i\} - s[t] + L(t),$$

(2)

where $D(t) := \sum_{i=1}^{\lfloor t \rfloor} \min(s,Q(i))$ is the cumulative number of scheduled appointments in $[0,t]$, including the patients who did not show up, and $L(t) := s[t] - D(t)$ is the cumulative number of unutilized appointment slots in $[0,t]$.

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 requirements:

Assumption 1. (a) $\sqrt{n}(\lambda_n - s) \to \eta \in \mathbb{R}$ as $n \to \infty$.
(b) $m_n s = n$.
(c) $\text{var}(u) := \theta^2 < \infty$.
(d) $F$ is differentiable about $0$ and $F'(0) < \infty$.

Any process associated with system $n$ is superscripted by $n$. Then, we define the fluid-scaled process

$$\bar{A}^n(t) := \frac{A^n(nt)}{n},$$

(3)

and the diffusion-scaled processes

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

$$\tilde{Q}^n(t) := \frac{Q^n(nt)}{\sqrt{n}}$$

$$\text{and } \tilde{L}^n(t) := \frac{L^n(nt)}{\sqrt{n}},$$

(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 and with left limits (RCLL), 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 say that \( X_n \xrightarrow{a.s.} X \) if \( P(\lim_{n \to \infty} ||X_n - X|| = 0) = 1 \). We say that \( X_n \xrightarrow{w} X \) if \( \int \mathcal{R} f dP_n \to \int \mathcal{R} f dP \) as \( n \to \infty \) for every bounded and continuous real function \( f \) defined on \((\mathcal{D}, \mathcal{B})\). Finally, the expressions “\( \overset{a.s.}{=} \)” and “\( \overset{w}{=} \)” will 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)

In (5), \( B(t) \) is a standard Brownian motion, and \( \{I(t) : t \geq 0\} \) is the minimal non-decreasing process such that \( Z(t) \geq 0 \) \( \forall t \geq 0 \) and satisfies \( \int_0^\infty 1\{Z(t) > 0\} dI(t) = 0 \). 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, \((Z(t), I(t)) = (\Phi_\gamma, \Psi_\gamma)(\alpha t + \sigma B(t))\). 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 \((\Phi_\gamma, \Psi_\gamma)\) the reader is referred to Ward and Glynn (2003, 2005).

**Theorem 1.** If \( \tilde{Q}^n(0) \xrightarrow{w} Z(0) \) as \( n \to \infty \), then \( \tilde{Q}^n(t) \xrightarrow{w} Z(t) \) as \( n \to \infty \), where \( Z(t) \) is a regulated Ornstein-Uhlenbeck process with initial position \( Z(0) \), infinitesimal drift \((\eta - sF'(0)z)\), 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 Normal\((\frac{\alpha}{\gamma}, \sigma^2)\) restricted to the interval \([0, \infty)\), i.e. \( Z_\infty \) has density

\[
f_{Z_\infty}(x) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left( -\frac{1}{2} \left( \frac{x - \alpha}{\sigma \gamma} \right)^2 \right), x \geq 0,
\]  (6)
where $\phi(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$ and $\Phi(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$ are the probability density function (pdf) and the cumulative density function (cdf) of a standard Normal random variable respectively. Let $h(y) := \frac{\phi(y)}{1-\Phi(y)}$ be the standard Normal hazard rate. The mean of $Z_{\infty}$ is

$$E[Z_{\infty}] = \frac{\alpha}{\gamma} + \sqrt{\frac{\sigma}{2\gamma}} h\left(-\alpha \sqrt{\frac{2}{\gamma \sigma^2}}\right). \quad (7)$$

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 $\lambda - s - \frac{\lambda F'(0)}{ms} \tilde{q}$, and infinitesimal variance $\lambda \theta^2$. In order to provide an intuition for the proposed approximation, note that for large $n$ and from Theorem 1

$$Q_n(t) \approx \sqrt{n} Z(t/n), \quad (8)$$

where

$$\sqrt{n} Z(t/n) = \sqrt{n} Z(0) + \sqrt{n} \eta n \frac{t}{n} - s F'(0) \int_{0}^{t/n} \sqrt{n} Z(\tau) d\tau + \sqrt{n} \sqrt{s} \theta B(t/n) + \sqrt{n} I(t/n).$$

Therefore, from (8) and (9),

$$Q(t) \approx Q(0) + (\lambda - s) t - \frac{\lambda F'(0)}{ms} \int_{0}^{t} \sqrt{n} Z(t/n) d\tau + \sqrt{n} \sqrt{s} \theta B(t) + \sqrt{n} I(t/n). \quad (10)$$

Suppose that $\lambda > 0$ and $s > 0$. In steady state, the backlog of appointments is approximated as $Q^a$ with pdf

$$f_{Q^a}(x) = \sqrt{\frac{2F'(0)}{ms \theta^2}} \frac{\phi\left(\sqrt{\frac{2F'(0)}{ms \theta^2}} \left(x - \frac{ms(\lambda - s)}{AF'(0)}\right)\right)}{1 - \phi\left(\frac{s - \lambda}{\lambda \sqrt{2msF'(0)\theta^2}}\right)}, x \geq 0, \quad (11)$$

and mean

$$E[Q^a] = \frac{ms(\lambda - s)}{AF'(0)} + \sqrt{\frac{ms \theta^2}{2F'(0)}} h\left(\frac{s - \lambda}{\lambda \sqrt{2msF'(0)\theta^2}}\right). \quad (12)$$

When either $\lambda = 0$ or $s = 0$, then $Q^a = 0$ with probability one.

Based on our diffusion approximation, we can further approximate the probability of balking as

$$P(\text{Balking}) = \int_{0}^{\infty} P(\text{Balking}|Q^a = x) f_{Q^a}(x) dx$$

$$= \int_{0}^{\infty} F\left(\frac{x}{ms}\right) f_{Q^a}(x) dx, \quad (13)$$
and the effective arrival rate as

$$\lambda_{\text{eff}} = \lambda[1 - P(\text{Balking})].$$  \hfill (14) 

Little’s Law provides an approximation for the expected (effective) appointment delay, $$E[W^a] := \frac{E[Q^a]}{\lambda_{\text{eff}}}.$$ 

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

**Lemma 1.** (a) $$E[Q^a]$$ is continuously differentiable in $$\lambda$$ at $$\lambda = 0 \forall s, m > 0.$$  
(b) $$E[Q^a]$$ is strictly increasing in $$\lambda$$ on $$(0, \infty) \forall s, m > 0.$$  
(c) $$E[Q^a]$$ is strictly increasing in $$m$$ on $$(0, \infty) \forall \lambda, s > 0.$$  
(d) $$E[Q^a]$$ is unimodal in $$s$$ on $$[0, \infty) \forall \lambda, m > 0.$$  

**Figure 2** $$E[Q^a]$$ is increasing in $$\lambda$$ and $$m,$$ and unimodal in $$s.$$  
(a) $$0 \leq \lambda \leq 40, \ s = 20.$$  
(b) $$\lambda = 20, \ 0 \leq s \leq 40.$$  

**Note.** The arrival process is Poisson, and the balking threshold is uniformly distributed with mean $$m.$$  

Aligned with our intuition, the average appointment backlog is increasing in the daily demand for medical appointments, $$\lambda.$$ Interestingly, $$E[Q^a]$$ 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. In Figure 2 we plot the expected appointment backlog $$E[Q^a]$$ with respect to $$\lambda$$ and $$s,$$ both in heavy and light traffic, for different patience levels.
4.3. Simulation Experiments

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 uniform patience, i.e. \( F \sim U[0, 2] \), and Poisson arrivals, i.e. \( G \sim \text{Exp}(1) \). In Figure 3 we demonstrate such a comparison for a medical practice where \( \lambda = 33 \) appointment requests per day, \( s = 30 \) offered appointments per day, and for different patience levels. Our simulation experiments confirm that for moderately large values of \( m \) the proposed probability distribution is very accurate.

Figure 3  The distribution of the appointment backlog: a comparison between the frequency histogram of a simulated appointment backlog and the pdf of the proposed approximation.

(a) \( m = 5 \) days  (b) \( m = 10 \) days

(c) \( m = 15 \) days  (d) \( m = 20 \) days

Our next experiment concerns the effective arrival rate, the rate at which patients are assigned to appointment slots. In Table 1 we compare \( \lambda_{\text{eff}} \) from (14) with the actual effective arrival of
a simulated appointment book, for a range of traffic intensities. Our results suggest that, for moderately large patience, $\lambda_{\text{eff}}$ approximates the effective arrival well, and

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

(15)

<table>
<thead>
<tr>
<th>$s$ = 40</th>
<th>$s$ = 45</th>
<th>$s$ = 50</th>
<th>$s$ = 55</th>
<th>$s$ = 60</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>40.00</td>
<td>39.82</td>
<td>44.82</td>
<td>42.92</td>
</tr>
<tr>
<td>4</td>
<td>40.00</td>
<td>40.00</td>
<td>44.98</td>
<td>44.71</td>
</tr>
<tr>
<td>6</td>
<td>40.00</td>
<td>40.00</td>
<td>45.00</td>
<td>44.96</td>
</tr>
<tr>
<td>8</td>
<td>40.00</td>
<td>40.00</td>
<td>45.00</td>
<td>45.00</td>
</tr>
<tr>
<td>10</td>
<td>40.00</td>
<td>40.00</td>
<td>45.00</td>
<td>45.01</td>
</tr>
<tr>
<td>12</td>
<td>40.00</td>
<td>40.00</td>
<td>45.00</td>
<td>45.01</td>
</tr>
<tr>
<td>14</td>
<td>40.00</td>
<td>40.00</td>
<td>45.00</td>
<td>45.01</td>
</tr>
<tr>
<td>16</td>
<td>40.00</td>
<td>40.00</td>
<td>45.00</td>
<td>05.01</td>
</tr>
<tr>
<td>18</td>
<td>40.00</td>
<td>40.01</td>
<td>45.00</td>
<td>45.01</td>
</tr>
<tr>
<td>20</td>
<td>40.00</td>
<td>40.01</td>
<td>45.00</td>
<td>45.01</td>
</tr>
</tbody>
</table>

5. The In-Clinic Queue

In this section we study the in-clinic queue in order to characterize the patients’ waiting times and physicians’ overtime. The evolution of the queue highly depends on how the patients are scheduled to arrive throughout the working day, on patient’s no-show behavior and punctuality, and on the distribution of the service times as well. As discussed in §2, there are no analytical expressions for the patients’ waits and physicians’ overtime in the literature. Only under certain assumptions on the service times distribution (deterministic or exponential), and by assuming punctual patients, recursive expressions can be derived.

We model the in-clinic queue as a $GI/GI/1$ queue. The length of a regular working day is $T$ hours, during which the offered appointments are allocated. Patients do not show up for their scheduled appointment with probability $1 - p \in [0, 1)$. On a particular day, if $\hat{s}$ patients are scheduled to see a physician, then the arrival rate to the clinic is $\lambda_c = \frac{\hat{s}p}{T}$ patients per hour. 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.

5.1. Diffusion Approximation

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 with scheduled appointments, has finite mean $\lambda_c^{-1}$ and finite squared coefficient of variation $c_u^2$. As a 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 \( c_v^2 \).

The workload right before the \( i^{th} \) arrival, denoted as \( W_i \), can be expressed by the classical Lindley recursion as

\[
W_i = \max\{W_{i-1} + v_{i-1} - u_{c,i}, 0\}, \quad \text{for } i \geq 2, \tag{16}
\]

and \( W_1 = 0 \).

As in Chen and Yao (2001) p. 142-144, the workload at time \( t \), denoted as \( W(t) \), can be approximated as \( W(t) \approx Y(t) \), where \( Y(t) \) is a regulated Brownian motion (RBM) with initial position \( Y(0) = 0 \), drift \( \alpha := \lambda_c - \mu \), and infinitesimal variance \( \beta^2 := \lambda_c c_u^2 + \min(\lambda_c, \mu) c_v^2 \), reflected at zero. One representation of \( Y(t) \) is

\[
Y(t) = X(t) + L(t), \tag{19}
\]

where \( L(t) = \sup_{0 \leq s \leq t} [-X(t)] \) and \( X(t) = \alpha t + \beta B(t) \), with \( B(t) \) being a standard Brownian Motion. The representation of \( Y(t) \) in (19) is referred to as the Skorokhod representation. The pdf of \( Y(t) \), as given by Harrison (1985), is

\[
f_{Y(t)}(x) = \frac{2}{\beta \sqrt{\pi t}} \phi\left( \frac{x - \alpha t}{\beta \sqrt{t}} \right) - \frac{2}{\beta^2} e^{2\alpha x} \Phi\left( \frac{x - \alpha t}{\beta \sqrt{t}} \right), \quad x \geq 0. \tag{20}
\]

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 as

\[
\sum_{i=1}^{N_c(T)} W_i = \int_0^T W(t) dN_c(t) \\
= \int_0^T nW(nt) d\frac{N_c(nt)}{n} \\
\approx \int_0^T nW(nt) d\lambda_c t \quad \text{(for large } n \text{ and from FSLLN)} \\
= \lambda_c \int_0^T W(t) dt \\
\approx \frac{2+\mu}{\mu} \int_0^T Y(t) dt =: W^c. \tag{21}
\]
The physician’s overtime is equal to $W(T)$, the workload at the end of the regular working day, and is approximated as

$$O^c := \frac{1}{\mu} Y(T).$$  \hspace{1cm} (22)

In what follows, we derive explicit expressions for $E[W^c]$ and $E[O^c]$.

**Lemma 2.** Assume that $\beta^2 > 0$. Then $E[\int_0^T Y(t) dt] = \begin{cases} \frac{E[Y^2(T)] - \beta^2 T}{\frac{2\alpha}{\beta^2}} & \text{if } \alpha \neq 0 \\ \frac{1}{\mu^2} \beta \sqrt{T} \phi(\frac{\alpha \sqrt{t}}{\beta}) + (\alpha T + \frac{\alpha^2}{\beta^2}) \Phi(\frac{\alpha \sqrt{T}}{\beta}) - \frac{\alpha^2}{2\beta} & \text{if } \alpha = 0, \beta > 0, \alpha \neq 0 \\ \frac{1}{\mu^2} \beta \sqrt{T} \phi(\frac{\alpha \sqrt{T}}{\beta}) + (\alpha T + \frac{\alpha^2}{\beta^2}) \Phi(\frac{\alpha \sqrt{T}}{\beta}) - \frac{\alpha^2}{2\beta} & \text{if } \alpha = 0, \beta > 0, \alpha = 0. \end{cases}$

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 of 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 serve patients for 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 when $\lambda_c > \mu$.

**Lemma 3.** The first two moments of $Y(t)$ are as follows:

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

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

**Lemma 4.** $E[Y(t)]$ and $E[Y^2(t)]$ are continuous in $\alpha$ at $\alpha = 0$.

Note that our proof of Lemma 3 provides an alternative proof for Theorem 1.1 of Abate and Whitt (1987), where $\alpha = -1$ and $\beta = 1$. Now we can express the performance measures of interest as follows.

**Theorem 2.**

$$E[O^c] = \begin{cases} 0 & \text{if } \beta = 0, \alpha \leq 0 \\ \frac{1}{\mu^2} \beta \sqrt{T} \phi(\frac{\alpha \sqrt{T}}{\beta}) + (\alpha T + \frac{\alpha^2}{\beta^2}) \Phi(\frac{\alpha \sqrt{T}}{\beta}) - \frac{\alpha^2}{2\beta} & \text{if } \beta = 0, \alpha > 0 \\ \frac{1}{\mu^2} \beta \sqrt{T} \phi(\frac{\alpha \sqrt{T}}{\beta}) + (\alpha T + \frac{\alpha^2}{\beta^2}) \Phi(\frac{\alpha \sqrt{T}}{\beta}) - \frac{\alpha^2}{2\beta} & \text{if } \beta > 0, \alpha \neq 0 \\ \frac{1}{\mu^2} \beta \sqrt{T} \phi(\frac{\alpha \sqrt{T}}{\beta}) + (\alpha T + \frac{\alpha^2}{\beta^2}) \Phi(\frac{\alpha \sqrt{T}}{\beta}) - \frac{\alpha^2}{2\beta} & \text{if } \beta > 0, \alpha = 0. \end{cases}$$

$$E[W^c] = \begin{cases} 0 & \text{if } \beta = 0, \alpha \leq 0 \\ \frac{2\alpha}{\beta^2} \beta^2 T^2 & \text{if } \beta = 0, \alpha > 0 \\ \frac{2\alpha}{\beta^2} \beta^2 T^2 & \text{if } \beta > 0, \alpha \neq 0 \\ \frac{2\alpha}{\beta^2} \beta^2 T^2 & \text{if } \beta > 0, \alpha = 0. \end{cases}$$

**Lemma 5.** $E[W^c]$ and $E[O^c]$ are continuous in $\alpha$ at $\alpha = 0$. 
Our diffusion approximation for the in-clinic queue in transient state is tractable, and provides explicit expressions for the performance measures of interest. It also captures patients’ non-punctuality and no-shows (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 $\mu^{-1} = 15.5$ 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, as demonstrated by our simulation experiments, under both light and heavy traffic. In Figure 4 we compare the expressions of $E[W^c]$ and $E[O^c]$ in Theorem 2, with the corresponding performance measures of a simulated clinic, for different traffic intensities and for different levels of service times’ variability. For low values of the $c_v$ (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.

5.3. An Extension to Account 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 renewal processes $N_w(t) = \max\{k \geq 0 : \sum_{i=1}^k u_{w,i} \leq t\}$. 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 $$\frac{\lambda_c^2 + \lambda_w^2}{(\lambda_c + \lambda_w)^2}.$$ 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 $\hat{\beta}^2 := \lambda_c c_u^2 + \lambda_w c_w^2 + \min(\lambda_c + \lambda_w, \mu) c_v^2$. For the rest of the paper we consider that there are no walk-ins.
Figure 4  Physician’s overtime and aggregate waiting time: 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) $\mu = 4$, $\lambda_c \in \{3, 4, 5\}$, $0 \leq c_v \leq 1$.

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

(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.

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 $N$, and therefore it is equivalent to perform the optimization with respect to $\lambda$ and $s$. 
As a reminder, we consider a reward $r > 0$ generated per patient served, 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.

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 backlog in (11), 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

$$R(\lambda, s) := rp\lambda \alpha - c_a \lambda \alpha - c_w E[Q^a] - c_w E[Q^c] - c_o E[O^c],$$

where $E[Q^a]$ is as in (12), and $E[Q^c]$ and $E[O^c]$ are as in Theorem 2, with $\alpha = \lambda_c - \mu$ and $\beta^2 = \lambda_c c_w^2 + \min(\lambda_c, \mu) c_w^2$. We consider the optimization problem

$$\max_{\lambda, s} R(\lambda, s)$$

s.t. $\lambda, s \leq M$

$$\lambda, s \geq 0.$$ (P)

The constraint $\lambda, s \leq M$ ensures that demand and supply for appointments cannot be arbitrarily large. Lemmas 1(a) and 5, 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 (15) and for the sake of tractability, we are able to provide a neat characterization of the optimal solution.

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

(a) $\lambda^* = 0$ and $\frac{\partial R}{\partial \lambda}(0, s^*) \leq 0$.

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

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

Under the assumptions 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. Either:
(a) It is not beneficial for the clinic to maintain a panel of patients and schedule appointments at all, or

(b) The clinic offers as many appointment slots as possible, and the optimal daily demand for medical appointments depends on the objective function’s coefficients and the clinic’s characteristics, or

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

In our extensive numerical experiments the optimal pair \((\lambda^*, s^*)\) almost always lies within regime (b), which we refer to as the “Open Access” regime, since the clinic offers as many appointment slots as possible. One scenario that yields a solution in regime (c), which we refer to as the “Limited Access” regime, is when the average service time is large, and the patients are very sensitive to the appointment delay. In practice, such a scenario could correspond to a clinic that performs long, complex, urgent procedures.

Throughout our numerical/simulation experiments, we normalize the objective function with respect to \(c_w\), i.e., \(c_w = 1\). Following Robinson and Chen (2010), we consider an overtime cost coefficient which is five times as much as the patients’ waiting cost coefficient, i.e., \(c_o = 5\). Further, we regard that the cost of waiting one day for an appointment is in the same order as the cost of waiting one hour in the clinic, i.e., \(c_a \sim c_w\). Finally, the average reward generated per hour is in the same order as the cost of staying one hour overtime in the clinic, i.e., \(r\mu \sim c_o\).

For fixed \(s\), the objective function appears to be concave in \(\lambda\) on \([0, s]\) as demonstrated by our numerical analysis. This suggests that a first order condition would uniquely determine the approximated global maximum. Figure 5(a) demonstrates the behavior of the objective function with respect to \(\lambda\) for different values of \(r\), while the cost coefficients are kept constant. Figure 5(b) demonstrates the optimal solution as given by the MATLAB R2012b optimization toolbox. It is evident that the no-show probability has a significant, non-linear effect on the panel sizing decision and should be taken under consideration. Further, as \(r\) increases, with values greater than \(\frac{c_o}{\mu}\), the arrival rate increases in a concave manner towards heavy traffic, i.e., \(\lambda \rightarrow s\). For reasonable values of \(r\), the system operates in light traffic, and hence an Open Access policy (satisfying today’s demand today) is evidently optimal.

6.2. A Simulation Experiment

In order to characterize the optimal solution in Theorem 3, we assumed that the arrival rate to the clinic from scheduled appointments is the same every day and equal is to \(p \times \min(\lambda, s)\) patients per day. In Figure 6 we compare the optimal pair \((\lambda, s)\) for a simulated system with the one obtained under the assumptions of Theorem 3, for different cost/reward coefficients. Our experimental setup is as follows:
Figure 5  Numerical analysis of the objective function and the optimal solution.

(a) Objective as function of $\lambda$, $p = 0.9$, $r \in \{0.5, 1, 2, 3, 4, 5\}$.

(b) Optimal $\lambda$ as a function of $r$, $p \in \{0.7, 0.8, 0.9, 1.0\}$.

Note. The objective function appears to be concave in $\lambda$ on $[0, s]$, as suggested by our numerical experiments. For this example: $s^* = M = 30$, the balking threshold is uniformly distributed between 0 and 10 days, the arrivals to the appointment book and the clinic are Poisson, $T = 8$ hours, $c_a = c_w = \frac{1}{2}c_o = 1$, $\mu = 3$ patients per hour (24 patients per day), $c_v = 0.325$.

- **Appointment book**: Requests for 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 of $m = 5$ days.

- **Clinic**: The length of the working day is $T = 8$ hours. Depending on the state of the appointment book at the beginning of day $k$, $\lambda_k$ patients are scheduled to arrive, $1 \leq k \leq 10000$. There is a no-show rate $1 - p$. As argued earlier in this paper, the arrival process depends on how the patients are scheduled to arrive throughout the working day and on their punctuality. For our experimental setup, we consider a Poisson arrival process from scheduled appointments with rate $\frac{p\lambda_k}{T}$ per hour. Finally, the service times are iid, lognormally distributed with mean 20 minutes and coefficient of variation 0.325.

We observe that indeed the optimal solution lies within the Open Access regime for both systems, and further, our approximated solution to the panel sizing decision is close to the one that optimizes the simulated system.

7. Conclusion

We study 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. 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). We
analyze the two queueing systems associated with each type of delay, and provide explicit expressions for the performance measures of interest based on diffusion approximations. In our analysis we capture many features of the complex reality of the healthcare system, including patients’ non-punctuality, no-shows, and stochastic service times. Two additional distinctive characteristics of this study are the balking behavior of the patients who face long appointment backlogs, and the transient-state analysis of the in-clinic queue, which allow the study of a system with traffic intensity greater than one and bear unique technical challenges.

Concerning the panel sizing and appointment scheduling decisions, our analysis provides theoretical and numerical support that either an “Open Access” policy is optimal, or supply and demand are perfectly matched and are both very small (“Limited Access” regime). Under our Open Access regime, the clinic offers as many appointment slots as possible per day, and the optimal panel size depends on the cost structure and the clinic’s characteristics. A solution within the Limited Access regime arises when the service times are long, and the patients are very sensitive to the appointment delay.

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), Dreier et al. (2008), Norris et al. (2012)). 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 expression for $\gamma(k)$ is given in Green and Savin (2008): $\gamma(k) = \gamma_{\text{max}} - (\gamma_{\text{max}} - \gamma_{\text{min}}) e^{-\frac{k}{C}t}^{-1}$, 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.

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 (2013)). Such a clinic with **front loaded** schedules can be approximated as follows: the working day is partitioned in 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 reflected diffusion with a piecewise constant drift, and a piecewise constant infinitesimal variance. Further, it is of interest to analyze 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.

A careful treatment of the emergency walk-ins is of interest as well. As demonstrated in §5.3, the analysis of the in-clinic queue can be readily adjusted to capture emergency walk-ins, given that we know their arrival rate. A relationship should be established first between the walk-in rate, and 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. An optimal appointment assignment rule may be developed for incoming requests, 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 (for example flu) on the demand for medical care.

**Appendix. Proofs**

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

\[
M(i) := \sum_{j=1}^{i} \left[ \mathbf{1}\left\{ \frac{Q(t_j^-)}{s} \geq mw_j \right\} - E\left( \mathbf{1}\left\{ \frac{Q(t_j^-)}{s} \geq mw_j \right\} | \mathcal{F}_{j-1} \right) \right]
\]

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

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

\[
Q(t) + \int_0^t F\left( \frac{Q(\tau^-)}{s \nu} \right) dA(\tau) = A(t) - M(A(t)) - s [t] + L(t).
\]
Consider further the diffusion scaled process $\tilde{M}^n(t) := \frac{M^n(nt)}{\sqrt{n}}$. From the pathwise equation for the backlog of appointments in (25), Assumption 1, the scaling in (3) and (4), and some algebra, we obtain

$$\tilde{Q}^n(t) + sF'(0) \int_0^t \tilde{Q}^n(\tau)d\tau = \tilde{X}^n(t) + \tilde{L}^n(t),$$

(26)

where

$$\tilde{X}^n(t) = \tilde{A}^n(t) + \frac{1}{\sqrt{n}}(n\lambda_n t - s|nt|) - \tilde{M}^n(\tilde{A}^n(t)) + sF'(0) \int_0^t \tilde{Q}^n(\tau)d\tau - \int_0^t \sqrt{n}F(\frac{\tilde{Q}^n(\tau)}{\sqrt{n}})d\tilde{A}^n(\tau).$$

(27)

To provide an intuition for the representation of the workload in (26), note that from L'Hôpital's rule $\lim_{y \to \infty} yF(\frac{y}{x}) = F'(0)x$, so that

$$\frac{1}{\sqrt{n}} \int_0^{nt} F(\frac{Q^n(\tau)}{\max})dA(\tau) = \int_0^t \sqrt{n}F(\frac{\tilde{Q}^n(\tau)}{\sqrt{n}})d\tilde{A}^n(\tau) \approx \int_0^t sF'(0)\tilde{Q}^n(\tau)d\tau \quad \text{(from (29)).}$$

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 $\gamma = sF'(0)$ 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,

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

(28)

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

(29)

and

$$\frac{1}{\sqrt{n}}(n\lambda_n t - s|nt|) \to \eta t,$$

(30)

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

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

(31)

and

$$\int_0^t sF'(0)\tilde{Q}^n(\tau)d\tau - \int_0^t \sqrt{n}F(\frac{\tilde{Q}^n(\tau)}{\sqrt{n}})d\tilde{A}^n(\tau) \Rightarrow 0.$$  

(32)

Combining (27), (28), (29), (30), (31), (32) 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 $\gamma = sF'(0)$. □
Proof of Lemma 1  
(a) Suppose that \( m, s > 0 \). Firstly we show that \( \lim_{\lambda \to 0} E[Q^a] = 0 \).

\[
\begin{align*}
\lim_{\lambda \to 0} E[Q^a] &= \lim_{\lambda \to 0} \left[ \frac{ms(\lambda-s)}{\lambda F(0)^{s}} + \sqrt{\frac{ms^2}{2F(0)^s}} h \left( \frac{\lambda-s}{\lambda} \sqrt{2ms F'(0)^{s}} \right) \right] \\
&= \lim_{\lambda \to 0} \left[ \frac{ms(\lambda-s)}{\lambda F(0)^{s}} + \sqrt{\frac{ms^2}{2F(0)^s}} \cdot 2m s \Phi \left( \frac{\lambda-s}{\lambda} \sqrt{2ms F'(0)^{s}} \right) \right] \\
&= \lim_{\lambda \to 0} \left[ \frac{ms(\lambda-s)}{\lambda F(0)^{s}} + \sqrt{\frac{ms^2}{2F(0)^s}} \cdot 2m s \Phi \left( \frac{\lambda-s}{\lambda} \sqrt{2ms F'(0)^{s}} \right) \right] \\
&= \lim_{\lambda \to 0} \left[ \frac{ms(\lambda-s)}{\lambda F(0)^{s}} + \sqrt{\frac{ms^2}{2F(0)^s}} \cdot 2m s \Phi \left( \frac{\lambda-s}{\lambda} \sqrt{2ms F'(0)^{s}} \right) \right] \\
&= 0,
\end{align*}
\]

where (33) comes from the fact that \( \lim_{\lambda \to 0} \Phi \left( \frac{\lambda-s}{\lambda} \sqrt{2ms F'(0)^{s}} \right) = 0 \) and L'Hôpital's rule.

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

\[
\lim_{\lambda \to 0} \frac{\partial E[Q^a]}{\partial \lambda} = \lim_{\lambda \to 0} \left[ \frac{ms}{F'(0)^s} \cdot \frac{s}{2} - \frac{ms}{F'(0)^s} \cdot \frac{s}{2} h'(y) \left( \frac{\lambda-s}{\lambda} \right) \sqrt{2ms F'(0)^{s}} \right] \\
= \lim_{\lambda \to 0} \left[ \frac{ms}{F'(0)^s} \cdot \frac{s}{2} - \frac{ms}{F'(0)^s} \cdot \frac{s}{2} \right] \\
= 0.
\]

(b) Suppose that \( \lambda, m, s > 0 \). Then \( E[Q^a] = \sqrt{\frac{ms^2}{2F(0)^s}} [h(y) - y] \), where \( y := \sqrt{\frac{2ms}{F'(0)^s}} \frac{(s-\lambda)}{\lambda} \). Therefore

\[
\frac{\partial E[Q^a]}{\partial \lambda} = \sqrt{\frac{ms^2}{2F(0)^s}} \frac{\partial}{\partial \lambda} [h'(y) - 1] \\
= - \sqrt{\frac{ms^2}{2F(0)^s}} \frac{s}{2} \Phi \left[ h'(y) - 1 \right] \\
= \frac{m s^2}{F'(0)^s} \left[ 1 - h'(y) \right].
\]

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

(c) Suppose that \( \lambda, m, s > 0 \). Then \( E[Q^a] = \sqrt{\frac{ms^2}{2F(0)^s}} [h(y) - y] \), where \( y := \sqrt{\frac{2ms}{F'(0)^s}} \frac{(s-\lambda)}{\lambda} \). Therefore

\[
\frac{\partial E[Q^a]}{\partial m} = \frac{1}{2\sqrt{m}} \sqrt{\frac{ms^2}{2F(0)^s}} [h(y) - y] + \sqrt{\frac{ms^2}{2F(0)^s}} \frac{\partial h}{\partial m} [h'(y) - 1].
\]

From (34) and some algebra, the condition \( \frac{\partial E[Q^a]}{\partial m} > 0 \) is true if and only if \( h(y) - 2y + yh^2(y) - y^2h(y) > 0 \). 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

\[
\left[ xh(x) \left[ \frac{1}{h(x)} \right]' \right]' < 0
\]
Then where (40) follows from (38) and (39). Barrow and Cohen (1954) proved that inequality (41) holds.

Lemma 6. where (35) follows from the fact the $h(x) > 0$ for all $x \in \mathbb{R}$, and (36) follows from:

$$h'(x) = h(x)[h(x) - x]$$
and
$$h''(x) = h(x)[h(x) - x]^2 + h^2(x)[h(x) - x] - h(x).$$

(d) For fixed $\lambda, m > 0$, it suffices to show that every critical point of $E[Q^a]$ (with respect to $s$) is a local maximum, i.e., $\frac{\partial E[Q^a]}{\partial s} < 0$ whenever $\frac{\partial E[Q^a]}{\partial s} = 0$. Then, since $E[Q^a]$ is continuous (and differentiable) in $s$, there can be at most one local maximum, concluding that $E[Q^a]$ is unimodal in $s$. Firstly, we prove the following intermediate result:

**Lemma 6.** Let $h(x) := \frac{\Phi(x)}{1 - \Phi(x)}$ be the hazard rate of a standard Normal random variable. Then

$$\frac{h''(x)[h(x) - x]}{[h'(x) - 1]^2} < 2$$
for all $x \in \mathbb{R}$.

**Proof of Lemma 6**

$$\frac{h''(x)[h(x) - x]}{[h'(x) - 1]^2} < 2$$
$$\iff \frac{h(x)[h(x) - x]^3 + h^2(x)[h(x) - x]^2 - h(x)[h(x) - x]}{1 - 2h(x)[h(x) - x] + h^2(x)[h(x) - x]^2} < 2$$
$$\iff 0 < -h(x)[h(x) - x]^3 + h^2(x)[h(x) - x]^2 - 3h(x)[h(x) - x] + 2,$$
where (40) follows from (38) and (39). Barrow and Cohen (1954) proved that inequality (41) holds.

Recall that for $\lambda, s > 0$

$$E[Q^a] = \sqrt{\frac{\lambda}{2\pi s}} \sqrt{s}[h(y) - y],$$
where $y := \sqrt{\frac{2m}{\lambda s}} \sqrt{\frac{\lambda}{\pi}}.$

Then

$$\frac{\partial y}{\partial s} = \sqrt{\frac{2m}{\pi s \lambda}} \left( \frac{3\sqrt{s}}{2s} - \frac{1}{2\sqrt{s}} \right) = \frac{3\sqrt{s}}{2s} + \frac{1}{\sqrt{s} \lambda} \sqrt{\frac{2m}{\pi s \lambda}},$$
$$\frac{\partial^2 y}{\partial s^2} = \frac{1}{2s^2} \frac{\partial y}{\partial s} - \frac{1}{2s} \frac{\partial^2 y}{\partial s^2} \sqrt{\frac{2m}{\pi s \lambda}},$$
and
$$\frac{\partial E[Q^a]}{\partial s} = \sqrt{\frac{\lambda}{2\pi s}} \left[ -h(y) - y \right] - \sqrt{\frac{\lambda}{\pi}} [1 - \Phi(y)] \frac{\partial y}{\partial s}. \tag{46}$$
From (46), whenever $\frac{\partial E[Q^n]}{\partial s} = 0$ then

$$\frac{\partial u}{\partial s} = \frac{h(y) - y}{2s[1 - h'(y)]},$$

and

$$\sqrt{2m\frac{P(0)}{\partial s} \frac{Q}{\partial s}} = \frac{|h(y) - y| - 3y[1 - h'(y)]}{2\sqrt{2s[1 - h'(y)]}} = \frac{|h(y) - y|}{2\sqrt{2s[1 - h'(y)]}} - 3y.$$  

(47)

(48)

From (45), (47) and (48), whenever $\frac{\partial E[Q^n]}{\partial s} = 0$ then

$$\frac{\partial^2 y}{\partial s^2} = \frac{3|h(y) - y|}{4s^2[1 - h'(y)]} - \frac{3y}{2s^2} - \frac{|h(y) - y|}{2s^2[1 - h'(x)]} + \frac{3y}{2s^2}$$

$$= \frac{2|h(y) - y[1 - h'(y)]}{4s^2[1 - h'(y)]}.$$  

(49)

Then, from (46), (47), and (49), whenever $\frac{\partial E[Q^n]}{\partial s} = 0$ then

$$\frac{\partial^2 E[Q^n]}{\partial s^2} = \sqrt{\frac{m\partial^2}{2P(0)}} \left[ - \frac{1}{4s^2} [h(y) - y] - \frac{h(y) - y}{2s[1 - h'(y)]} - \frac{1}{2s^2} [1 - h'(y)] \frac{h(y) - y}{2s[1 - h'(y)]} \
+ \sqrt{s} h''(y) \frac{|h(y) - y|}{2s^2[1 - h'(y)]^2} - \sqrt{s} [1 - h'(y)] \frac{2|h(y) - y| - 3y[1 - h'(y)]}{4s^2[1 - h'(y)]^2} \right]$$

$$= \sqrt{\frac{m\partial^2}{2P(0)}} \frac{1}{4s^2} \left[ - 5[h(y) - y] + 3y[1 - h'(y)] + \frac{h''(y)|h(y) - y|^2}{1 - h'(y)^2} \right]$$

$$< \sqrt{\frac{m\partial^2}{2P(0)}} \frac{3}{4s^2} \left[ h(y) - 2y + yh^2(y) - y^2h(y) \right]$$

(50)

(51)

(52)

Inequality (50) follows from Lemma 6 and the fact that $h(x) > x$ for all $x \in \mathbb{R}$. Equality (51) follows from (38). Inequality (52) follows directly from (37). Therefore, $\frac{\partial^2 E[Q^n]}{\partial s^2} < 0$ whenever $\frac{\partial E[Q^n]}{\partial s} = 0$. 

\[\square\]

Proof of Lemma 2  
By Ito formula (see Theorem 3.3 of Karatzas and Shreve (1991), p.149), we get

$$f(Y(T)) = f(Y(0)) + \int_0^T f'(Y(t))dX(t) + \int_0^T f'(Y(t))dL(t) + \frac{1}{2} \int_0^T f''(Y(t))d < X >_t$$

$$= f(0) + \alpha \int_0^T f'(Y(t))dt + \beta \int_0^T f'(Y(t))dB(t) + \int_0^T f'(Y(t))dL(t) + \frac{\partial^2}{2} \int_0^T f''(Y(t))dt,$$

where $f$ is a twice continuously differentiable function. Firstly assume that $\alpha \neq 0$. For $f(x) := x^2$ we get

$$Y^2(T) = 2\alpha \int_0^T Y(t)dt + 2\beta \int_0^T Y(t)dB(t) + 2 \int_0^T Y(t)dL(t) + \beta^2 \int_0^T dt$$

$$= 2\alpha \int_0^T Y(t)dt + 2\beta \int_0^T Y(t)dB(t) + \beta^2 T.$$  

(53)
From (53) we get that \( Y^2(T) - 2\alpha \int_0^T Y(t) dt - \beta^2 T = 2\beta \int_0^T Y(t) dB(t) \), a martingale, concluding that \( E \left[ \int_0^T Y(t) dt \right] = \frac{EY^2(T)}{2\alpha} - \beta^2 T. \)

Next, assume that \( \alpha = 0 \). For \( f(x) := x^3 \) we get

\[
Y^3(T) = 3\beta \int_0^T Y^2(t) dB(t) + 3 \int_0^T Y^2(t) dL(t) + 3\beta^2 \int_0^T Y(t) dt
\]

\[
= 3\beta \int_0^T Y^2(t) dB(t) + 3\beta^2 \int_0^T Y(t) dt. \tag{54}
\]

From (54) we get that \( Y^3(T) = 3\beta^2 \int_0^T Y(t) dt \), a martingale, concluding that \( E \left[ \int_0^T Y(t) dt \right] = \frac{EY^3(T)}{3\beta^2}. \) \( \square \)

**Proof of Lemma 3**  When \( \beta = 0 \) there is no variability, and therefore \( Y(t) = \max(0, t) \) with probability one.

Then consider the case where \( \beta > 0 \). First, suppose that \( \alpha \neq 0 \) and let \( g(x) := \frac{2\alpha}{\beta^2} e^{2\alpha x} \Phi \left( \frac{x-\alpha t}{\beta \sqrt{t}} \right) \). From the probability density of \( Y(t) \) in (20) we get

\[
\int_0^\infty g(x) dx = \int_0^\infty \left( \frac{2}{\beta \sqrt{t}} \phi \left( \frac{x-\alpha t}{\beta \sqrt{t}} \right) \right) dx = 2\Phi \left( \frac{\sqrt{2} \alpha t}{\beta} \right) - 1. \tag{55}
\]

By definition,

\[
E[Y(t)] = \int_0^\infty x f_Y(x) dx = \int_0^\infty \frac{2x}{\beta \sqrt{t}} \phi \left( \frac{x-\alpha t}{\beta \sqrt{t}} \right) dx - \int_0^\infty x g(x) dx. \tag{56}
\]

For the first integral in (56) we use the transformation \( u = \frac{x-\alpha t}{\beta \sqrt{t}} \) to get

\[
\int_0^\infty \frac{2x}{\beta \sqrt{t}} \phi \left( \frac{x-\alpha t}{\beta \sqrt{t}} \right) dx = 2\beta \sqrt{t} \phi \left( \frac{\alpha \sqrt{t}}{\beta} \right) + 2\alpha t \Phi \left( \frac{\alpha \sqrt{t}}{\beta} \right). \tag{57}
\]

Then, we note that

\[
g(x) = \frac{\beta^2}{2\alpha} g'(x) + \frac{1}{\beta \sqrt{t}} \phi \left( \frac{x-\alpha t}{\beta \sqrt{t}} \right), \tag{58}
\]

and therefore

\[
\int_0^\infty x g(x) dx = \int_0^\infty \frac{\beta^2}{2\alpha} g'(x) dx + \int_0^\infty \frac{1}{\beta \sqrt{t}} \phi \left( \frac{x-\alpha t}{\beta \sqrt{t}} \right) dx. \tag{59}
\]

The second integral in (59) is the same as the one in (57), divided by 2. For the first integral in (59) we apply integration by parts to get

\[
\int_0^\infty \frac{\beta^2}{2\alpha} g'(x) dx = \lim_{x \to \infty} \frac{\beta^2}{2\alpha} g(x) - \int_0^\infty \frac{\beta^2}{2\alpha} g(x) dx \tag{60}
\]

\[
= -\frac{\beta^2}{2\alpha} \left( 2\Phi \left( \frac{\alpha \sqrt{t}}{\beta} \right) - 1 \right). \tag{61}
\]

The equality in (61) comes from the fact that \( \lim_{x \to \infty} \frac{\beta^2}{2\alpha} g(x) = 0 \) and from (55). If we combine (56), (57), (59), (61) we get

\[
E[Y(t)] = \beta \sqrt{t} \phi \left( \frac{\alpha \sqrt{t}}{\beta} \right) + (\alpha t + \frac{\beta^2}{2\alpha}) \Phi \left( \frac{\alpha \sqrt{t}}{\beta} \right) - \frac{\beta^2}{2\alpha}. \]
We apply similar methods to compute the second moment

$$E[Y^2(t)] = \int_0^\infty x^2 f_Y(x)dx = \int_0^\infty \frac{2x^2}{\beta \sqrt{\pi}} \phi\left(\frac{x - \alpha t}{\beta \sqrt{t}}\right)dx - \int_0^\infty x^2 g(x)dx. \quad (62)$$

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

$$\int_0^\infty \frac{2x^2}{\beta \sqrt{\pi}} \phi\left(\frac{x - \alpha t}{\beta \sqrt{t}}\right)dx = 2\alpha \beta t \sqrt{\pi} \Phi\left(\frac{\alpha \sqrt{t}}{\beta}\right) + 2(\beta^2 t + \alpha^2 t^2) \Phi\left(\frac{\alpha \sqrt{t}}{\beta}\right). \quad (63)$$

For the second integral in (62), we make use of (58) to get

$$\int_0^\infty x^2 g(x)dx = \int_0^\infty \frac{\beta^2 x^2}{2 \alpha} g(x)dx + \int_0^\infty \frac{x^2}{\beta \sqrt{\pi}} \phi\left(\frac{x - \alpha t}{\beta \sqrt{t}}\right)dx. \quad (64)$$

The second integral in (64) is the same as the one in (63), divided by 2. For the first integral in (64) we apply integration by parts to get

$$\int_0^\infty \frac{\beta^2 x^2}{2 \alpha} g(x)dx = \lim_{x \to \infty} \frac{\beta^2 x^2}{2 \alpha} g(x) - \int_0^\infty \frac{\beta^2 x}{\alpha} g(x)dx = \frac{\beta^2}{2 \alpha} \left(2 \Phi\left(\frac{\alpha \sqrt{t}}{\beta}\right) - 1\right) - \frac{\beta^2}{\alpha^2} \Phi\left(\frac{\alpha \sqrt{t}}{\beta}\right) - \beta^2 t \Phi\left(\frac{\alpha \sqrt{t}}{\beta}\right). \quad (65)$$

The equality in (66) comes from the fact that $\lim_{x \to \infty} \frac{\beta^2 x^2}{2 \alpha} g(x) = 0$ (applying L'Hôpital’s rule) and from (59). If we combine (62), (63), (64), (66) we get

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

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

Proof of Lemma 4 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 $\lim_{\alpha \to 0} E[Y(t)] = \frac{2\alpha \sqrt{t}}{\beta \sqrt{2\pi}}$ and that $\lim_{\alpha \to 0} E[Y^2(t)] = \beta^2 t$.

$$\lim_{\alpha \to 0} E[Y(t)] = \lim_{\alpha \to 0} \left[ \beta \sqrt{t} \phi\left(\frac{\alpha \sqrt{t}}{\beta}\right) + (\alpha t + \frac{\beta^2}{\alpha}) \Phi\left(\frac{\alpha \sqrt{t}}{\beta}\right) - \frac{\beta^2}{\alpha^2}\right]$$

$$= \frac{\beta \sqrt{t}}{\sqrt{2\pi}} + \beta^2 \frac{\phi\left(\frac{\alpha \sqrt{t}}{\beta}\right)}{\alpha} - \frac{\beta^2}{\alpha^2}$$

$$= \frac{\beta \sqrt{t}}{\sqrt{2\pi}} + \beta^2 \Phi\left(\frac{\alpha \sqrt{t}}{\beta}\right) - \frac{\beta^2}{\alpha^2} \quad (67)$$

$$\lim_{\alpha \to 0} E[Y^2(t)] = \lim_{\alpha \to 0} \left[ \beta \sqrt{t} (\alpha t + \frac{\beta^2}{\alpha}) \phi\left(\frac{\alpha \sqrt{t}}{\beta}\right) + (2\beta^2 t + \alpha^2 t^2 - \frac{\beta^4}{\alpha^2}) \Phi\left(\frac{\alpha \sqrt{t}}{\beta}\right) + \frac{\beta^4}{\alpha^4}\right]$$

$$= \lim_{\alpha \to 0} \left[ \beta \sqrt{t} \phi\left(\frac{\alpha \sqrt{t}}{\beta}\right) + \frac{2\beta^2 t}{\alpha} \Phi\left(\frac{\alpha \sqrt{t}}{\beta}\right) \right] + \beta^2 t$$

$$= \beta^2 t. \quad (68)$$

Equations (67) and (68) follow from the fact $\lim_{\alpha \to 0} \Phi\left(\frac{\alpha \sqrt{t}}{\beta}\right) - \frac{1}{2} = 0$ and L'Hôpital’s rule. \square
Proof of Theorem 2  Consider the case where $\beta > 0$ and $\alpha = 0$. We apply integration by parts to get
\[
E[Y^3(t)] = \int_0^\infty \frac{2t^3}{\beta \sqrt{2\pi}} e^{-\frac{t^2}{2\beta^2}} dt
\]
\[
= \frac{4\beta^3 \sqrt{2}}{\sqrt{2\pi}} .
\]
(69)

The rest of the proof follows from Lemmas 2 and 3 and equation (21).

Proof of Lemma 5  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 $\lim_{\alpha \to 0} E[W^c] = \frac{4\beta T \sqrt{T}}{3 \sqrt{2\pi}}$, and that $\lim_{\alpha \to 0} E[O^c] = \frac{2\beta \sqrt{T}}{3 \sqrt{2\pi}}$.

\[
\lim_{\alpha \to 0} E[W^c] = \lim_{\alpha \to 0} \left[ \frac{1}{\alpha^3} \left[ (\beta \alpha^2 T + 2\beta^2 \alpha T^2 + \alpha^4 T^2) \phi(\frac{\alpha T}{\beta}) + (2\beta^2 \alpha^2 T + 2\beta^2 \alpha T^2) \phi(\frac{\alpha T}{\beta}) - 2\beta^2 \alpha T \right] \right]
\]
\[
= \lim_{\alpha \to 0} \left[ \frac{1}{\alpha^3} \left[ 4\beta^2 T \sqrt{T} \phi(\frac{\alpha T}{\beta}) + 4\beta^2 T \phi(\frac{\alpha T}{\beta}) \right] \right]
\]
\[
= \lim_{\alpha \to 0} \left[ \frac{1}{\alpha^3} \left[ 8\beta^2 T \sqrt{T} \phi(\frac{\alpha T}{\beta}) \right] \right]
\]
\[
= \frac{4\beta^3 \sqrt{T}}{3 \sqrt{2\pi}} .
\]

\[
\lim_{\alpha \to 0} E[O^c] = \lim_{\alpha \to 0} \left[ \frac{1}{\alpha^3} E[Y(T)] \right]
\]
\[
= \frac{2\beta \sqrt{T}}{3 \sqrt{2\pi}} .
\]

Equation (70) comes from L’Hôpital’s rule and some algebra.

Proof of Theorem 3  Suppose that $(\lambda^*, s^*)$ is an optimal solution to (P). We denote the feasible region with $D := \{ (\lambda, s) : \lambda \leq M, \lambda \geq 0, \lambda \geq 0 \} \subset \mathbb{R}^2$, 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.,
\[
\max_{\lambda, s} R(\lambda, s)
\]
\[
s.t. \hspace{1cm} s \leq \lambda
\]
\[
\lambda \leq M
\]
\[
s \geq 0
\]
(P1)

Under the assumption that $\lambda_0 = \frac{\min(\lambda, s)}{T}$ with probability one, the objective function and its partial derivatives can be written as
\[
R = rp s - c_a E[Q^a] - c_w E[W^c] - c_o E[O^c]
\]
\[
\frac{\partial R}{\partial \lambda} = -c_a \frac{\partial E[Q^a]}{\partial \lambda} - c_w \frac{\partial E[W^c]}{\partial s} - c_o \frac{\partial E[O^c]}{\partial s}
\]
\[
\frac{\partial R}{\partial s} = rp - c_a \frac{\partial E[Q^a]}{\partial \lambda} - c_w \frac{\partial E[W^c]}{\partial s} - c_o \frac{\partial E[O^c]}{\partial s}
\]
The first order conditions for optimality are

\[ \frac{\partial R}{\partial s} + \xi_1 - \xi_2 = 0 \]  
\[ \frac{\partial R}{\partial s} - \xi_1 + \xi_3 = 0 \]  
\[ s - \lambda \leq 0, \lambda - M \leq 0, -s \leq 0 \]  
\[ \xi_1(s - \lambda) = \xi_2(\lambda - M) = \xi_3 s = 0 \]  
\[ \xi_1, \xi_2, \xi_3 \geq 0 \]  
\[ \frac{\partial E[Q^a]}{\partial s} = 0, \frac{\partial E[W^c]}{\partial s} = 0, \frac{\partial E[O^c]}{\partial s} = 0 \]  

Suppose (for contradiction) that \( s < \lambda \). From (74) we have that \( \xi_1 = 0 \), and (71) implies that \( c_a \frac{\partial E[Q^a]}{\partial s} + \xi_2 = 0 \), which is a contradiction, since from Lemma 1(b) we have that \( \frac{\partial E[Q^a]}{\partial s} > 0 \). Therefore, if \( (\lambda^*, s^*) \in D_1 \) then \( s^* = \lambda^* \), concluding that \( (\lambda^*, s^*) \) is in \( D_2 \).

Consider now the optimization problem

\[ \max_{\lambda, s} R(\lambda, s) \]  
\[ \text{s.t. } \lambda \leq s \]  
\[ s \leq M \]  
\[ \lambda \geq 0 \]  

(P2)

The objective function and its partial derivatives can be written as

\[ R = rp\lambda - c_a E[Q^a] - c_w E[W^c] - c_o E[O^c] \]  
\[ \frac{\partial R}{\partial s} = rp - c_a \frac{\partial E[Q^a]}{\partial s} - c_w \frac{\partial E[W^c]}{\partial s} - c_o \frac{\partial E[O^c]}{\partial s} \]  
\[ \frac{\partial R}{\partial s} = -c_a \frac{\partial E[Q^a]}{\partial s}. \]

The first order conditions for optimal solution to (P2) are

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

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

\{ (\lambda, s) : 0 < \lambda < s < M \}, \{ (\lambda, s) : 0 = \lambda \leq s \leq M \}, \{ (\lambda, s) : 0 < \lambda = s < M \}, \{ (\lambda, s) : 0 < \lambda \leq s = M \}.

Case 1: Suppose that \( 0 < \lambda^* < s^* < M \). From (79) we have that \( \xi_1 = \xi_3 = 0 \) and from (77) we get that \( \frac{\partial R}{\partial s} = \frac{\partial E[Q^a]}{\partial s} = 0 \). In Lemma 1(d) we prove that for fixed \( \lambda > 0 \), \( E[Q^a] \) 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 \(E[Q^a] = 0\) for all \(s\), and from (77) and (79) we get that \(\xi_1 = \xi_2 = 0\). From (76) we get \(\xi_3 = -\frac{\partial R}{\partial s}_{|_{(0,s^*)}}\). Therefore, the conditions (76)-(80) are satisfied only if \(\frac{\partial R}{\partial s}_{|_{(0,s^*)}} \leq 0\).

**Case 3:** Suppose that \(0 < \lambda^* = s^* < M\). From (79) we have that \(\xi_2 = \xi_3 = 0\), and from (77) we get that \(\xi_1 = c_a \frac{\partial E[Q^a]}{\partial s} \geq 0\). Therefore, from (44) and (46), the condition \(\frac{\partial E[Q^a]}{\partial s} \geq 0\) is satisfied if and only if

\[
\frac{1}{2\sqrt{s}}[h(0) - 0] - \sqrt{s}[1 - h'(0)] \frac{1}{\sqrt{2\pi}} \sqrt{\frac{2m}{E'(0)m}} \geq 0 \iff \frac{1}{\sqrt{2\pi}} - \frac{\pi - 2}{\pi} \sqrt{\frac{2m}{E'(0)m}} \geq 0 \iff s \leq \frac{E'(0)m^2\pi}{4m(\pi - 2)^2}.
\]

\[\square\]

**References**


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


