Nurse lines are call centers that provide advice to patients about what the most appropriate course of action is according to their symptoms: For example, visiting the Emergency Department (ED) versus visiting a physician. For nurse lines, as well as many other service centers, diagnostic accuracy is a key performance measure, along with waiting time due to congestion. Furthermore, since longer service typically entails higher accuracy but also more congestion, managers must decide on the optimal service depth and the staffing level to effectively perform a diagnostic process between agents and customers. We study this type of “diagnostic service center” in a strategic queueing setting, where customers have autonomy to decide whether to use the service center or not, based on their expectation of diagnostic accuracy and waiting time. We model this problem as a multiple-server queueing system with the servers performing a sequential testing process. Analysis of our model shows that the optimal staffing level is such that all potential patients are enticed to use the nurse-line. Also, we show that when the optimal staffing level increases due to a change in the environment, it is not necessarily accompanied by a decrease of the wait times. Finally, the intuition obtained from the analysis of our model leads to improved rules of thumb for practical staffing decisions.

**Key words:** service operations, strategic queueing, diagnostic process, call center

1. **Introduction**

In response to sky-rocketing health care costs and increased competition, the current US health care system is moving toward greater engagement of patients in health care choices (Hibbard 2004). Besides the benefit of making patients feel more empowered, helping patients select appropriate services can reduce unnecessary costs incurred when patients present themselves for inappropriate treatments. Providing *help* in making these decisions is crucial: One defining feature in health care...
markets is that consumers have difficulties in identifying their needs, and matching their needs to
the most appropriate treatment option (Arrow 1963). As a result, a large proportion of health care
costs are caused by unnecessary physician or emergency room visits that arise because consumers
lack medical information, and thus tend to overestimate the severity of their illness (Herzog 2003,
Lynch 2000).

How big of a problem is this? According to statistics reported by the Centers for Disease Con-
trol and Prevention (McCaig and Burt 2005), during 2003, of estimated 113.9 million emergency
room visits, 13% were non-urgent. At an estimated average cost of $300 per emergency room visit
(Machlin 2006), a cost of $4.4 billion can be managed more efficiently by directing patients to the
appropriate care center. This explains why health plans, managed care organizations, hospitals
and physicians\(^1\) are investing in a special type of service center that provides high-quality decision
counseling to patients, enabling them to make more informed and thus more cost-effective treat-
ment decisions (Sabin 1998). We call these service centers “diagnostic service centers\(^2\).” To date,
some evidence-based studies conducted in the medical field have shown cost savings from using
such centers (NurseResponse 2007).

Diagnostic health care service centers have been used in many cities, under different names, such
as nurse line, nurse hot line, nurse triage line, telephone triage, and nurse help line. They provide
a medical triage phone service, to help patients choose the appropriate care, at the appropriate
place, at the appropriate time (NurseResponse 2007). Patients may call a health coach (usually a
registered nurse) 24 hours a day, 7 days a week. Health coaches provide support over the telephone
to help patients interpret and act on symptoms, deciding among treatment options. For example,
*Should I go to the emergency room? Should I schedule an appointment with my doctor? Should I
try self-care?*\(^3\) In addition to the triage function, health coaches provide evidence-based, unbiased
“diagnostic information” in a much broader sense: Helping individuals manage their conditions,
making the best treatment decisions in the context of their values and preferences.

\(^1\) Whereas health plans and managed care organization suffer direct costs due to unnecessary visits, hospitals and
physicians suffer from the indirect inconvenience cost (staff stress, bad reputation) from the increased crowding, and
the opportunity cost incurred by failing to provide care to other patients.

\(^2\) We use “diagnostic” to characterize the information (advice) provided by these service centers, which is valuable for
making decisions, and is not bundled with treatment. It is a diagnosis within nurses’ capacity as professionals, not
equivalent to medical or clinical diagnosis performed by doctors, where legal issues may be involved (Kabala 1998).

\(^3\) A similar description of these options can be found at nurse line vendors such as HealthLine Systems, Inc., IntelliCare,
This paper analyzes how health service providers (health plans, managed care organizations, hospitals or physicians) can make operational decisions to maximize the potential of diagnostic service both in terms of cost savings and patient satisfaction. From a practical point of view, one major challenge for providers is financial: “Dedicating a highly skilled, experienced registered nurse to work the phone lines in a triage center is an expensive undertaking, and one that is often hard to justify to cost-cutting managers” (Hellinghausen 2000). The benefit of offering this service (outside of the intangibles such as patient empowerment) greatly depends on the demand volume aggregated from individual patient’s decisions whether to use the service or not: At a high level, cost savings are proportional to the volume of calls that successfully avert inappropriate treatments. Thus the greater the volume of calls effectively served, the greater the potential savings.

In order to boost call volume, providers must pay attention to the two primary factors that influence patients’ decisions to use the service: Waiting time and accuracy of advice. Nurse-line vendors often market their products along these two dimensions. For example, OnCall (2006) advertises that their system makes it easier to satisfy “the patient’s need for timely, dependable medical advice.” To guarantee high diagnostic accuracy, nurse-lines typically hire experienced registered nurses who are trained to provide advice based on protocols/guidelines (Mayo et al. 2002). Protocols use algorithm-like frameworks that help nurses produce yes/no answers to sets of assessment questions, leading to definitive patient dispositions (Mayo 1998). The manager selects protocols to ensure high diagnostic accuracy; but providing this highly accurate advice takes time, increasing the service time of the patient being served as well as the waiting time of those patients behind her. Thus the fundamental tension in a diagnostic service center is wait versus accuracy.

Nurse-lines have received recent attention in the medical literature. Researchers and practitioners in the Europe as well as in Northern America are interested in the economic viability of nurse-lines. In the UK, researchers have conducted field experiments in which they estimate costs of a nurse-line and cost savings that can be generated. Lattimer et al. (1998) and Lattimer et al. (2000) study nurse-lines for after-hours primary care. They conclude that nurse-lines are cost-effective from the UK’s National Health Services’ perspective: Savings are generated through a reduction of short stay admissions to the hospital, but the authors did not consider patient related costs. Poole (2003) has written a book about nurse-lines for pediatric office practices. He provides rules of thumb for the appropriate staffing levels. A more general survey of nurse-line research and Canadian call
center programs is given by Stacey et al. (2003). Most of the research cited, however, lacks an underlying theoretical basis for assessing the economic viability of nurse-lines.

The two fundamental concerns for nurse-lines are largely treated in different academic literatures: The waiting time performance measure in congested service/queueing systems has been widely studied in the operations management literature, while the accuracy of a (hypothesis) test has been widely studied in the statistics literature. But the link between the cost saving and overall service performance — both accuracy and waiting time — is still largely missing. We provide this link by embedding a corner-stone model of hypothesis testing, capturing the accuracy of the diagnostic protocol, within a queueing framework which captures the patient’s wait time. In addition, by endowing customers with the ability to choose to use the service center or not, we model the effect of the tradeoff between customer wait and accuracy, the two measures of greatest interests in a nurse-line context.

To accomplish this, we abstract the protocols into a certainty threshold to specify the service center’s accuracy: A nurse’s belief about the patient’s pathology, obtained through the interview with the patient, should reach this certainty threshold before she terminates the diagnostic process and gives advice. A high certainty threshold implies high accuracy, but also a longer service time (for greater diagnostic depth) and consequently greater system congestion. Due to the trade-off between patients’ desire for accuracy but aversion to waiting, it remains unclear how patients will react to this certainty threshold, and how the certainty threshold should be set to maximize the service provider’s benefit. As in many service centers, the service provider can influence this tradeoff by determining the staffing level: Hiring more nurses can alleviate congestion, but, is more expensive. Thus the provider must jointly decide on the service depth and staffing level, taking patients’ desire for high accuracy and prompt service as well as staffing costs all into account.

The objective of our research is to construct a model for such diagnostic service centers, which will allow us to understand the complex interplay of diagnostic accuracy, system congestion and staffing within a strategic queueing setting (in which customers have autonomy in making attendance decisions). From our analysis, we obtain the following new insights relevant for managerial practice:

(i) For a broad range of parameters, there exists a nurse staffing level such that the savings derived from in a nurse-line covers the nurse-line wage cost — that is, it is economically desirable to invest in a nurse line.
(ii) If it pays to invest, it is always optimal to staff the nurse-line such that all potential patients are enticed to use the nurseline.

(iii) Optimal staffing levels and equilibrium wait times are not necessarily substitutes: hiring more nurses due to changes in the operating environment may not reduce congestion in the system.

(iv) When the cost of wrongly selecting one option increases, while the cost of wrongly selecting the other option remains constant, the optimal error levels on both options do not necessarily behave as substitutes, they may also behave as complements.

The remainder of the paper is organized as follows. We review relevant literature in Section 2. A description of the model is presented in Section 3. We present the analytical results in Section 4, 5 and 6 respectively. In Section 7, we complement our analysis by means of numerical studies, and in Section 8, we conclude the paper highlighting its theoretical and managerial contributions.

2. Literature Review

This paper is related to several streams of literature. These include information marketing, service operations management, sequential probability ratio testing and strategic queueing.

Researchers in information marketing explicitly model diagnostic information and how to price it. Arora and Fosfuri (2005) analyze the optimal pricing scheme for selling diagnostic information to buyers, modeling the value of information as the difference in expected payoffs with and without it. By contrast, we ignore the pricing issue and focus on setting the certainty threshold, where accuracy comes at the cost of congestion, which depresses demand. Another paper explicitly modeling diagnostic information is Sarvary (2002). This paper considers the market for second opinions and analyzes competition between information sellers, who can provide information with different levels of quality. Queueing delays at the service are not factored into the buyer’s costs, which is a fundamental difference from our paper.

In the service operations management literature, there exists a large body of work addressing service quality, either wait-related measures or customer-satisfaction related measures (for an excellent overview, see Gans et al. (2003)). Our paper complements this literature by modeling the accuracy of the diagnostic information as another aspect of service quality, and by studying its impact on endogenous patients’ decisions. Furthermore, different from traditional service models where customer service time is associated with an objective completion criterion, in our model the
service provider controls the service time distribution by defining the certainty threshold.

There have been papers explicitly considering situations in which the provider can control the service time distribution, and thus impact quality of a “discretionary service” (for an overview, see Hopp et al. (2006)). In Hopp et al. (2007), an agent decides when to terminate processing a task; the reward of the task is an increasing function of time. They characterize the optimal control policy and derive insights; for example that adding capacity may actually increase congestion, and processing time variability can improve system performance. Bouans (2003) considers admission and early abortion of jobs in a multi-phase service center with concave increasing rewards as a function of phases completed. They consider dynamic control decisions, showing that under some regularity conditions, both the optimal admission control policy and the optimal termination policy have a threshold structure. Finally, Debo et al. (2008) link discretionary task completion of an expert service provider to the payment structure. They find that when the expert can stretch the service time without being detected by a customer, a variable service rate may be selected; the expert may slow down when the workload decreases.

Our paper is distinct from the previous papers on discretionary services: In Debo et al. (2008), no value is added (that is, there is no customer benefit) with longer service. Both Hopp et al. (2007) and Bouans (2003) focus on a centralized system in which a real-time task termination policy is determined by the service provider, with no decisions made by customers. In our paper customers do make decisions — whether to use the service or not, based on expected wait times and accuracy.

Recently, Anand et al. (2008) introduced in a different context a relationship between service quality and speed. They assume that the service times are exponentially distributed, but, the value for the customer is a linearly decreasing function of the service rate: Shorter services provides less customer value. Anand et al. (2008) find the equilibrium joining and pricing strategies. In our paper, the value for the customer is also decreasing in the service rate (as diagnosis is less accurate). However, instead of working with exponentially distributed service times, we endogenize the service times (both the first and second moment) as the outcome of a continuous time hypothesis testing problem, which is more accurate for services involving diagnosis. Furthermore, Anand et al. (2008) are concerned about pricing, while we are not.

Our paper is also related to service operations models with a triage server. Shumsky and Pinker (2003), motivated by a health care problem, model a gatekeeper who can diagnose the customer’s
problem and then may or may not refer the customer to a specialist. Their focus is to design a wage contract between the system manager and gatekeepers in a principal-agent setting, such that gatekeepers will choose system-optimal referral rates; there is no explicit queueing. Other significant differences between their model and ours are: (1) Our nurse line has only the triage function, no treatment function; (2) we model the tradeoff of diagnosis accuracy versus congestion; (3) in our model customers make decisions to use this service or not. In a follow-up paper Hasija et al. (2005), the gatekeeper model is extended to include queueing at both the gatekeeper and at the specialist. This paper determines the staffing levels and referral rates that minimize the sum of staffing, customer waiting, and mistreatment costs, exploring when the gatekeeper system is better than a system with only experts. They also give a close approximation of the optimal referral rate. But as in their first paper, the customer demand rate is exogenously given.

Within the call center literature, there are a series of papers studying the cross-selling control problem, which can be seen as a form of the trade-off between increasing revenue by increasing service depth and reducing waiting. In order to answer when and to whom the center should attempt to cross-sell, the cross-selling literature investigates customer segmentation, incentive issues and payment schemes to agents, and how the manager should make these decisions together with staffing and cross-selling control decisions. Moreover, the customer arrival process is exogenous. In contrast, with the health-care motivation in mind, we do not segment patients according to their revenue potentials, and instead we focus on other issues such as quality (accuracy), endogenizing customer arrival decisions. Paper in this stream include Günes and Aksin (2004), Armony and Gurvich (2006) and Gurvich et al. (2006). One of the insights derived from these latter papers is that if staffing levels are appropriately adjusted, the introduction of cross-selling does not necessarily add to customer waiting times. Similarly, we find that as an implication of our sensitivity results, if skill level is appropriately adjusted, the increase of the certainty threshold may or may not increase customer waiting times.

The diagnostic model we develop draws on the sequential probability ratio test literature. In this area, the seminal book is Wald (1947), where the sequential probability ratio test (SPRT) (likelihood-ratio test) was developed as a hypothesis test for sequential analysis. Furthermore, since approximating sums of independent observations by a Brownian motion process has proved to bring considerable simplification and appreciable qualitative insights (see e.g. Siegmund (1985),
Dvoretzky and Wolfowitz (1953) or Sirjaev (1973)), we build our diagnostic model on a sequential analysis of a Brownian motion. This leads to a closed form analysis of the first and second moments of the stopping time (testing time). By considering this stopping time as the service time in a multi-server queueing system, we provide a link between the hypothesis testing and the queueing literatures.

The customer autonomy in making decisions about whether or not to use the service center stems from the strategic queueing literature. Naor (1969) first studies customer autonomy in service systems; he demonstrates that an aggregate equilibrium pattern of behavior exists, which may not be optimal from the point of view of the whole society. Following this paper, many other papers study the equilibrium behavior of customers and servers in queueing systems. Hassin and Haviv (2003) provide a comprehensive review of this literature.

3. The Model

We consider a physician’s group, health plan or managed care organization, which we refer to as the “Health Organization” (HO). The HO has a population of enrolled members; this generates a stream of patients with the need for treatment. In order to help patients make more informed treatment decisions, the HO manages a diagnostic service center — the nurse line. The nurse line assesses the pathology of a patient by asking a series of questions and then dispenses treatment advice. Before seeking treatment each patient chooses to use the nurse line or not, by weighing their perceived value of the diagnostic information versus the inconvenience of waiting to be served.

In the following subsection, we outline our assumptions regarding to patient characteristics, cost structure, diagnostic process and service characteristics. Based on these, we then formulate the HO’s optimization problem taking into account the patients’ decisions.

3.1. Elements of the Model

**Patient characteristics and treatment options.** Each patient is of two possible pathologies, \( \theta \in \{-1, +1\} \). Each patient’s pathology is unknown. The prior probability that a patient has pathology \( \theta = +1 \), is equal to \( \pi \); \( \pi \) is common knowledge and identical for all patients.

There are two treatment options for each patient: \( A \) or \( B \). Every patient must select one of the two treatments, possibly with the help of the nurse-line. Pathology \(-1\) can be interpreted as a “healthy” patient, and \(+1\) as “sick.” Option \( A \) can then be interpreted as “selfcare,” and \( B \) as
Table 1  Costs for the patient and HO as a function of the patient’s pathology and selected treatment.

<table>
<thead>
<tr>
<th>(Patient’s cost, HO’s cost)</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta = -1$</td>
<td>$(0,0)$</td>
<td>$(c_B, C_B)$</td>
</tr>
<tr>
<td>$\theta = +1$</td>
<td>$(c_A, C_A)$</td>
<td>$(0,0)$</td>
</tr>
</tbody>
</table>

“visit the Emergency Department (ED).” These are often the main outcomes for nurse-lines.

**Cost structure.** The appropriate treatment of pathology $-1$ is $A$, and $+1$ is $B$. If a patient selects the appropriate treatment, no costs are incurred\(^4\). The cost incurred by a $+1$ patient wrongly selecting option $A$ is $c_A$; the cost incurred by a $-1$ patient wrongly selecting option $B$ is $c_B$. In addition, patients incur an inconvenience cost $c_w$ per unit of time waiting to be served by the nurse-line. In Section 7, we generalize this to a strictly convex waiting cost. The cost incurred by the HO is determined in a similar fashion: $C_A$ is the HO’s cost when a $+1$ patient wrongly selects option $A$ and $C_B$ is the HO’s cost when a $-1$ patient wrongly selects $B$. Similar to the patient, when a patient chooses the appropriate option, the HO does not incur any costs. Table 1 summarizes the HO and patient costs consequent to the treatment option selected. The HO must hire a number of nurses to operate the nurse line. Each nurse costs $c_n$ per unit of time. We do not consider any other operating costs that are not influenced by the staffing level or service depth selection.

We assume the following relation holds between the patient cost and the probability (prior knowledge) that a patient is sick.

**Assumption 1.** We assume that the patients have a “difficult” decision to make in the sense that with their prior knowledge, both options $A$ and $B$ are equally expensive:

\[(1 - \pi) c_B = \pi c_A.\]

As patients are indifferent between the two treatment options, we assume they select each treatment with probability $p_0 = \frac{1}{2}$ if they do not use the nurse-line\(^5\).

Assumption 1 analytically simplifies our model. It is supported by the fact that the patients satisfying $(1 - \pi) c_B = \pi c_A$ are most likely to patronize the nurse line because they are most ambiguous

\(^4\) We do not include unavoidable costs related to the treatment itself; these costs are irrelevant for the management of the nurse-line.

\(^5\) In principle, when patients are indifferent, they can randomize with any probability. For simplicity, we select $\frac{1}{2}$ as the focal randomization probability; all results are insensitive to this randomization probability.
about the two treatment options. If $(1 - \pi) c_B \gg \pi c_A$ or $(1 - \pi) c_B \ll \pi c_A$, patients are confident in their choice and are less likely to seek advice from a nurse line. For example, for road-side accidents with serious injuries, going to the ED is an obvious choice. That segment of the patient population will call 911, not a nurse-line. Hence, the potential demand rate, $\Lambda$, does not include such obvious situations. We relax Assumption 1 in some of our numerical experiments in Section 7.

**Diagnostic process.** The diagnostic process is modelled as a sequential probability ratio test for the drift of a Brownian Motion: $H_0: \theta r = -r$ against $H_1: \theta r = r$. We assume that for a patient pathology $\theta$, the nurse observes, through questions and answers, a Brownian Motion (BM) with drift $\theta r$ and variance $\sigma^2$ per unit time: $Y_\theta(t) = \theta rt + \sigma X(t)$, where $X(t)$ is the standard Wiener process with $X(0) = 0$. The nurse updates her belief about the patient’s pathology $\theta = +1$ from $\pi$ at time zero to $\pi_t$ at time $t$ by observing $\{Y_\theta(t'), 0 \leq t' \leq t\}$. We refer to $r (\geq 0)$ as the nurse’s skill level; the larger $r$ implies the difference between the two drifts representing the two pathologies is larger, and hence the patient pathology is more easily recognized. Under this formulation, Sirjaev (1973) shows that $\pi_t$ can be written as 

$$
\pi_t = \frac{\pi e^{\frac{1}{2} (y - x)} \exp \left( \frac{1}{2} \frac{Y_\theta(t)}{\sigma} \right) \exp \left( \frac{1}{2} \frac{Y_\theta(t)}{\sigma} \right)}{1 + \pi e^{\frac{1}{2} (y - x)} \exp \left( \frac{1}{2} \frac{Y_\theta(t)}{\sigma} \right) \exp \left( \frac{1}{2} \frac{Y_\theta(t)}{\sigma} \right)}.
$$

We fix $y \leq 0 \leq x$ and have the nurses continue asking questions as long as $y < Y_\theta(t) < x$. If at some time $\tau$, $Y_\theta(\tau)$ hits $x$ (or $y$) for the first time, the patient is advised to seek treatment $B$ (or $A$). We assume the patient follows the nurse’s advice. Thus the service time of the patient, $\tau$, is determined by the following stopping rule:

$$
\tau = \inf \{ t : Y_\theta(t) \notin (y, x) \}.
$$

We refer to the stopping boundary vector $x = (x, y)$ as the **certainty threshold** or the **service depth** set by the service provider.

**Service characteristics.** The enrolled members of the HO fall ill according to a Poisson process with rate $\Lambda$. All such patients independently consider whether to use the nurse line. The call rate of the patient pool is denoted by $\lambda (\leq \Lambda)$. We assume there is no reneging, i.e., once a patient is waiting for the service from the nurse-line, the patient will not leave until served.

The performance of this service center is measured by both the probability of misdiagnosis (error probabilities) and the patient waiting time characteristics. The error probabilities are denoted by $\alpha$ and $\beta$; $\alpha$ is the probability that $Y_{+1}$ first hits $y$ before hitting $x$, and $\beta$ is the probability that

---

6 This excludes patients whose priors do not satisfy Assumption 1.
Y_{-1} first hits x before hitting y. We occasionally refer to \( \alpha \) (\( \beta \)) as the type I (II) error. In the nurse-line context, \( \alpha \) is the probability of advising self care to a sick patient and \( \beta \) is the probability of advising a healthy patient to go to the ED. Both errors lead to costs for the patient and the HO, see Table 1. The error probabilities are functions of the certainty threshold \( x \).

We model the service delivery process as an \( M/G/m \) queueing system; \( m \) is the number of nurses employed to serve the patient pool. Each nurse performs the diagnostic process adopting the same certainty threshold, \( x \). The expected waiting time before being served is denoted as \( W \), which depends on the certainty threshold \( x \), the staffing level of the nurse-line \( m \), the nurse skill level \( r \), and the call arrival rate \( \lambda \).

**Patient’s cost savings.** For a given certainty threshold, the patient’s cost savings *excluding waiting costs* are:

\[
\Delta_P(x) = c_0 - (1 - \pi) \beta(x) c_B - \pi \alpha(x) c_A, \tag{1}
\]

where \( \alpha(x) \) and \( \beta(x) \) are the error probabilities and \( c_0 = p_0 (1 - \pi) c_B + (1 - p_0) \pi c_A \). The cost savings are the difference between the expected patient cost without a nurse-line ("pre-call cost") \( c_0 \), and the expected cost with a nurse-line ("post-call cost") \((1 - \pi) \beta c_B + \pi \alpha c_A \). Should the nurse-line provide perfect advice (i.e. no errors; \( \beta = \alpha = 0 \)), the maximum savings would be \( c_0 \).

**HO’s cost savings.** Similarly, the savings *per patient* to the HO are:

\[
\Delta_{HO}(x) = C_0 - (1 - \pi) \beta(x) C_B - \pi \alpha(x) C_A, \tag{2}
\]

where \( C_0 = p_0 (1 - \pi) C_B + (1 - p_0) \pi C_A \). Note that \( p_0 \) (determined by the patient’s cost structure) determines the HO’s cost without a nurse-line.

**Equilibrium conditions.** Patients behave as rational economic agents who maximize their expected utility. This utility depends not only on their decision and that of the HO, but also on the decision of every other patient. If we denote the expected waiting time in queue when the aggregated demand rate is \( \lambda \) given \( x \) and \( m \) as \( W(\lambda; x, m) \), then the expected utility of a patient calling the nurse line is the difference between the benefit and the waiting cost incurred:

\[
U(\lambda; x, m) = \Delta_P(x) - c_w W(\lambda; x, m). \tag{3}
\]

\(^7\) If \( x = y = 0 \), then we hit both barriers at the same time. In this case, we assume the HO may use any rule to dispense recommendations.
Each patient has two pure strategies: To use or not to use the nurse line. The **calling probability** \( p \in [0,1] \) is a pure strategy if \( p = 0 \) or \( 1 \), and is a mixed strategy otherwise. In this paper we analyze the symmetric equilibrium because patients are assumed to be homogeneous; in a symmetric equilibrium, given the certainty threshold \( x \), all patients have the same calling probability \( p_e(x,m) \) (Hassin and Haviv 2003); the subscript \( e \) represents this is an equilibrium decision.

With this notation, for a given certainty threshold \( x \) and staffing level \( m \), \( \lambda_e(x,m) = p_e(x,m) \Lambda \) and the symmetric patient equilibrium decision \( p_e(x,m) \) satisfies the following conditions (Hassin and Haviv 2003):

\[
p_e(x,m) = \begin{cases} 
1, & U(\Lambda;x,m) > 0, \\
\in [0,1], & U(0;x,m) \geq 0 \geq U(\Lambda;x,m), \\
0, & U(0;x,m) < 0.
\end{cases}
\]

(4)

The above conditions define three situations: (i) When \( U(\Lambda;x,m) > 0 \), even if all patients call the nurse line, the expected utility of a patient is positive, therefore always calling \( (p_e(x,m) = 1) \) is a unique dominant strategy for all patients and thus a unique equilibrium strategy. (ii) When \( U(0;x,m) < 0 \), even if no other patient uses the nurse line, the expected utility of a patient who uses the nurse line is negative, therefore not calling \( (p_e(x,m) = 0) \) is a unique equilibrium strategy. (iii) When \( U(0;x,m) \geq 0 \geq U(\Lambda;x,m) \), there exists a unique equilibrium strategy \( p_e(x,m) \in [0,1] \), which solves \( U(p_e(x,m) \Lambda;x) = 0 \); given traffic \( p_e(x,m) \Lambda \), individual patients are indifferent to calling or not.

In the third situation, the intuition behind zero utility is as follows. Patients anticipate the traffic at the service center and call the nurse line as long as the benefit is greater than the expected disutility due to wait (that is, utility is positive). As demand increases, the waiting disutility increases until it equals with the benefit (that is, utility is zero). At this point, the system reaches equilibrium because patients are indifferent about calling or not and no one will deviate from his/her current decision. Similarly, when utility is negative, some patients will drop the service. As a result, congestion is reduced and waiting disutility decreases until it is equalized with the benefit again. So at the equilibrium, the patient utility is always zero and each patient calls independently with the same probability \( p_e(x,m) \) (Hassin and Haviv 2003).

The HO’s profit rate generated by the nurse line is defined as the difference between the rate of benefit (cost savings) from a nurse line and the staffing cost rate:

\[
J(x,m) = \Delta_{HO}(x) \lambda_e(x,m) - c_n m.
\]

(5)
Table 2  Summary of notations.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_A$ ($C_A$)</td>
<td>Cost to patient (HO) of a sick patient not visiting the health care provider.</td>
</tr>
<tr>
<td>$c_B$ ($C_B$)</td>
<td>Cost to patient (HO) of a healthy patient falsely visiting the health care provider.</td>
</tr>
<tr>
<td>$r$</td>
<td>Skill level of a nurse.</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>Variance of the noise in the diagnostic process.</td>
</tr>
<tr>
<td>$c_w$</td>
<td>Cost rate of patient waiting (per unit of time).</td>
</tr>
<tr>
<td>$c_n$</td>
<td>Cost rate of an employed nurse (per unit of time per nurse).</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>Maximum possible call volume for the nurse-line.</td>
</tr>
<tr>
<td>$\pi$</td>
<td>Prior probability that the patient needs to visit the health care provider.</td>
</tr>
</tbody>
</table>

We consider $m$ as a real number, i.e. we allow hiring a part-time nurse; this assumption has been used in nurse line practice (see Poole (2003)). We can now define the optimal HO decision $(x^*, m^*)$, which is the solution of:

$$J^* = \max_{y \leq 0 \leq x, m \geq 0} J(x, m).$$ (6)

Having established our model, we now proceed, in the next section, to our analysis. Pursuant to this, table 2 summarizes the main parameters that will be used in the remainder of the paper.

4. General Analysis

In this section, we first derive expressions that determine the equilibrium demand rates that arise in response to a given HO’s strategy. To make the decision wether to use the nurse line or not, a patient trades off the benefit of the service (diagnostic accuracy) against the waiting cost. To evaluate this, we first define for a given patient pathology $\theta$ and service depth $x$,

$$u_\theta(x) = \frac{1 - \exp\left(-\frac{2x\theta}{\sigma^2}y\right)}{\exp\left(-\frac{2x\theta}{\sigma^2}x\right) - \exp\left(-\frac{2x\theta}{\sigma^2}y\right)};$$ (7)

$u_\theta(x)$ is the probability of hitting the upper threshold $x$ before hitting the lower threshold $y$ (Stockey 2008). Then, the errors $\alpha(x)$ and $\beta(x)$ are

$$\alpha(x) = 1 - u_{+1}(x) \quad \text{and} \quad \beta(x) = u_{-1}(x).$$ (8)

We can now calculate the first two moments of the stopping time (conditioned on the drift, $r\theta$). Proof of this and all other results appear in the Appendix.

**Lemma 1 (1st and 2nd moments of the stopping time).** When the initial value of the Brownian Motion is zero, conditional on the drift being $\theta r$, the first moment of the stopping time is

$$E_\theta(x) = \frac{1}{\theta r} (xu_\theta(x) + y (1 - u_\theta(x)))$$ (9)
and the second moment of the stopping time is

\[
V_\theta (x) = \frac{4}{\sigma^2} \left\{ u_\theta (x) \int_0^x G(z; x) \exp \left( -\frac{2\theta}{\sigma^2} z \right) dz - (1 - u_\theta (x)) \int_y^0 G(z; x) \exp \left( -\frac{2\theta r}{\sigma^2} z \right) dz \right\} \tag{10}
\]

where

\[
G(z; x) = \int_y^z E_\theta (x - w, y - w) \exp \left( \frac{2\theta r}{\sigma^2} w \right) dw.
\]

With the expressions derived in Lemma 1, we can determine the first and second moments of the stopping time; \(E(x)\) and \(V(x)\):

\[
E(x) = \pi E_{+1}(x) + (1 - \pi) E_{-1}(x), \tag{11}
\]

and

\[
V(x) = \pi V_{+1}(x) + (1 - \pi) V_{-1}(x). \tag{12}
\]

In Lemma 2 we give basic sensitivity properties of the error terms and the first two moments of the service time.

\begin{enumerate}[label=(\roman*)]
  \item \(\frac{\partial}{\partial x} \alpha(x) \geq 0; \frac{\partial}{\partial x} \beta(x) \leq 0; \frac{\partial}{\partial x} (\alpha(x) + \beta(x)) \leq 0; \frac{\partial}{\partial y} \alpha(x) \geq 0; \frac{\partial}{\partial y} \beta(x) \leq 0; \text{ and } \frac{\partial}{\partial y} (\alpha(x) + \beta(x)) \geq 0.
  
  \item \(\frac{\partial}{\partial x} E(x) \geq 0; \frac{\partial}{\partial x} V(x) \geq 0; \frac{\partial}{\partial y} E(x) \leq 0; \text{ and } \frac{\partial}{\partial y} V(x) \geq 0.
\end{enumerate}

Lemma 2 establishes that the main trade-offs in our model of protocol selection are consistent with reality: When the positive threshold \(x\) increases, the type II error \(\beta\) is reduced while the type I error \(\alpha\) increases because the Brownian motion is more likely to hit the negative threshold. But, the overall sum of error levels decreases when the protocol becomes deeper (i.e. \(x\) is larger or \(y\) is smaller): a deeper protocol — a longer test — logically implies the diagnosis is more likely to be correct. This error reduction comes at a cost though: Increased congestion. Both the first and second moments of service time increase when the protocol becomes deeper.

Thus, a deeper protocol creates more congestion, which reduces demand for the nurse-line, or, alternatively necessitates hiring extra nurses. To make these relationships more precise, we use Whitt’s simple heavy traffic approximation (Whitt 1993) to calculate the expected waiting time in queue for a nurse line with \(m\) nurses:

\[
W(\lambda; x, m) = \frac{1}{m^2} \frac{\lambda V(x)}{1 - \lambda \frac{V(x)}{m}}. \tag{13}
\]

\(^8\)Note that here we use Whitt formula for a positive, real \(m\).
Note that the special case of $m = 1$ reduces to the P-K formula (Ross 2002).

Given the expressions for patient benefit and delay, we can now derive the patient equilibrium decision in Lemma 3:

**Lemma 3.** Let $\lambda(x, m)$ denote the unique solution $\lambda$ solving $U(\lambda; x, m) = 0$:

$$\lambda(x, m) = \frac{m}{\frac{c_n}{2m} \frac{r(x)}{\Delta'(x)} + E(x)}.$$  \hspace{1cm} (14)

Then the patient equilibrium decision is:

$$p_e(x, m) = \begin{cases} 1, & \Lambda < \lambda(x, m), \\ \frac{\lambda(x, m)}{\Lambda}, & \Delta'(x) > 0 \text{ and } \lambda(x, m) \leq \Lambda, \\ 0, & \Delta'(x) \leq 0, \end{cases}$$  \hspace{1cm} (15)

and the equilibrium arrival rate is $\lambda_e(x, m) = p_e(x, m) \Lambda$.

Next we analyze the optimization problem (6) and derive the following result about the optimal staffing level relative to the optimal demand rate:

**Proposition 1 (Optimal Staffing).** If $J(x^*, m^*) > 0$, then $m^* = \min \{m | \lambda(x^*, m^*) = \Lambda\}$; if it is optimal to invest in a nurse line, it is optimal to capture all of the demand, and staffing should be set at the minimum level to do so.

The HO’s optimization problem (6) contains three decision variables: $(x, y, m)$. Proposition 1 shows that only two of the three decision variables can be chosen independently: $x$ and $y$. Once these are chosen, $m$ should be set to the minimum value that capture all of the demand.

The intuition behind Proposition 1 is as follows: Assume that at a staffing level $m$, the maximum call volume is not reached, $\lambda < \Lambda$. From Equation (14) we can see that for any fixed certainty threshold $x$, the call volume per nurse $\lambda/m$ increases as the number of nurses $m$ increases — additional nurses attract calls at a rate greater than one$^9$. As a result, the marginal profit per nurse $J/m = \Delta_{HO}\lambda/m - c_n$ also increases as the number of nurses increases. Therefore, it is optimal for the HO to invest in as many nurses as possible$^{10}$, for any protocol $x$. However, the profit for the HO cannot grow unboundedly. Since the total call volume $\lambda$ increases as the number of nurses increases (see also Equation (14)) and is bounded by $\Lambda$, the HO’s profit grows only until the nurse

---

$^9$Because of the quadratic relationship between the staffing and the mean waiting time in Equation (13).

$^{10}$Assuming that the total staffing cost increases linearly in the number of nurses.
line captures the maximum call volume \( \Lambda \). Hence, at the optimal solution, the nurse line always captures the maximum call volume \( \Lambda \).

Proposition 1 is useful as it reduces the degrees of freedom of the optimization problem of Equation (6). It is also interesting because it demonstrates that in queueing models with strategic agents (for example, as in Naor (1969) or in Hassin and Haviv (2003)), when the single-server assumption is generalized to a system in which more servers may be hired at a linear cost per server, the optimal investment in servers leads to a “degenerate” equilibrium, i.e. all agents use the service.

Wang et al. (2007) study the case in which there is exactly one nurse \((m = 1)\) and show that it is possible that at the optimal (symmetric) certainty threshold, the equilibrium calling probability is strictly less than one: \( p^*_e < 1 \); It is not always optimal to entice the whole potential population to call the nurse-line when the staffing level is exogenously given. Hence, adding the flexibility of selecting the staffing level is the driver of the above result.

We have analyzed the optimal demand rate and the staffing, assuming it is optimal to invest. In the following proposition, we derive conditions specifying when it is optimal to make investment:

**PROPOSITION 2 (When to Invest).** (i) It is optimal to invest in a nurse-line (i.e. \( J(x^*, m^*) > 0 \)) if and only if there exists a pair \( x \) that satisfies:

\[
\Delta^{HO}(x) > c_n E(x),
\]

(ii) For any \( c_A, c_B, \pi \) that satisfy Assumption 1, there exists a pair \( x \) that satisfies (16).

The condition (16) in Proposition 2(i) can be understood as follows: If the protocol \( x \) is fixed, \((\alpha, \beta), \Delta_P, \Delta^{HO}, E \) and \( V \) are fixed. Recall that the profit per nurse is \( \Delta^{HO} \lambda/m - c_n \) (see (5)). From Equation (14), we see that \( \lambda/m \) increases as the number of nurses increases, but it is bounded by the fixed service rate \( 1/E \). Thus when the upper bound of the profit per nurse is positive, i.e., \( \Delta^{HO}/E - c_n > 0 \) (see (5) and replacing \( \lambda/m \) with \( 1/E \)), then due to continuity, the number of nurses can be made high enough such that the marginal benefit \( \Delta^{HO} \lambda/m \) approaches \( \Delta^{HO}/E \), and is thus high enough to cover the marginal cost \( c_n \). Therefore there exists a staffing level \( m \) for this certainty threshold, \( x \), such that \( J(x, m) > 0 \).

Considering Proposition 2(ii), we see that under our model, for any unit nurse cost (even if it is very high) and for any cost saving potential (even if it is very low), some investment in a
nurse-line is optimal. This is somewhat surprising. The result can be understood as follows: When \( C_A \pi \neq C_B (1 - \pi) \), then the incentives of the patient and the HO are not aligned. The patient is indifferent between the two options (A and B) under Assumption 1, while the HO has a clear preference. In that case, the following policy generates positive benefits for the HO: Always advise the patient to select the HO’s preferred option. This can be done in “no time” and results in cost savings for the HO. Obviously, other nontrivial policies can likely be found that provide further improvements making investing in a nurse-line desirable.

When \( C_A \pi = C_B (1 - \pi) \), the HO’s and patient’s incentives are perfectly aligned: Both the HO and the patient are ambiguous about the best treatment option. In this case, the HO must take time and perform a diagnostic process to reduce the error rates. Note that the HO’s benefit function becomes \( C_A \pi (1 - \alpha - \beta) = C_A \pi \epsilon \) (linear in \( \epsilon = 1 - \alpha - \beta \)). Thus as \( \epsilon \) decreases toward zero (i.e. a shallow protocol with \( x \) and \( y \) close to 0), the HO’s benefit is linearly decreasing while the average service time \( E(x) \) is decreasing faster (quadratically, see the proof of Proposition 2(ii)). As a result, \( \Delta_{HO}(x) > c_n E(x) \) as \( \epsilon \) goes to zero.

Therefore, with either aligned or non-aligned incentives, there always exists a protocol for which \( \Delta_{HO}(x) > c_n E(x) \) and it is optimal to invest in a nurse-line; at least a fraction of the time of a nurse should be allocated to tele-triage (nurse-line) activities. Obviously, this result needs to be interpreted carefully within the boundaries of the model. Other important factors may affect this result under certain circumstances. We discuss some of these below.

First, recall from Section 3 that we only have included human resources costs for the HO as these are reported to be the most substantial (see e.g. Hellinghausen 2000). If there exists a fixed cost rate that needs to be covered when investing in a nurse-line, for example relating to call center technology (hardware and software) that needs to be acquired and maintained, the cost savings will need to be sufficiently large — not just positive — to cover the fixed cost rate. In this case there will exist threshold levels of \( C_0 \) and \( c_0 \) below which, and a threshold level of \( c_n \) above which, the nurse-line savings do not cover a strictly positive fixed cost rate and hence, investment is not optimal. In this case, part (i) of Proposition 2 will still hold (with the additional fixed cost), but part (ii) may not.

Second, notice that Proposition 2 may prescribe very shallow protocols, i.e. short protocols with a high error probability. Practically, an external regulator would impose upper bounds on the error
probabilities, then, evidently, Proposition 2 might not hold. Imposing such constraint on the HO
would add an additional constraint to part (i) of Proposition 2, and again might cause part (ii) to
fail to hold.

Having established the elemental behavior of the nurse line, we now explore how the outcomes of
the optimization problem of Equation (6) vary as a function of the parameters of the environment.
These results have implications for managers as well as for potential regulators. As the general
sensitivity analysis is quite challenging, we first analyze the case of a completely symmetric cost
structure (Section 5). Then we analyze the infinitesimal asymmetry (Section 6): Starting from the
symmetric case, we change the HO’s cost of one error infinitesimally while keeping the cost of the
other error constant. We complement these results in Section 7, in which we numerically explore
settings with general asymmetric costs, as well as the case of convex waiting cost.

5. Analysis of the Symmetric Case

In this section we derive the comparative statics when the cost structure is symmetric; \( c_A = c_B \),
\( C_A = C_B \) and the prior is \( \pi = 1/2 \). In this case the optimization problem of Equation (6) is
completely symmetric in \( x \) and \( -y \) for a given nurse staffing level \( m \). It follows then that \( y^* = -x^* \)
and \( \alpha^*(x) = \beta^*(x) \). Therefore with a slight abuse of notation we drop \( y \) from the arguments, and
let \( \alpha(x) \) be the total error probability (type I + type II) in the symmetric case as a function of
\( x \). Then, the HO and patients’ savings can be written as: \( \Delta_{HO}(x) = C_A \left( \frac{1}{2} - \alpha(x) \right) \) and \( \Delta_P(x) =
c_A \left( \frac{1}{2} - \alpha(x) \right) \) (see (1) and (2)). The health organization’s problem thus becomes:

\[
J^* = \max_{x \geq 0, m \geq 0} J(x, m) = \Delta_{HO}(x) \lambda_e(x, m) - c_n m
\]

As the optimal call rate is \( \Lambda \) (see Proposition 1), we can rewrite and further reduce the problem
of Equation (17) as follows:

**Lemma 4.** *The optimal number of nurses is determined by \( m^* = \Lambda N(x^*) \), where
\[
J^* = \max_{x \geq 0} \tilde{J}(x) = (\Delta_{HO}(x) - c_n N(x)) \Lambda
\]
and
\[
N(x) = E(x) \left( \frac{1}{2} + \frac{1}{2} \sqrt{1 + \frac{2c_w 1 + C^2(x)}{\Lambda \Delta_P(x)}} \right)
\]
where \( C^2(x) = \frac{V(x) - E^2(x)}{E^2(x)} \).

\(^{11}\) Notice that Assumption 1 is satisfied in this case.
### Table 3  Sensitivity Analysis when $c_w = 0$.

<table>
<thead>
<tr>
<th>Parameter, increasing</th>
<th>Interpretation</th>
<th>Impacts on Error Rate</th>
<th>Impacts on Staffing</th>
<th>Impacts on Service Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_A$</td>
<td>HO’s cost</td>
<td>Negative</td>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>$c_n$</td>
<td>Staffing cost</td>
<td>Positive</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>$r$</td>
<td>Nurse skill</td>
<td>Negative</td>
<td>Positive, Neg.</td>
<td>Pos./Neg.</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>Population size</td>
<td>Zero</td>
<td>Positive</td>
<td>Zero</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>Variance of the diagnostic noise</td>
<td>Positive</td>
<td>Pos./Neg.</td>
<td>Pos./Neg.</td>
</tr>
</tbody>
</table>

As defined above, $\Lambda N(x)$ is the minimum number of nurses required to obtain the optimal equilibrium arrival rate $\Lambda$ when the service depth is $x$; it is the solution for $m$ of $U(\Lambda; x, m) = 0$ or equivalently, $\Delta_P(x) = c_w W(\Lambda; x, m)$. With the formulation of Lemma 4, the decision variable $m$ is eliminated from the optimization problem, and only the optimal protocol, $x^*$ remains. This result is possible due to the simple dependency on $m$ of the elegant heavy traffic approximation of Whitt (1993). With this simplified form, we can now establish that $\hat{J}(x)$ has a unique optimizer over $[0, +\infty]$:

**Proposition 3.** $\hat{J}(x)$ has a unique optimizer $x^*$.

Next, we analyze $\max_{x \geq 0} \hat{J}(x)$ in two phases: Since practitioners often set capacities assuming no congestion externalities (Poole 2003), we first study the case when $c_w = 0$. Then we introduce congestion externalities, that is, $c_w > 0$.

**Ignoring congestion externalities: $c_w = 0$.**

**Proposition 4 (Ignoring congestion externalities).** When $c_w = 0$, we have the comparative statics illustrated in Table 3.

As mentioned above, in practice the case $c_w = 0$ provides a heuristic for determining the optimal nurse staffing level: The capacity is set such that the service rate exactly equals to the arrival rate (Poole 2003); the mean service time, $E$, is estimated and then $m = E\Lambda$ is the number of nurses required when the call arrival rate is $\Lambda^{12}$. In this case, from Lemma 4, we can see the objective function becomes $(\Delta_{HO}(x) - c_n E(x))\Lambda$, and the protocol selection problem reduces to a classical continuous time hypothesis testing problem (see e.g. Sirjaev (1973) and Dvoretzky and

---

12 In practice, nurse-line managers may multiply the estimated mean service time by a factor that is larger than one in order to provide some slack to the system. We do not take such a factor into account for purposes of exposition. In Proposition 5, we will derive a theoretical factor taking the congestion effect into account when solving the hypothesis testing problem. The factor we derive is valid for nurse-lines with high accuracy levels.
Wolfowitz (1953) for an analysis. In that setting, a trade-off is made between the cost of additional experimentation and the type I/type II error reduction: \( \max_{x \geq 0} \Delta_{HO}(x) - c_n E(x) \); the cost per unit of time of experimenting becomes the nurse’s wage \( c_n \). Hence, the sensitivity results in Proposition 4 are the same as those for the classical hypothesis testing problem. We discuss these in our setting now.

First, as the HO’s error cost \( C_A \) increases (simultaneously and identically with \( C_B \)), the error decreases while the testing (diagnosis) time and staffing level increase: The HO’s error cost is the main motivation to invest in nurse-line (perform a test), so a higher cost implies higher cost reduction, which makes nurse-lines more valuable.

Second, note that optimally the system captures \( \Lambda \). As \( m = \Lambda E(x) \), most comparative statistics of staffing level and service time (service depth) are identical. Intuitively, as the marginal nurse cost increases, the staffing level and service depth decrease, and the error probabilities increase. But, as the population size increases, the staffing level increases while the service depth and error are unchanged.

Finally, as the nurse skill \( r \) increases, the error probabilities decrease, but the staffing level and the service depth may either increase or decrease. Fixing other variables and parameters, an increase in the nurse skill \( r \) (the magnitude of the Brownian Motion’s drift) both reduces error and speeds up the testing process. The objective value \( \Delta_{HO}(x) - c_n E(x) \) will increase and the system will move to a new optimum. But the specific form of the new optimum, or in other words, the new optimal service depth, remains unclear; the optimal service depth \( x \) can increase or decrease. As \( m = \Lambda E(x) \), if \( x \) decreases and \( r \) increases, the expected service time decreases and thus the staffing level decreases. But if both \( x \) and \( r \) increase, the expected service time and the staffing may increase: An increase of nurse skill may trigger a more in-depth test, which lasts longer and achieves higher profit.

Because the noise of the Brownian motion \( \sigma \) plays a similar role as \( r \), only in an opposite direction (increasing errors and slowing down the testing), the comparative statics results for \( \sigma \) are opposite of for \( r \).

Note that none of above results are driven by congestion externalities, they are pure consequences of the underlying tradeoff between error reduction and testing costs in the hypothesis testing problem. Next, we study the insights taking the congestion externalities into account.
Table 4 Sensitivity Analysis when $c_w > 0$.

<table>
<thead>
<tr>
<th>Parameter, increasing</th>
<th>Interpretation</th>
<th>Impact on Error Rate</th>
<th>Impact on Staffing</th>
<th>Impact on Wait</th>
<th>Impact on Service Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_A$</td>
<td>HO’s cost</td>
<td>Negative</td>
<td>Positive</td>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>$c_n$</td>
<td>Staffing cost</td>
<td>Positive</td>
<td>Negative</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>$r$</td>
<td>Nurse skill</td>
<td>Negative</td>
<td>Pos./Neg.</td>
<td>Negative</td>
<td>Pos./Neg.</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>Variance of the diagnostic noise</td>
<td>Positive</td>
<td>Pos./Neg.</td>
<td>Negative</td>
<td>Pos./Neg.</td>
</tr>
<tr>
<td>$c_A$</td>
<td>Patient’s cost</td>
<td>Negative</td>
<td>Pos./Neg.</td>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>$c_w$</td>
<td>Waiting cost</td>
<td>Positive</td>
<td>Pos./Neg.</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>Population size</td>
<td>Negative</td>
<td>Positive</td>
<td>Positive</td>
<td>Positive</td>
</tr>
</tbody>
</table>

Taking congestion externalities into account: $c_w > 0$. Patients endogenously take their waiting into account which gives them a disincentive to use the service. In this case, the specifics of the waiting behaviors, $C^2$ and $\alpha$, play a role in the optimal staffing decision, (18). Proposition 5 provides insight (bounds) on how $C^2$ depends on the certainty threshold, $x$:

**Proposition 5 (Properties of $C^2$).** (i) $C^2(x)$ is decreasing in $x$ over $(0, +\infty)$. (ii) $\lim_{x \to 0} C^2(x) = \frac{2}{3}$; (iii) $\lim_{x \to +\infty} C^2(x) = 0$.

At a high service depth, the expected stopping time increases linearly in the threshold, while the standard deviation of the Brownian Motion increases linearly in the square root of the threshold. Hence, the coefficient of variation tends to zero when the threshold tends to infinity.

With strictly positive waiting costs ($c_w > 0$), some sensitivity results are different from the case of $c_w = 0$. Let $W^* = W(\Lambda; x^*, m^*)$ denote the waiting time when the staffing level and protocol are selected optimally. Now, we complete the comparative statics analysis:

**Proposition 6.** When $c_w > 0$, we have the comparative statics in Table 4.

Looking at Table 4, we notice that waiting time and error probability are substitutes: When one increases, the other decreases. This is because in equilibrium the patient does not obtain any strictly positive surplus. If there were a positive surplus, the HO would be able to increase the certainty threshold and obtain more cost savings and all patients would still join. As a result, the expected waiting cost for the patient is exactly equal to the expected benefit of calling the nurse-line, that is, the patient’s surplus is driven back to zero again. Hence, when the benefit increases (i.e. the error decreases), the waiting cost (i.e. the expected waiting time) increases.

Other results about the impact on error rates and waiting are intuitive: When the patient’s
mismatch cost increases, the nurse-line becomes more attractive to the patient and thus the HO can increase wait and decrease the error probability without losing the demand. In addition, we see that when the population size increases, the error level decreases. This indicates that the optimal error levels cannot be determined at the level of an individual patient using hypothesis testing theory, but should reflect the total call volume for which the nurse-line is designed. In Subsection 7.4, we explore numerically how the two key performance parameters, accuracy and wait times, change as the scale of the nurse-line grows.

The impact on staffing is complicated: The optimal staffing level monotonically increases as the HO’s cost increases, because the HO’s cost is a major motivation for investment in a nurse-line. In contrast, the patient’s cost has an ambiguous impact on the optimal staffing level. On one hand, a larger patient’s cost will increase the savings for the patient, making it more attractive for the patient to call and thus the number of nurses required to entice the demand can be decreased. On the other hand, if more patients call, this may require a higher staffing level. Thus the net effect can go in either direction. The impact of the waiting cost on the staffing level is similar as can be seen from Equation (18): It is also ambiguous, but opposite to the impact of the patient’s cost. This comparative statics analysis also illustrates how fundamentally different the patient’s cost and the HO’s cost drive the optimal staffing decisions.

We conclude this section by comparing the optimal staffing level $m^* = \Lambda N(x^*) = \Lambda \mathbb{E}(x^*)(\frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{2c_w(1 + C^2(x^*))}{\Lambda \Delta p(x^*)}})$ (18) with the heuristic for determining the optimal nurse staffing level by ignoring congestion (the classical hypothesis testing problem), $\Lambda \mathbb{E}(x^*)$. We can see that the optimal staffing has a “safety factor” $\frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{2c_w(1 + C^2(x^*))}{\Lambda \Delta p(x^*)}}$ that inflates the expected service time in the staffing decision. For a fixed service depth $x^*$, the “safety factor” is high when the patient waiting cost $c_w$ is high, the maximum call volume $\Lambda$ is low or the patient cost $c_A$ is low (since $\Delta p(x^*) = \frac{c_A}{2}(1 - \alpha(x^*))$). As the service depth approaches zero, $\frac{1 + C^2(x^*)}{\Lambda \Delta p(x^*)}$ can be arbitrarily high (because $\Delta p(x^*)$ is close to zero), and hence the deviation from that heuristic can be arbitrarily high.

6. Analysis of the Asymmetric Case

In this section we study how the derived certainty thresholds, error probabilities, waiting time and optimal investment in nurses change with the type I mismatch costs $C_A$, i.e. the cost of a sick

patient erroneously not visiting a health provider. Recall the symmetric case is defined in Section 5: \( C_A = C_B, \, \pi = \frac{1}{2}, \, c_A = c_B \), and consequently, \( x^* = -y^* \). To make the asymmetric analysis tractable, we perform the comparative statics in the small neighborhood of the symmetric case. Specifically, the optimal number of nurses and certainty thresholds are determined by:

\[
J^* = \max_{y \leq 0 \leq x} \left( \Delta_{HO}(x) - c_n N(x) \right) \Lambda,
\]

where \( N(x) \) is determined by Equation (18) in Lemma 4. We change \( C_A \) to \( C_A + \delta \), where \( \delta \) is very small, keeping all other parameters, \( (C_B, \, c_A, \, c_B, \, \pi, \, r, \, \sigma, \, c_w, \, c_n, \, \Lambda) \), constant, then we study the changes of error rates, waiting and staffing. In Section 7, we will see that the analytical comparative statics obtained at \( x^* = -y^* \) also numerically hold when \( x^* \neq -y^* \).

**Proposition 7.** When \( C_A \) changes, we have the comparative statics illustrated in Table 5.

From Table 5, we can see the major difference between \( c_w = 0 \) case and \( c_w > 0 \) case. In the \( c_w = 0 \) case (the problem reduces to a classical hypothesis testing problem), both errors are substitutes: When the HO’s cost of one option increases, the corresponding error decreases, and the other error always increases to control the service time, and thus the testing cost \( c_n E(x) \). In the \( c_w > 0 \) case, when the HO’s cost of one option increases, the error of this option again decreases, but the error of the other option could either increase or decrease. The congestion externalities change the fundamental characteristics of the hypothesis testing problem. Specifically, the testing cost structure is changed from \( c_n E(x) \) to \( c_n N(x) \), in which \( C^2 \) plays an important role in addition to the expected service time. So as \( C_A \) increases, the two errors may decrease together.

### Table 5  Impact of Asymmetry (increasing \( C_A \) to \( C_A + \delta \)).

<table>
<thead>
<tr>
<th>( c_w &gt; 0 )</th>
<th>Impact on Type I Error Rate</th>
<th>Impact on Type II Error Rate</th>
<th>Impact on Wait</th>
<th>Impact on Staffing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>Negative</td>
<td>Pos./Neg.</td>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>( c_w = 0 )</td>
<td>Negative</td>
<td>Positive</td>
<td>Positive</td>
<td>Positive</td>
</tr>
</tbody>
</table>

7. Numerical Experiments and Extensions

After discussing parameter values and a baseline setting in Subsection 7.1, we conduct numerical experiments that allow us to obtain insights into the following questions: (1) How strong is the substitution between the two error probabilities when the asymmetric HO cost structure is introduced? (Subsection 7.2) (2) What are the sensitivity results when the patient’s cost changes and
Table 6 Parameter settings used for the numerical experiments.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>unit</th>
<th>Range</th>
<th>Base Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient Population; $\Lambda$</td>
<td>calls per minute</td>
<td>$20/60 \sim 500/60$</td>
<td>20/60</td>
</tr>
<tr>
<td>Skill Level; $r$</td>
<td>NA</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>Noise on diagnosis; $\sigma$</td>
<td>NA</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Insurer’s Cost; $C_A = C_B$</td>
<td>$$ per visit</td>
<td>$20 \sim 250$</td>
<td>50</td>
</tr>
<tr>
<td>Patient’s Cost; $c_A = c_B$</td>
<td>$$ per visit</td>
<td>$1 \sim 40$</td>
<td>25</td>
</tr>
<tr>
<td>Staffing Cost; $c_n$</td>
<td>$$ per minute per person</td>
<td>$0.2 \sim 0.5$</td>
<td>0.5</td>
</tr>
<tr>
<td>Waiting Cost; $c_w$</td>
<td>$$ per minute</td>
<td>$0 \sim 10$</td>
<td>4</td>
</tr>
</tbody>
</table>

Assumption 1 may not be satisfied? (Subsection 7.3) (3) How do non-linear waiting costs change the optimal error probabilities and waiting times? (Subsection 7.4)

7.1. Parameters and Base Case

We construct our numerical study based on ball-park data inferred from an American Academy of Pediatrics guide for developing a telephone advice system for a pediatric office practice (Poole 2003). In this guide (page 121), the peak call volume to a 5-pediatrician practice is 145 calls per hour. Since large scale operations may aggregate the demand of more physicians, we set $\Lambda$ in a broader range: from $20/60$ to $500/60$ calls per minute in our numerical experiments. We use the range of hourly compensation for registered nurses across the United States reported in the same guide (page 161): $13$ to $30$ per hour (or $0.2$ to $0.5$ per minute). In the following experiments, we test the impact of changing parameter values around the base case, in the range we show in Table 6.

We validate the data of the base case by estimating its performance measures and staffing level. We rewrite the objective function as a function of the error rate $\alpha$ and solve the first order condition. Ignoring the congestion externalities, the optimal error level is $2.30\%$, the average service time is 7.15 minutes, and 2.38 nurses are needed. When the waiting costs are $4/minute^{13}$, the error level increases to 3.28$, the number of nurses increases to 3.17, the service time is 6.32 minutes and the waiting time is 2.92 minutes. These numbers (call duration, call rate, staffing level) are of the same order of magnitude as reported by Poole (2003).

7.2. Changing the HO’s costs

In this subsection we verify and extend the results in Section 6. by changing the HO’s cost of one error while fixing the HO’s cost of the other error.

---

$^{13}$ We have run experiments with other values of waiting costs, and the results were qualitatively the same as these reported in this section.
Figure 1  In the left panel, optimal error levels, \((\alpha^*, \beta^*)\) when fixing \(C_B = 50\) at \(c_w = 0\) (bottom curve), \(c_w = 4\) (top curve) while \(C_A\) ranges from 20 to 250. At the solid line, \(\alpha^* = \beta^*\). In the right panel, the optimal staffing, \(m^*\), when \(c_w = 0\) (bottom curve), \(c_w = 4\) (top curve) for \(C_A\) ranging from 20 to 250. At the vertically dashed line, \(C_A = C_B = 50\).

Example 1: \(\sigma = 1, r = 1/2, \pi = 1/2, c_A = c_B = 25, \Lambda = 20/60, c_w = 4, c_n = 0.5\). Fixing \(C_B = 50\), we let \(C_A\) range from 20 to 250 for \(c_w = 0\) and 4 respectively, see Figure 1.

Observations: Figure 1 (left panel) confirms that without congestion externalities \((c_w = 0)\), when the mismatch cost of one option \((C_A)\) increases while the other \((C_B)\) remains constant, there is a substitution effect on the errors as shown in the bottom curve of Figure 1 (left panel): One error increases while the other error decreases. With congestion externalities \((c_w > 0)\), this substitution effect is greatly weakened (top curve of Figure 1) and may even become a (weak) complementary effect. These results have been proved to hold in a small neighborhood of the symmetric cost structure (Proposition 7). Our numerical studies indicate that when the cost parameters are highly asymmetric, these results remain true. It may be safe to say that the error probabilities become more loosely coupled due to congestion: Changes in the HO’s cost of one option mainly impacts the error of that option and almost not of the other option. Figure 1 (right panel) indicates that in the asymmetric case, the optimal staffing level increases as the HO’s cost of one option increases, which is consistent with the symmetric case (Table 4).

Taking both figures together, comparing solutions at a fixed HO’s cost \((C_A)\), if \(c_w\) is increased from \(c_w = 0\) to \(c_w > 0\), two things may happen: The staffing may increase and/or the type II error may increase. This is because the congestion cost drives customers away from the system. To
Figure 2  When \( c_A \) ranges from 1 to 40, in the left panel, optimal error levels, \((\alpha^*, \beta^*)\) for both options. In the right panel, the optimal staffing level, \( m^* \). At the vertically dashed line, \( c_A = c_B = 25 \).

When recapture them, the system reduces the wait in one or both of the above ways.

7.3. Changing the patients’ costs

Another way of introducing an asymmetric cost structure is to change the patients’ costs. In this case, Assumption 1 may not be satisfied and thus \( p_0 \), the probability of choosing the option \( A \) without consulting a nurse line may not be \( \frac{1}{2} \); Since \( \pi \) is defined as the probability that the patient’s pathology matches the treatment \( B \), the expected error cost from using treatment \( A \) is \( \pi c_A \) and the expected error cost from using treatment \( B \) is \( (1 - \pi) c_B \). Hence, if the expected cost of selecting option \( A \) is lower than the expected cost of selecting option \( B \), \( p_0 = 1 \); if the expected cost of selecting option \( A \) is higher than the expected cost of selecting option \( B \), \( p_0 = 0 \); otherwise, we set \( p_0 = \frac{1}{2} \). Now we illustrate the role that the two errors play when changing one of the patient’s cost parameters while keeping the other constant.

**Example 2:** \( \sigma = 1, r = 1/2, \pi = 1/2, C_A = C_B = 50, \Lambda = 20/60, c_w = 4, c_n = 0.5 \). Fixing \( c_B = 25 \), we let \( c_A \) range from 1 to 40, see Figure 2.

**Observations:** Similarly as changing the HO’s costs, the error probabilities are either substitutes or complements with respect to changes of the patients’ costs (due to a similar congestion effect as discussed in Proposition 7). But different from the case of changing the HO’s costs, the staffing level is non-monotone: From Figure 2 (right panel), we also see the staffing level decreases rapidly when \( \pi c_A < (1 - \pi) c_B \) (or equivalently in this example \( c_A < c_B \)), and increases slowly when \( \pi c_A > \)
(1 − π)c_B (or equivalently in this example c_A > c_B). When πc_A < (1 − π)c_B, without a nurse line, the patient always selects option A. Thus c_A plays an important role in both the pre-call and the post-call expected costs, πc_A and (1 − π)β(x)c_B + πα(x)c_A. The difference between the pre-call and the post-call expected costs is \( \Delta_P(x) = -(1 − \pi)\beta(x)c_B + \pi(1 − \alpha(x))c_A \), which tends to increase in c_A (disregarding second order effects), which gives patients more incentive to use the nurse line. Thus the nurse line manager can decrease the number of nurses to lower the staffing cost. In contrast, when (1 − π)c_B < πc_A, without a nurse line, the patient always selects the option B, and has a pre-call expected cost of (1 − π)c_B. Thus c_A only plays a role in the post-call expected costs, as the nurse-line could mistakenly advise option A to a type I patient. So the increase of c_A reduces the patients’ incentive to use the service. In order to keep the demand, the nurse line manager has to (slowly) increase the staffing level.

Hence, starting from a symmetric cost structure, changing the patient’s cost is much more subtle than changing the HO’s cost.

7.4. Convex waiting costs

Recall from Section 3 that we assumed that the waiting cost is a linear function of the waiting time. Since some studies (Larson 1987, Taylor and Claxton 1994) argue that the waiting cost may be a non-linear function of the waiting time, in this subsection we set waiting cost equal to \( c_wW_0^\gamma \left( \frac{W}{W_0} \right)^\gamma \), where \( \gamma \in (0, 1] \) and \( W_0 > 0 \). When \( \gamma = 1 \), the waiting cost function is linear; otherwise, the waiting cost function is strictly convex. In addition, when \( \gamma < 1 \), waiting times shorter than \( W_0 \) are less expensive, but waiting times longer than \( W_0 \) are more expensive than when the waiting costs is linear, i.e. \( W_0 \) is a “soft” waiting time constraint. As \( \gamma \) goes to zero, the convex waiting costs become a degenerate function: Zero when \( W < W_0 \) and infinity when \( W \geq W_0 \). Thus, \( W_0 \) becomes a hard constraint.

Solving \( U(\lambda; x, m) = 0 \) with the convex waiting cost, we obtain the corresponding equilibrium demand rate:

\[
\lambda(x, m) = \frac{m}{2c_w V(x) \Delta_P(x)^\gamma + E(x)}. \tag{19}
\]

The corresponding N (see Lemma 4) is then

\[
N = E \left( \frac{1}{2} + \frac{1}{2} \sqrt{1 + \frac{2c_w W_0^{\gamma - 1} 1 + C^2}{\Lambda \Delta_P}} \right).
\]
Given this formulation, the underlying logic of Proposition 1 and Proposition 2 still holds; the call volume per nurse $\lambda/m$ increases as the number of nurses $m$ increases, and it is bounded by the fixed service rate $1/E$ (see Equation (19)). Next, we explore the effect of different curvatures of the waiting cost function as the call volume increases.

**Example 3:** $\sigma = 1$, $r = 1/2$, $\pi = 1/2$, $C_A = C_B = 50$, $c_A = c_B = 25$, $c_w = 4$, $c_n = 0.5$, $W_0 = 2$. $\Lambda$ ranges from 20 calls per hour to 500 calls per hour, see Figure 3.

**Observations:** From Figure 3 (left panel), it is confirmed that the error rate and the waiting time are substitutes as the scale of the nurse line ($\Lambda$) increases (Proposition 6): the symmetric error rate decreases while the waiting time increases. This again reflects the primary tradeoff: high accuracy vs. low wait. In addition, the more convex the waiting cost function is (that is, lower $\gamma$), the less sensitive the optimal waiting time is to the change in scale. When the waiting cost function is more convex, patients care more about the waiting time in their utility functions, and the nurse line manager cannot increase waiting time as much as she increases accuracy as traffic grows. As a result, the waiting time does not increase much as $\Lambda$ increases. We also see that the error rate of small-scale nurse-lines is more sensitive to fluctuations in the potential demand ($\Lambda$) than the error rate of large-scale nurse-lines. In experiments when $W_0 = 2$ and $\gamma = 1/10$, when $\Lambda$ increases from 20 to 270 calls per hour, the error rates change from 0.025 to 0.038, while for $\Lambda$ between 270 and 500 calls per hour, the error rates range from 0.023 to 0.025. Thus, Hence, if a regulator aims at ensuring an exogenously specified accuracy level, large nurse-lines will be more likely to reach that level consistently, while smaller nurse-lines may not. This creates incentives for individual practitioners (each with a low potential call rate) to form a large nurse-line (Poole 2003).

Figure 3 (right panel) illustrates the impact on the optimal nurse staffing level when the potential call rate ($\Lambda$) and the convexity of the waiting cost change. We plot the optimal number of nurses per unit of potential call rate ($m^*/\Lambda$). First, notice that small-scale nurse-lines need more nurses per call per minute than large scale nurse-lines. This is due to “economies of scale” gained at large-scale nurse-lines. Second, notice that the more convex the waiting cost function is (that is, at lower $\gamma$), the higher $m^*/\Lambda$ is. This is because patients are more sensitive to the waiting time, and as a result, the nurse line must increase the staffing level to reduce congestion.
8. Discussion and Conclusion

We analyzed a model in which a service provider can select the duration and quality of a provided service. Our main motivation is given by nurse-lines: Call centers that provide patients advice on the most appropriate course of action according to their symptoms. Nurse-lines play a role in reducing the pressure on many Emergency Departments (EDs) throughout the US by filtering the patients that go to the ED. Our model combines fundamental insights from both the hypothesis testing and queueing literature to clarify the complex interaction between the key variables that are of managerial interest. Our primary contributions are as follows:

- We provide insight in the management of discretionary service processes (Hopp et al. 2006). We model the service process in detail as a hypothesis testing problem of the drift of a Brownian Motion. The stopping boundary determines both the service value and the stopping time distribution, which in turn determines the queue congestion. Hopp et al. (2007) showed via simulation that increasing capacity may increase the waiting time if the extra capacity is deployed to increase the service value (accuracy). We prove how, in general, changes to parameters of the environment may lead to changes in experienced wait times and optimal capacity investment that go in different directions: More capacity can be accompanied by either longer or shorter wait times.

- While the hypothesis testing literature considers a trade-off between accuracy and testing
costs (time) on a case-by-case base (Wald 1947, Sirjaev 1973), we analyze how the large scale hypothesis testing problem (nurse line management problem) can be implemented with (i) a number of “experimenters” (nurses) and (ii) endogenously determined demand for experiments (traffic). The implication for large scale hypothesis testing is that negative congestion externalities are created that must be considered when determining the staffing levels and optimal error probabilities. Our analysis shows that the staffing level depends on the squared coefficient of variation of service time and the patients’ valuation of the experiments.

- Finally, we introduce multi-server queues within a strategic queueing framework (Naor 1969, Hassin and Haviv 2003) and find that, using Whitt’s simple heavy traffic approximation, the equilibrium calling probability is 1 when the investment costs are linear in the number of nurses. This is due to the existence of stochastic economies of scale: When increasing the number of nurses proportional to the demand rate, the average waiting time decreases. This is different from the classical strategic queueing framework by Naor (1969), which considers only a single server.

For practice, our comparative static results reveal insights for managing nurse-lines. While clinical literature has shown that nurse-lines may generate savings (Lattimer et al. 2000), little is known about how to manage them. We show explicitly how our model improves the rule-of-thumb given by Poole (2003) to determine the staffing levels of a nurse-line: We take into account (i) protocol selection and (ii) congestion externalities. In general, the optimal protocol selection is intuitive: More mismatch cost on one treatment option leads to lower error on that option; when hiring a nurse becomes more expensive, errors increase; more highly skilled nurses lead to reduced errors. However, the staffing decision is much more subtle than protocol selection: When the error decreases, either more or fewer nurses may be required. Our model provides the intuition behind these findings, and hence provides the first steps toward a better understanding of the management and design of nurse-lines (and similar service centers). These insight are also valuable for potential regulators. They may be interested in the error levels achieved by the nurse-line as well as the expected waiting times.

Our model is the first analytical model explicitly linking queueing and hypothesis testing theory. More elaborate decision models could be developed from our model, to incorporate tactical staffing decisions, multiple pathologies and treatments (e.g. call 911, visit ED, visit physician, self-care) or different quality measures (e.g. the busy probability) for example. In this paper, we have focused on
unconstrained optimization of the HO’s profit. Should a regulator impose some accuracy (and/or wait) level, then a constrained optimization problem may have to be considered, in which not only the staffing level is determined, but also the nurse skill or nurse pool skill mix.

9. Acknowledgement

Laurens Debo wishes to acknowledge summer support from the Graduate School of Business of the University of Chicago.

References


10. **Appendix**

A proof of the Propositions can be obtained from the authors.