In the health care domain, diagnostic service centers provide advice to patients over the phone about what the most appropriate course of action is based on their symptoms. Managers of such centers must strike a balance between accuracy of advice, callers’ waiting time and staffing costs by setting the appropriate capacity (staffing) and service depth. We model this problem as a multiple-server queueing system with the servers performing a sequential testing process, and the customers deciding whether to use the service or not, based on their expectation of accuracy and congestion. We find the dual concerns of accuracy and congestion lead to a counterintuitive impact of capacity: Increasing capacity might increase congestion. In addition, (i) Patient population size is an important driver in management decisions, not only in staffing, but also in accuracy of advice; (ii) Increasing asymmetry in error costs may not increase asymmetry in the corresponding error rates; and (iii) The error costs for the two major stake-holders — the service manager and the patient — may impact the optimal staffing level in different ways. Finally, we highlight the relevance of our model and results to challenges in practice elicited during interviews with current clinical researchers and practitioners.

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 thus find it hard to match 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 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 Control and Prevention (McCaig and Burt 2005), during 2003, of an 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 treatment decisions (Sabin 1998). We call these service centers “diagnostic service centers\(^2\).”

Diagnostic 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.

This paper analyzes how health service providers (health plans, managed care organizations, hospitals or physicians) can make operational decisions to maximize the benefits 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

---

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, as well as 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, Inc., Nurse Response, LVM Systems (Courson 2005).
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 can manage 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 (Mayo et al. 2002). There are a variety of protocols in the market. These protocols are symptom based algorithm-like frameworks that help nurses produce yes/no answers to sets of assessment questions, leading to definitive patient dispositions (Mayo et al. 2002).

Nurse lines implement protocols, either within a computer format or book format, to ensure high diagnostic accuracy; but providing 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. The service provider can also influence this tradeoff by determining the staffing level: Hiring more nurses can improve service, but is more expensive. Thus the provider must jointly decide on the service depth (protocol-related choices) and staffing level, taking patients’ desire for high accuracy and prompt service as well as staffing costs all into account. This is the fundamental question we explore in this paper.

Nurse lines have received much attention in the medical literature. Many studies have estimated positive cost-savings that can be attributed to nurse lines through directing callers to the appropriate level of health care and thus relieving unnecessary traffic at the expensive/urgent options (Bunik et al. 2007, Bogdan et al. 2004, Cariello 2003). Another important set of questions addressed in this literature pertain to the safety and appropriateness of the nurse line advice (Kempe et al. 2000). Performance on waiting times and call durations has also been studied (The Quality Improvement of Literacy and Collaborative 2006, Valanis et al. 2003). Poole (2003) provides a comprehensive guide for developing a telephone advice system for a pediatric office practice. It describes the methods and detailed steps which have been developed and tested over 15 years in children’s hospital telephone advice programs nationwide.

4 We use the term “accuracy” to refer to the concepts of “safety” and “appropriateness” mentioned in the nurse-line literature.
To date, most of these research studies, however, are based on field experiments or empirical surveys, adopting the perspective of only one stake-holder (e.g. Bunik et al. (2007) focus on the HMOs, Kempe et al. (2000) focus on the patients). Due to the interaction between patients’ incentive to use the service and managers’ incentive to invest, a better understanding of the benefits/costs for different stake-holders involved (patients, HMOs, physicians or hospitals) is needed to promote more efficient and effective deployment of nurse lines. In addition, these extant studies decompose the interrelated decisions in the management problem: At the strategic level, protocol implementation and nurse training policies address the appropriateness of the recommendation. Then, at the tactical level, nurse staffing addresses the waiting time performance. However, it is generally agreed that different protocol implementations result in different diagnostic performance for a given symptom (Wheeler and Siebelt 1997) as well as call duration (Bunik et al. 2007). Hence, a model that can jointly optimize these management decisions — protocol implementation and staffing — which both impact the appropriateness of the recommendation and the waiting times, is required.

The objective of our research is to construct such an analytical model, integrating perspectives of different stake-holders involved in diagnostic service centers. This model will allow us to understand the complex interplay of diagnostic accuracy, system congestion, staffing, and customer autonomy in a unified framework. To accomplish this, we embed a corner-stone model of hypothesis testing, capturing the accuracy of advice, within a multiple-server strategic queueing framework which balances the patient’s benefit against her/his waiting time, which is in turn influenced by the nurse line’s staffing decision. Existing analytical models have treated the two fundamental concerns for nurse lines separately: The waiting time in service/queueing systems in the operations management literature, and the accuracy of a (hypothesis) test in the statistics literature.

Specifically, we model the service center’s accuracy, or service depth as being controlled by a certainty threshold: 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, we need to analyze how patients will react to this certainty threshold, and consequently how the certainty threshold and staffing level should be jointly set to maximize the service provider’s benefit net of staffing costs. Our analysis answers the following types of research questions: What are the optimal staffing level and the optimal service depth the service provider should set? When should the service provider
invest in a nurse line? How do nurse skill level, patient population size and other parameters impact these decisions?

The main contributions of this paper are:

1. We develop an analytical model to evaluate cost savings by linking call duration with recommendation accuracy. Our analysis allows the exploration of the impact of staffing and protocol implementation on accuracy of advice and waiting times. With this link, we analyze the benefit/cost for the different nurse line stake-holders in a unified framework.

2. Results we obtain confirm that the tradeoff between accuracy and congestion motivated by diagnostic service centers in the health care domain changes many aspects of traditional call center design and staffing. By linking queueing and hypothesis testing theory, analysis of our model extends these theories and derives new insights which are useful in nurse line design and staffing.

3. Through interviews with nurse line researchers and practitioners, we establish the crucial questions and challenges facing these lines in practice. We then relate our model to these challenges, creating a bridge between research and practice.

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 link our paper with current challenges in practice. In Section 9, 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 test, strategic queueing and nurse line clinical studies. We have discussed the relevant nurse line clinical papers in the introduction; we introduce the other four streams here.

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. We ignore the pricing issue and focus on setting the certainty threshold, where accuracy comes at the cost of congestion, which depresses demand. Sarvary (2002) considers the market for second opinions and analyzes competition between information sellers, who can provide information with different levels of quality. Queueing delays are not factored into the buyer’s costs.

In the service operations management literature, there exists a large body of work address-
ing 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 diagnostic information as another aspect of service quality, and by studying its impact on patients’ decisions to patronize the center. 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 which is an increasing function of time. They characterize the optimal control policy and derive insights; for example adding capacity may actually increase congestion, and processing time variability can improve system performance. Bouns (2003) considers admission and early abortion of jobs in a multi-phase service center with concave increasing rewards as a function of phases completed. He considers dynamic control decisions, showing that under some regularity conditions, both the optimal admission control policy and the optimal termination policy have a threshold structure. Both Hopp et al. (2007) and Bouns (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 contrast, our paper customers do make decisions — whether to use the service or not, based on expected waiting times and accuracy. 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, a variable service rate may be selected; the expert may slow down when the workload decreases. In that paper, customers do not obtain additional benefit from longer service, which is distinct from ours.

Anand et al. (2008) introduce 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. They find 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 moments) as the outcome of a continuous time hypothesis testing problem, which is more accurate for services involving diagnosis. Recently, de Véricourt and Sun (2009) study a similar judgement accuracy/congestion tradeoff in a decision process. They explicitly model the judgement process, as we do for the diagnostic process. However, differently, they focus
on a centralized system with no customer decisions and within a dynamic model. In addition, to simplify the analysis, they assume a single-sided hypothesis test (i.e. no false positives), which is another significant difference (our model captures both false positives and negatives).

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 a 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; there is no explicit queueing and demand is exogenous. A follow-up paper (Hasija et al. 2005) extends their model 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. 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 is a form of the trade-off between increasing revenue by increasing service depth and reducing waiting. Papers in this stream include Günes and Aksin (2004), and Gurvich et al. (2009). In order to answer when and to whom the center should attempt to cross-sell, this literature investigates customer segmentation, incentive issues and payment schemes to agents, together with staffing and cross-selling control decisions, when the customer arrival process is exogenous. In contrast, we do not segment patients according to their revenue potentials, and instead focus on other issues such as quality (accuracy), endogenizing customer arrival decisions.

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. Since approximating sums of independent observations by a Brownian motion process has proved to bring considerable simplification and appreciable qualitative insights (Siegmund 1985, Dvoretzky and Wolfowitz 1953, Sirjaev 1973), we build our diagnostic model on a sequential analysis of a Brownian motion. This leads to a closed-form expression for 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.

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 the 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 “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 error costs are incurred, either for the patient or for the HO. The error cost incurred by a $+1$ patient wrongly selecting option $A$ is $c_A$; if a $-1$ patient wrongly selecting option $B$ it is $c_B$. In addition, patients incur an inconvenience cost $c_w$ per unit of time waiting to be served by the nurse line. This can be generalized to a strictly convex waiting cost.

The error costs incurred by the HO are similar: $C_A$ when a $+1$ patient wrongly selects option $A$ and $C_B$ when a $-1$ patient wrongly selects $B$. Table 1 summarizes the HO’s and the patient’s error costs consequent to the treatment option selected.

---

5 We do not include unavoidable costs related to the correct treatment itself; these costs are irrelevant for the management of the nurse line.
In addition, 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.

Table 1  Error costs for the patient and HO as a function of the patient’s pathology and selected treatment.

<table>
<thead>
<tr>
<th>(Patient’s error cost, HO’s error cost)</th>
<th>A (Self care)</th>
<th>B (ED)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta = -1$ (Healthy)</td>
<td>$(0, 0)$</td>
<td>$(c_B, c_B)$</td>
</tr>
<tr>
<td>$\theta = +1$ (Sick)</td>
<td>$(c_A, c_A)$</td>
<td>$(0, 0)$</td>
</tr>
</tbody>
</table>

We assume the following relation holds between the patient’s error cost and the prior 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 have equal expected error costs:

$$\pi c_A = (1 - \pi) c_B.$$  

Under the above assumption, patients are indifferent between the two treatment options without using the nurse line. Thus we assume they select each treatment with prior probability $\frac{1}{2}$ if they do not use the nurse line$^6$.

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 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 faced by a nurse line 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 modeled as a sequential probability ratio test for the drift of a Brownian Motion: $H_0 : \theta r = -r$ against $H_1 : \theta r = r$ ($r$ is the nurse skill level, discussed below). 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 $X (0) = 0$. This Brownian Motion captures the dynamic process of belief updating (Sirjaev 1973), which in our setting is the nurse updating her

$^6$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.
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 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\(^8\). 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 $r \geq 0$ as the nurse’s \textbf{skill level}. A larger $r$ implies that the difference between the two drifts representing the two pathologies is larger, and hence the patient pathology is more easily recognized: A shorter service time is needed to serve a patient in average and the diagnostic accuracy is higher. We refer to the stopping boundary vector $x = (x,y)$ as the \textbf{certainty threshold} or the \textbf{service depth} set by the service provider.

\textbf{Call arrival and service characteristics.} The enrolled members of the HO fall ill (in accordance with Assumption 1) according to a Poisson process with rate $\Lambda$. All such patients independently consider whether to use the nurse line. Their aggregate call rate is denoted by $\lambda (\leq \Lambda)$. We assume there is no reneging.

The performance of this service center is measured by both the probability of misdiagnosis (error probabilities) and the patient waiting time. 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 $Y_{-1}$ first hits $x$ before hitting $y$\(^9\). We also 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, that is, the \textit{under-referral} rate; and $\beta$ is the probability of advising a healthy patient to go to the ED, that is, the \textit{over-referral} rate. 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 $x$ and with the same skill level $r$. 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 $\lambda$.

\(^7\) Under this formulation, Sirjaev (1973) shows that $\pi_t$ can be written as $\frac{\exp(\frac{r}{2}Y_\theta(t))}{1+\frac{r}{2}\exp(\frac{r}{2}Y_\theta(t))}$.

\(^8\) If each patient accepts the advice with a certain probability (see Bunik et al. (2007)), our model still applies.

\(^9\) If $x = y = 0$, then we hit both barriers instantaneously. In this case we assume the HO may use any rule to dispense recommendations.
Patient’s cost savings. The patient’s expected error cost without a nurse line (“pre-call cost”) is \(\frac{1}{2}(1 - \pi) c_B + \frac{1}{2} \pi c_A\) because the probabilities of both type I and II errors are \(\frac{1}{2}\). For a given certainty threshold, the expected error cost with a nurse line (“post-call cost”) is \((1 - \pi) \beta(x) c_B + \pi \alpha(x) c_A\), where \(\alpha(x)\) and \(\beta(x)\) are the error probabilities. Therefore, the patient’s expected cost savings excluding waiting costs are the difference between the “pre-call cost” and the “post-call cost”:

\[
\Delta_p(x) = (1 - \pi) \left( \frac{1}{2} - \beta(x) \right) c_B + \pi \left( \frac{1}{2} - \alpha(x) \right) c_A. \tag{1}
\]

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

\[
\Delta_{HO}(x) = (1 - \pi) \left( \frac{1}{2} - \beta(x) \right) C_B + \pi \left( \frac{1}{2} - \alpha(x) \right) C_A. \tag{2}
\]

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 as \(W(\Lambda; x, m)\) when the aggregated demand rate is \(\lambda\) given \(x\) and \(m\), then the expected utility of a patient calling the nurse line is the difference between the benefit and the waiting cost incurred:

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

Each patient has two pure strategies: To call or not to call 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 as 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):

\[
p_e(x, m) = \begin{cases} 1, & U(\Lambda; x, m) > 0, \\ \in [0, 1], & U(0; x, m) \geq 0 \geq U(\Lambda; x, m), \\ 0, & U(0; x, m) < 0. \end{cases} \tag{4}
\]

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 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 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:

\[
J (x, m) = \Delta_{\text{HO}} (x) \lambda_e (x, m) - c_n m. \tag{5}
\]

We consider \( m \) as a real number, because nurses might work part-time in practice (Poole 2003).

We can now define the optimal HO decision \((x^*, m^*)\), which is the solution of:

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

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c_A ) (( C_A ))</td>
<td>Error cost to patient (HO) of a sick patient not visiting the ED.</td>
</tr>
<tr>
<td>( c_B ) (( C_B ))</td>
<td>Error cost to patient (HO) of a healthy patient falsely visiting the ED.</td>
</tr>
<tr>
<td>( r )</td>
<td>Skill level of a nurse.</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>Variance (noise) of 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 demand rate (patient population size) for the nurse line.</td>
</tr>
<tr>
<td>( \pi )</td>
<td>Prior probability that the patient needs to visit the ED.</td>
</tr>
</tbody>
</table>
4. General Analysis

We first derive expressions that determine the equilibrium demand rates that arise in response to a given HO’s strategy. To do so, we define for a given patient pathology \( \theta \) and service depth \( x \),

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

\( u_\theta (x) \) 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

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

Next we 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, and the drift \( r\theta \), the first moment of the stopping time is

\[
    E_\theta (x) = \frac{1}{\theta r} (x u_\theta (x) + y (1 - u_\theta (x))) \tag{9}
\]

and the second moment of the stopping time is

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

where

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

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 elementary sensitivity properties of the error terms and the first two moments of the stopping time:

**Lemma 2.** (i) \( \frac{\partial}{\partial x} \alpha (x) \geq 0; \frac{\partial}{\partial y} \alpha (x) \geq 0; \frac{\partial}{\partial x} \beta (x) \leq 0; \frac{\partial}{\partial y} \beta (x) \leq 0; \beta (x) \geq 0; \alpha (x) \geq 0; \text{ and} \frac{\partial}{\partial y} (\alpha (x) + \beta (x)) \geq 0. \)

(ii) \( \frac{\partial}{\partial x} E(x) \geq 0; \frac{\partial}{\partial x} V(x) \geq 0; \frac{\partial}{\partial y} E(x) \leq 0 \); and \( \frac{\partial}{\partial y} V(x) \leq 0. \)
Lemma 2 establishes that the main trade-offs in our model are consistent with reality: Recall that $\alpha$ is the type I error probability ($Y_{+1}$ first hits $y$ before hitting $x$), and $\beta$ is the type II error probability ($Y_{-1}$ first hits $x$ before hitting $y$). When the positive threshold $x$ increases, $\beta$ is reduced while $\alpha$ increases because the Brownian motion is more likely to hit the negative threshold. In addition, the overall sum of error levels decreases as $x$ is larger or $y$ is smaller: a larger service depth — a longer test — logically implies the diagnosis is more likely to be correct. This error reduction comes at a cost of increased congestion: Both the first and second moments of service time increase when the service depth becomes larger.

As the service depth impacts both the first and second moments of the service time, it will impact the customer waiting time and hence also the number of nurses to hire. To quantify these relationships, we use Whitt’s simple heavy traffic approximation (Whitt 1993) for the expected waiting time in queue for a nurse line with $m$ nurses\textsuperscript{10}:

$$ W(\lambda; x, m) = \frac{1}{m^2} \frac{\lambda V(x)}{2(1 - \frac{\lambda F(x)}{m})}. $$

(13)

Note that the special case of $m = 1$ reduces to the P-K formula (Ross 2002).

Given the expressions for the patient’s 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 $U(\lambda; x, m) = 0$:

$$ \lambda(x, m) = \frac{m}{\lambda x + E(x)}. $$

(14)

Then the patient equilibrium decision is:

$$ pe(x, m) = \begin{cases} 1, & \lambda < \lambda(x, m), \\ \frac{\lambda(x, m)}{\Lambda}, & \Delta_p(x) > 0 \text{ and } \lambda(x, m) \leq \Lambda, \\ 0, & \Delta_p(x) \leq 0, \end{cases} $$

(15)

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

Next we analyze the optimization problem (6) and derive the following result concerning 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.

\textsuperscript{10}Whitt’s formula is valid for a positive, real $m$. 
This “capture-all-demand” staffing result is driven by the statistical economies of scale in multi-server queueing systems: “beyond nominal requirements, in order to achieve the target service level in the face of stochastic variability... the required excess capacity grows less than proportionately with the load of calls to be handled (Garnett et al. 2002).” We describe how this idea is reflected in the strategic queueing setting of nurse lines: From Equation (14), we can see that at any target accuracy (any fixed certainty threshold $x$), the call volume per nurse $\lambda/m$ increases as the number of nurses $m$ increases — thus the revenue from the increased demand rate grows more than proportionately with the staffing cost$^{11}$. As a result, the profit rate is increasing with the staffing level until the call volume reaches the upper bound $\Lambda$. Once $\lambda = \Lambda$, the profit rate declines as the staffing level increases. Hence, the optimal staffing level should be set at the minimum level to capture the maximum call volume$^{12}$.

Extending queueing models with strategic agents (for example, as in Hassin and Haviv 2003) to incorporate the statistical economies of scale in multi-server queueing systems changes the characteristic of the equilibrium demand. In the widely-studied single server setting (Hassin and Haviv 2003), the equilibrium can be one of three cases: “everybody calls,” “no-one calls” or “some fraction of customers call.” In contrast, as shown in our model, when more servers may be hired at a linear cost per server, the equilibrium resulting from the optimal investment in servers is always “everybody calls,” not any other case$^{13}$.

Recall that 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 captures all of the demand. Thus, the results in Proposition 1 significantly simplifies our analysis by reducing the degrees of freedom of the optimization problem. Next, based on this simplification, we derive conditions specifying when it is optimal to invest:

**PROPOSITION 2.** 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:

$^{11}$ Assuming that the total staffing cost increases linearly in the number of nurses.

$^{12}$ This result does not depend on the assumption of homogeneity of customers: For patients who are heterogeneous in their prior, unit waiting cost or error cost, this result still holds. This is because the key condition for Proposition 1 remains true: the call volume per nurse $\lambda/m$ still increases as the number of nurses $m$ increases.

$^{13}$ Complementary to our paper, Wang et al. (2007) study the nurse line problem assuming the staffing level is fixed at $m$. They show that it is possible that at the optimal certainty threshold, the equilibrium calling probability is strictly less than one: It is not always optimal to have “everybody 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 result in our paper.
\[
\Delta_{HO}(x) > \frac{1}{2} c_n E(x) \left( 1 + \sqrt{1 + \frac{2c_w}{\Lambda} \cdot \frac{V(x)}{\Delta P(x) E^2(x)}} \right) 
\] (16)

Staffing cost is one of the major concerns in investing in a call center (Gans et al. 2003). Condition (16) in Proposition 2 indicates that if at any accuracy (fixed value of \(x\)), the benefit from serving each individual patient is greater than the staffing cost per patient to maintain that accuracy level and capture all the demand, then the overall profit is positive. Due to the concavity of the benefit function and the convexity of the expected service time in the small neighborhood of \(x = 0\), we can always find (numerically) a small value of \(x\) such that the overall profit is positive\(^{14}\). This means for any unit staffing cost (even if it is very high) and for any cost saving potential (even if it is very low), it is optimal to invest in a nurse line.

This is somewhat surprising and needs to be interpreted carefully within the boundaries of the model, as other factors may affect this result. First, if there exists a fixed cost when investing in a nurse line, for example relating to call center technology (hardware and software), the cost savings will need to be sufficiently large — not just positive — to cover the fixed cost\(^{15}\). In this case there will exist a threshold level of unit waiting cost above which, and a threshold level of cost saving potential below which, the nurse line savings do not cover the strictly positive fixed cost and hence, investment is not optimal. Second, an external regulator might impose upper bounds on the error probabilities, implying the certainty threshold \(x\) cannot be very small. In this case, again, there may not exist a certainty threshold \(x\) such that the profit rate is positive. Third, nurses may spend some time that is required to greet the caller and identify the symptom from the caller before invoking the protocol. This extra time is independent of the service depth and will imply that the unit waiting cost cannot exceed a threshold level in order to make investing in the nurse line optimal.

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 a model with infinitesimal asymmetry (Section 6): Starting from the symmetric case, we change the HO’s cost of one error infinitesimally while keeping the

---

\(^{14}\) We can prove analytically that for the symmetric case (\(c_A = c_B\), \(C_A = C_B\) and the prior is \(\pi = 1/2\)) there always exists a pair \(x\) that satisfies (16). We conjecture this holds for the asymmetric case as well.

\(^{15}\) In this case, condition (16) should be changed to \(\Delta_{HO}(x) > \frac{1}{2} c_n E(x) \left( 1 + \sqrt{1 + \frac{2c_w}{\Lambda} \cdot \frac{V(x)}{\Delta P(x) E^2(x)}} \right) + \frac{K}{\Lambda}\), where \(K\) is the fixed cost.
cost of the other error constant. We complement these results in Section 7, in which we numerically explore settings with generally asymmetric costs.

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 \( \tau = 1/2 \)\(^{16} \). 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 error probability (either type I or type II) in the symmetric case as a function of \( x \). Thus 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:

\[
J^* = \max_{x \geq 0; m \geq 0} J(x, m) \doteq \Delta_{HO}(x) \lambda_{e}(x, m) - c_n m. \tag{17}
\]

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

**Lemma 4.** The optimal number of nurses is determined by \( m^* = \Lambda N(x^*) \), where

\[
J^* = \max_{x \geq 0} \hat{J}(x) \doteq (\Delta_{HO}(x) - c_n N(x)) \Lambda
\]

and

\[
N(x) = E(x) \left( \frac{1}{2} + \frac{1}{2} \sqrt{1 + \frac{2c_w 1 + C^2(x)}{\Lambda \Delta_{P}(x)}} \right) \tag{18}
\]

where \( C^2(x) = \frac{V(x) - E^2(x)}{E^2(x)} \).

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, \( \Delta_{P}(x) = c_w W(\Lambda; x, m) \). Following Lemma 4, the staffing decision variable is eliminated from the optimization problem, and only the optimal service depth remains. With this simplified form, we can establish that \( \hat{J}(x) \) has a unique optimizer over \([0, +\infty]\) (see the appendix for a proof).

Next, we study how the optimal decisions change with parameters in Proposition 3.

**Proposition 3.** When \( c_w > 0 \), we have the comparative statics in Table 3.

Our result suggests that accuracy designed for the nurse line should be based on the population size, which is typically ignored by managers (Poole 2003). As shown in Table 3, when the population

\(^{16}\) Notice that Assumption 1 is satisfied in this case.
Table 3  Sensitivity Analysis for Symmetric Case 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>$\Lambda$</td>
<td>Population size</td>
<td>Negative</td>
<td>Positive</td>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>$r$</td>
<td>Nurse skill</td>
<td>Negative</td>
<td>Pos./Neg.</td>
<td>Positive</td>
<td>Pos./Neg.</td>
</tr>
<tr>
<td>$c_A$</td>
<td>Patient’s error 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>$C_A$</td>
<td>HO’s error 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>
</tbody>
</table>

size $\Lambda$ increases, the optimal error rate decreases. This means that a diagnostic service center located in a high-patient-density area or one which aggregates demand from many regions should operate at a lower error rate; in these cases, a diagnostic service center can bring higher return-on-investment in accuracy for the manager. (Recall Proposition 1 states the optimal demand rate is always equal to the population size.) Thus the greater the population size, the larger the increase in the profit rate resulting from the increase of the individual benefit (accuracy of advice).

Note that without congestion externality ($c_w = 0$), the optimization problem in Lemma 4 can be reduced to a classical continuous time hypothesis testing problem (see e.g. Sirjaev (1973) and Dvoretzky and Wolfowitz (1953) for an analysis)

$$\max_{x \geq 0} \Delta_{HO}(x) - c_n E(x),$$

where a trade-off is made between the cost of additional experimentation $c_n E(x)$ and the type I/type II error reduction $\Delta_{HO}(x)$. According to traditional hypothesis testing theory, the optimal error rate is not related to the population size. But for diagnostic service centers, due to congestion externality, the optimal error rate 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 addition, Proposition 3 shows that if the nurse line hires more highly skilled nurses, callers will experience a longer wait on average. Recall that in our model, fixing other variables and parameters, an increase in the nurse skill $r$ (the magnitude of the Brownian Motion’s drift) both reduces error and accelerates the testing process. Thus directly, increasing the skill level leads to an increase of accuracy $(\frac{1}{2} - \alpha(x))$ and thus to the individual benefit $\Delta_P(x)$. At the equilibrium, the customer’s waiting cost is always balanced with the individual benefit (because utility is zero), so the increased individual benefit implies the longer waiting time. The intuition is that increased accuracy (individual benefit) makes patients more tolerant of long waiting times. We also notice that in contrast to the waiting time, the change in the optimal staffing level or the service depth remains unclear.

$^{17}$ In this case, from Lemma 4, we can see the HO’s objective function becomes $(\Delta_{HO}(x) - c_n E(x))\Lambda$. 
Similarly, Proposition 3 also shows that adding capacity (staffing) may lead to the deterioration of performance measures: increasing congestion or error, but not both (as again, utility must balance at zero). Due to the discretionary property of diagnostic service, capacity (staffing) not only impacts congestion but also impacts the error rate. As the manager optimizes the overall effect, in different parameter settings the added capacity is used to focus on one of the two different effects: either reducing congestion or increasing accuracy. Thus as shown in Table 3, counter-intuitively, (i) adding capacity may actually increase congestion, and (ii) adding capacity may actually increase error rates. Our result proves analytically an observation by Hopp et al. (2007) in a centralized discretionary service system: adding capacity may increase congestion.

Following this point, we notice that the two performance measures — waiting time and error probability — are substitutes: When one becomes worse, the other becomes better. This again stems from the fact that 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 error 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 probability decreases), the waiting cost (i.e. the expected waiting time) increases.

Finally, Proposition 3 shows that the sensitivity of staffing is complicated. The optimal staffing level monotonically increases as the HO’s error cost $C_A$ increases, because the HO’s error cost is a major motivation for investment in a nurse line: A larger error cost implies a high cost saving potential. In contrast, the patient’s error cost $c_A$ has an ambiguous impact on the optimal staffing level: On one hand, a higher patient’s error cost will increase the cost saving potential 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, the HO also cares (indirectly) about the patient’s cost savings (as it influences the nurse line demand volume and hence profits). Thus, the HO wants to reduce the error rate when it becomes more expensive for the patient. However, a decrease in the optimal error rate requires deeper service (longer time) and hence might require an increase in the number of nurses. Thus, the net impact of an increase of the patient’s error costs on the staffing level can go in either direction. Similarly, the impact of the waiting cost on the staffing level is also ambiguous, but opposite to the impact of the patient’s error cost.
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 error costs $C_A$, i.e. the cost to the HO 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:

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

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, $(C_B, c_A, c_B, \pi, \sigma, c_w, c_n, \Lambda)$, constant. Then we study the changes in error rates, waiting time and staffing. In Subsection 7.2, we will see that the analytical comparative statics obtained in the small neighborhood of the symmetric case also all numerically hold in the case with a generally asymmetric structure.

**Proposition 4.** When $C_A$ changes from the symmetric case, we have the comparative statics illustrated in Table 4.

### Table 4 Impact of Asymmetry (increasing $C_A$ to $C_A + \delta$).

<table>
<thead>
<tr>
<th>$c_w$</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>$c_w &gt; 0$</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>

We compare the general case of a positive waiting cost (congestion externality exists) with a classical hypothesis testing problem (the $c_w = 0$ case) in Table 4. We find that in the general case, increasing asymmetry in error costs may impact the two error rates either in the same or opposite directions: When the HO’s error cost of one option increases, the corresponding error decreases, and the other error may increase or decrease. In contrast, in the classical hypothesis testing problem, both errors are always substitutes: When the HO’s error cost of one option increases, the corresponding error decreases, and the other error always increases to control the testing cost $c_n E(x)$. Thus congestion externalities change the fundamental characteristics of the hypothesis testing problem. Specifically, the testing cost structure is changed from $c_n E(x)$ to $c_n N(x)$. Thus as shown in Equation (18), the squared coefficient of variation of the service time ($C^2(x)$) plays an
important role in addition to the expected service time\textsuperscript{18}. So as the HO’s error cost $C_A$ increases above $C_B$, the two errors may actually decrease together.

The intuition behind our result is that due to congestion externality, the staffing and error rate decisions are related. As one error cost increases, the overall benefit of investing in a nurse line increases, which justifies a higher investment in staffing. The increased capacity might be used to decrease the other error rate. Our result reminds managers to be aware of the complexity of the cost factor’s impact on the system. For example, although the cost structure in an emergency nurse line is more asymmetric than in a dietitian nurse line, we cannot easily conclude that the asymmetry in error rates resulting from the former is higher than the latter.

7. Numerical Experiments and Extensions

After discussing parameter values and a baseline setting in Subsection 7.1, we conduct numerical experiments to obtain insights into the following questions: (1) Will the analytical comparative statics obtained in the small neighborhood of the symmetric case still hold under a generally asymmetric HO error cost? (Subsection 7.2.) (2) What are the sensitivity results when the patient’s cost changes and Assumption 1 may not be satisfied? (Subsection 7.3.)

7.1. Parameters and Base Case

We construct our numerical study using data ranges for call volume and staffing compensation inferred from an American Academy of Pediatrics guide for developing a telephone advice system for a pediatric office practice (Poole 2003, page 121 and page 161). We test the impact of changing parameter values around the base case, in the ranges we show in Table 5.

<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>$20 \sim 500$</td>
<td>$60$</td>
</tr>
<tr>
<td>Skill Level; $r$</td>
<td>NA</td>
<td>$1 \sim 2$</td>
<td>$1$</td>
</tr>
<tr>
<td>Noise on diagnosis; $\sigma^2$</td>
<td>NA</td>
<td>$1$</td>
<td>$1$</td>
</tr>
<tr>
<td>Insurer’s Cost; $C_A = C_B$</td>
<td>$\text{$ per visit}$</td>
<td>$20 \sim 250$</td>
<td>$50$</td>
</tr>
<tr>
<td>Patient’s Cost; $c_A = c_B$</td>
<td>$\text{$ per visit}$</td>
<td>$1 \sim 40$</td>
<td>$25$</td>
</tr>
<tr>
<td>Staffing Cost; $c_n$</td>
<td>$\text{$ 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>$\text{$ per minute}$</td>
<td>$0 \sim 10$</td>
<td>$4$</td>
</tr>
</tbody>
</table>

We validate the data of the base case by estimating its performance measures and staffing level and comparing these with values in practice. We rewrite the objective function as a function of $\text{x}$.

\textsuperscript{18} Although Equation (18) is derived in the symmetric case, it can be extended to the asymmetric case. In the asymmetric case, $x$ is replaced with $\text{x}$. 
the error rate $\alpha$ and solve the first order condition\textsuperscript{19}. Ignoring 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 $4/minute\textsuperscript{20}$, the error level increases to 3.28%, the service time decreases to 6.32 minutes, the waiting time is 2.92 minutes, and the number of nurses increases to 3.17. These numbers are of the same order of magnitude as reported by Poole (2003).

### 7.2. Changing the HO’s error 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.

**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.

![Figure 1](image-url)  
*Figure 1*  
Left: 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, $\alpha^* = \beta^*$. Right: 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, $C_A = C_B = 50$.

**Observations:** Figure 1 (left panel) confirms that without congestion externalities ($c_w = 0$) (the bottom curve), when the HO’s error cost of one option ($C_A$) increases while the other ($C_B$) remains constant, there is a substitution effect among the errors: One error rate increases while the other decreases. In the same figure, with congestion externalities ($c_w > 0$) (the top curve), this substitution effect is greatly weakened and may even become a (weak) complementary effect: Two error

\textsuperscript{19}See Equation (??) in the appendix.

\textsuperscript{20}We have conducted experiments with other values of waiting costs, and the results were qualitatively the same as those reported in this section.
rates increase together. Figure 1 (right panel) indicates that in the asymmetric case, the optimal staffing level increases as one of the HO’s error costs increases. We proved these results in a small neighborhood of the symmetric cost structure (Proposition 4). Our numerical studies indicate that when the cost parameters are highly asymmetric, these results remain true: In the presence of congestion externality, the error rates may both decrease due to the increase of one error cost, as this triggers a higher investment in capacity, which might improve the other error rate.

7.3. Changing the patients’ error costs

Another way of introducing an asymmetric cost structure is to change the patients’ error costs. In this case, Assumption 1 may not be satisfied and thus the probability of choosing option A without consulting a nurse line may not be \( \frac{1}{2} \). We illustrate the impact of asymmetry in the patients’ cost structure 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.

![Figure 2](image_url)

**Observations:** As seen in Figure 2 (left panel), similarly to when changing the HO’s error costs, the error probabilities are either substitutes or complements with respect to changes of the patients’ error costs (due to a similar congestion effect as discussed in Proposition 4). But different from changing the HO’s error costs, the staffing level is non-monotonic: as shown in Figure 2 (right panel), 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 \( \pi c_A > (1 - \pi) c_B \) (or equivalently in this...
example $c_A > c_B$). When $\pi c_A < (1 - \pi) c_B$, 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, $\pi c_A$ and $(1 - \pi) \beta (x) c_B + \pi \alpha (x) c_A$. The difference between the pre-call and the post-call expected costs is $\Delta_P (x) = -(1 - \pi) \beta (x) c_B + \pi (1 - \alpha (x)) c_A$. This individual patient benefit $\Delta_P (x)$ tends to increase in the patient’s error cost $c_A$ (disregarding second order effects). Thus patients have more incentive to use the nurse line and the nurse line manager can decrease the number of nurses to lower the staffing cost as $c_A$ increases.

In contrast, when $(1 - \pi) c_B < c_A$, without a nurse line, the patient always selects the option $B$. The pre-call expected cost is $(1 - \pi) c_B$ and the post-call expected cost remains $(1 - \pi) \beta (x) c_B + \pi \alpha (x) c_A$. Thus $c_A$ only plays a role in the post-call expected costs, as the nurse line could mistakenly advise option $A$ to a $\theta = +1$ patient. The difference between the pre-call and the post-call expected costs is $\Delta_P (x) = (1 - \pi) (1 - \beta (x)) c_B - \pi \alpha (x) c_A$. So the increase of $c_A$ reduces the patients’ incentive to use the service. In order to keep the optimal demand rate $\Lambda$, the nurse line manager has to (slowly) increase the staffing level.

8. Relevance of Integrated Model and Analysis to Practice

In this section, we summarize our findings from interviews with current nurse line practitioners and our reviews of clinical studies, relating our research model, analysis and insights to nurse line practice. We provide more details in a companion paper, the authors (2010).

Typical Nurse Line Service Process. As the protocols used are symptom-based, first, after the greeting, the nurse recognizes symptoms. Examples of typical symptoms are abdominal pain, nausea and vomiting, or respiratory problems. The nurse interviews the patient and records a description of the problem in the patient’s own words. Then the nurse asks 6-7 general questions to gather more information. Based on the key words in these pieces of information, the software suggests appropriate protocols; there may be multiple protocols appropriate for one patient. The nurse then chooses what she thinks is the most appropriate one. Within a certain protocol, a series of Yes/No questions are asked and then the final recommendation is generated (911/ED/urgent care/home care). The nurse may either: (i) Accept this recommendation; (ii) Suggest a more severe recommendation or (iii) Initiate another protocol. In the third case, the nurse must recommend the more severe of the protocols’ outcomes. The nurse must follow the protocol exactly, but protocols may involve branches – for example the vomiting guideline may ask if there was a recent head injury. If the answer is yes then it advises the nurse to switch to the head injury protocol, and asks
if she would like to do so (Fiorenzio 2009). A nurse may also terminate one protocol and initiate another at any time.

**Accuracy/Response-Time Trade Off.** A consistent conclusion emerging from our interviews is that minimizing response times may actually *interact/conflict* with increasing the accuracy of the nurse’s recommendation, although these two targets are set independently in practice. We give two examples of such interactions:

First, with a frequency of $10^{-2} - 20\%$, nurses follow two (or more) protocols instead of just one for a given symptom. This is typically the case when symptoms are somewhat ambiguous. Using multiple protocols may be discouraged by the nurse line manager due to the *trade-off* between gaining more accuracy and decreasing waiting time. An interviewee (Fiorenzio 2009) mentions that during busy times, multiple protocols may be followed less frequently.

Another example is when coping with emergency events or disease outbreaks. Whenever an emergency event (for example, the H1N1 virus, the avian flu, etc.) occurs, nurse line volumes typically increase significantly, severely straining the staffing at nurse lines. Moreover, in the face of such emergent medical situations, protocols may need to be updated, to guarantee *appropriateness of the recommendation*. Prior to receiving an update, nurses err on the safe side and typically take a longer time to dispense callers (they also may have to look up other sources of information – such as from the Centers for Disease Control and Prevention – rather than operating from a self-contained protocol). Here again the trade-off between accuracy and waiting time is apparent. Note that the protocol decision can influence this tradeoff: Using a protocol that is quickly updated to incorporate emergent conditions eases the trade-off between accuracy and waiting time. Equipped with updated protocols, nurses can maintain high accuracy of advice without sacrificing efficiency (waiting time) in face of emergency events.

In our model, we use service depth to capture the different protocol-related choices affecting accuracy/response time tradeoff that the nurse-line can make.

**Importance of Nurse Skill.** Interviewees also confirm that nurses are making *discretionary* decisions: For example, whether/when to discourage multiple protocols depends on the benefits of more accurate decision vs. the extra time that is consumed by the additional protocol, or when to end a protocol before it makes a recommendation and switch to another.

To make the above complex discretionary decisions (Hopp et al. 2007), nurses must have the ability to think critically and quickly, and apply deductive reasoning. This is captured in our *nurse skill* level. For example, at the nurse line we interviewed, a well-trained nurse can serve 6 calls/hour
while a new nurse can only serve 4 calls/hour. Holding other conditions are the same, a nurse with
a higher skill level works both faster and better, consistent with our model. When deciding among
protocols, the nurse’s experience/skill is very important, as this helps the nurse determine which
protocols to select first, and also whether to use multiple protocols. In addition, experienced nurses
may know the different protocols by heart, which accelerates the diagnostic process.

Protocol Implementation. An important high-level decision is how protocols are to be selected
and implemented. For example, there exists significant variation in page length and soft-ware
features of currently available protocol sets and their implementations\(^{21}\). In addition, different
protocol implementations may result in different diagnostic performance for a given symptom
(Wheeler and Siebelt 1997) as well as call duration (Bunik et al. 2007). For example, compared
to a non-computerized protocol implementation, a computerized protocol can suggest the most
relevant protocols more promptly as it is based on automated key word searches instead of relying
on a nurse’s memory. In general, a lower-skilled nurse equipped with an advanced computer system
can perform as well as a more highly-skilled nurse using a book, in terms of speed and accuracy
(Fiorenzio 2009, Massaro 2009). Thus how to implement protocols in the service process is just as
important as other decisions such as staffing and training\(^{22}\).

Impact of Call Volume. From the interviews, we learned that decisions regarding protocols
(i.e., service depth) are not consciously related to call volume. Our paper provides a different
perspective: In order to maximize the total benefit from investing in a nurse line, the accuracy
target for the nurse line should be based on the population size. This result implies that large
nurse lines facing many potential patients have more incentive to improve the appropriateness of
recommendations. In practice, one nurse line we interviewed (NurseAdvice New Mexico) reaches
about 1 million inhabitants, half of New Mexico’s population. This nurse line works with computer-
ized implementation. Smaller nurse lines, for example cooperatives between individual physicians,
might use manual protocols instead of computerized systems, which some practitioners view as less
accurate. Further, some small nurse lines may even do telephone triage without protocols, which is

\(^{21}\) Examples of paper-based protocols (adult and pediatric books): (1) Simonsen, (2) Reisman & Stevens (3) Briggs,
Lafferty and Bair. Pediatric only books: (1) Schmitt, (2) Brown, (3) Katz. The books range from 94 to 668 pages.
Vendors of software for protocol implementation: (1) Clinical Solutions SW (uses Teleguide\(^{\text{TM}}\) Algorithms proto-
cols), (2) HealthLine Systems, Inc., Sharp Focus (uses the Cleveland Clinic Foundation protocols), (3) LVM Sys-
tems. (uses Schmitt/Thompson protocols), (4) McKesson’s, ASK-A-NURSE program and (5) Fonemed, LLC. (uses
Schmitt/Thompson protocols).

\(^{22}\) In practice the nurse’s skill level will be an important intermediate variable that influences the relationship between
the protocol form and the performance characteristics. The nurse’s skill level can be influenced by nurse training and
development.
considered very dangerous. This phenomenon suggests that as nurse line size increases, error rates may indeed decrease as predicted by our model.

**Cost Saving Model.** Recall in Subsection 3.1, we define the cost savings for the patient and the HO from a nurse line. This cost saving model complements current state-of-the-art methods in the clinical research (Bunik et al. 2007). These researchers typically measure the cost saving of a nurse line by taking the difference between costs of a patient’ first experienced level of care without a nurse line and with a nurse line (Bunik et al. 2007); they do not consider whether a patient’s choice is the most appropriate level of care or not. Their cost-saving model justifies the existence of nurse lines *only when* callers have a bias toward urgent care and nurse lines can direct them to non-urgent (less expensive) care. Therefore it under-estimates the cost-saving potential of nurse lines. For example, when a patient’s intention before calling is to stay at home and after calling is to seek urgent care, the cost saving is measured as the difference between costs of his first experienced level of care — staying at home versus urgent care, which is obviously negative. In addition, according to this cost-saving model, it would be optimal for the nurse line to always recommend the cheapest option (i.e. non-urgent care), which is not realistic. Thus this cost-saving model is not sufficient.

As improvement and complementation, our nurse line model counts all the costs incurred until patients are cured, including patients’ first experienced level of care. In other words, these costs depend on how far a patient’s choice is from the most appropriate level of care (i.e., how accurate this choice is), without a nurse line or with a nurse line. Then no matter what the patient’s intention before calling is, improved accuracy through taking nurse line’s recommendation always produces positive savings. Our model is consistent with the typical situation in practice: Assuming the nurse line’s recommendation is more appropriate than a caller’s initial intention, if without the nurse line the caller would have sought home care, then his/her health status would likely deteriorate and the patient might end up ultimately obtaining urgent care. Therefore, the caller’s cost before calling should include both home care and urgent care. Thus the cost saving is positive because the patient avoids the extra cost of home care by using the nurse line (as well as any deterioration of condition due to the treatment delay).

**Error Rates.** In practice, asymmetry in error rates does exist (under-referrals occur less frequently than over-referrals (Bunik et al. 2007)), but how they are influenced by error costs (i.e., cost per error) has not been studied. Our analysis shows that it is more complicated than what intuition or hypothesis testing theory may indicate, because changes in error costs can impact error
rates both directly and indirectly through staffing: Staffing decision and error rates decisions are related due to congestion externality. As one error cost increases, the overall benefit of investing in a nurse line increases, leading to a higher investment in staffing (Proposition 7). The increased capacity in turn might be used to decrease the other error rate. In practice, nurse lines might face different levels of asymmetry in error cost. For example, a pediatric nurse line is very different from a nurse line serving chronically ill adults. Our result reminds managers to weigh the two error rates carefully even when one error cost is obviously higher.

**Staffing.** In practice staffing decisions are solely based on “statistical analysis of historical demand data” (Massaro 2009), which typically follow call center theories without concerning appropriateness of advice or the discretionary nature of diagnostic service. But due to this discretionary nature, adding capacity (staffing) can actually allow for longer service which increases accuracy, thus having an impact on both congestion and the error rate. As managers strive to optimize the overall effects of service efficiency and accuracy, our model shows that when reducing congestion is the dominant factor for profitability, the added capacity should be used to relieve congestion, but may reduce accuracy. When increasing accuracy is the dominant factor for profitability, the added capacity improves accuracy, but may increase congestion.

We compare the optimal (symmetric) staffing level $m^* = \Lambda N(x^*) = \Lambda E(x^*)(\frac{1}{2} + \frac{1}{2} \sqrt{1 + \frac{2c_w(1+C^2(x^*))^2}{\Lambda \Delta p(x^*)}})$ (18) with a heuristic used in practice. As mentioned in Poole (2003), in practice, staffing is set as the product of the call volume and the estimated expected service time. This heuristic is exactly a simplified version of Equation (18) that ignores congestion. 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 capacity. Our model identifies the optimal “safety factor” as $\frac{1}{2} + \frac{1}{2} \sqrt{1 + \frac{2c_w(1+C^2(x^*))^2}{\Lambda \Delta p(x^*)}}$. This factor captures the impacts of both the service accuracy requirement and stochastic variability. For a fixed service depth $x^*$, the optimal “safety factor” is high when the patient’s unit waiting cost $c_w$ is high, the maximum call volume $\Lambda$ is low or the patient’s error cost $c_A$ is low (since $\Delta p(x^*) = c_A^2(1 - \alpha(x^*))$). In these cases, the deviation of that heuristic from optimal can be high.

9. **Discussion and Conclusion**

We model a type of diagnostic service center motivated by nurse lines: Call centers that provide medical advice on the most appropriate course of action according to callers’ symptoms. At these service centers, the value — for both the customer and the provider — results from the correctness
of the decision arrived at during the service. Therefore, different from traditional models, the value for the service provider is determined by the product of the accuracy of advice and the demand volume aggregated from individual patient’s choices of whether to use the service center or not. Since longer service typically entails higher accuracy but also more congestion, this unique feature coupled with customers’ choices presents an important operational challenge for managers: They must jointly optimize service depth (accuracy) and staffing. While some empirical studies in the clinical literature have shown that nurse lines may generate savings (Lattimer et al. 2000), there are no analytical models studying how to manage them to optimize cost savings net of staffing costs. Our paper helps build such an analytical framework incorporating the complex interplay of diagnostic accuracy, system congestion, staffing and agent skill level within a strategic queueing setting.

We build this framework using two elements — an $M/G/m$ queueing system and a continuous time sequential testing process. Previous research has only studied these two elements separately; by integrating the two we generate new insights, in addition to extensions of well-known principles. We find that the dual concerns of accuracy and congestion leads to a counter-intuitive impact of capacity investment: Service centers staffed with highly skilled (better trained and more experienced) workers are often believed to relieve system congestion. In the nurse line setting, however, increasing skill level actually increases congestion because the service provider can optimally increase the time/accuracy serving each individual without reducing demand. Similarly, in different situations, increasing staffing level might lead to a deterioration in system performance as the manager focuses on one of these dual concerns: adding capacity may actually increase congestion or adding capacity may actually increase error rates. Recent work on the management of discretionary service processes (Hopp et al. 2006 and Hopp et al. 2007) numerically show that in a centralized setting, increasing capacity may also intensify congestion. Our analysis extends this finding into a decentralized setting and gives a theoretical proof.

Our unique joint staffing and protocol implementation problem for nurse line practice also generates insights previous work has not identified, such as: (i) Population size is an important driver in nurse line decisions, not only affecting staffing, but also error probabilities. (ii) Increasing asymmetry in error costs may impact two error rates either in the same direction or in opposite directions. (iii) The optimal staffing level monotonically increases as the HO’s cost-saving potential (error cost) increases, while in contrast, the optimal staffing level might increase or decrease as the patient’s cost-saving potential (error cost) increases.
In summary, our work demonstrates that the tradeoff between accuracy and congestion motivated by diagnostic service centers in the health care domain changes many aspects of traditional call center design and staffing. Due to greater needs from dealing with emergency events such as disease outbreaks, nurse lines play a more important role in advising anxious people and reducing infection and congestion at key health facilities. Toward extending our work to these real applications, more elaborate decision models could be developed to incorporate dynamic staffing/service control decisions, diagnostic processes with multiple pathologies and treatments (e.g. call 911, visit ED, visit physician, self-care), different quality measures (e.g. the busy probability), non-linear waiting costs and/or non-linear staffing costs, or with accuracy (and/or wait) standards imposed by a regulator. Note that from a nurse line decision maker’s perspective, protocols are symptom based and symptoms determine different call durations and possible outcomes. Thus another possible extension is to study the optimal service depth for each major symptom with different error costs and waiting costs.

10. Acknowledgement
The authors thank the seminar participants at Carnegie Mellon, Columbia, Cornell, The Wharton School and The University of Chicago. The authors thank interviewees: Connie Fiorenzio at NurseAdvice New Mexico, Kelli Massaro and Teresa Hegarty at After Hours Telephone Care Program affiliated with the Children’s Hospital, Aurora, CO, and Maya Bunik at the child health clinic at Children’s Hospital, Denver, Colorado. Xiaofang Wang wishes to thank the people at Children’s Hospital at Pittsburgh whose help motivated this research. Laurens Debo wishes to acknowledge summer support from the University of Chicago Booth School of Business. The authors are grateful for the invaluable comments and suggestions provided by the associate editor and the referees on earlier versions of this paper.

References

23 Note that we focus on static policies in our model, i.e. the service depth does not depend on the number of waiting calls. De Vericourt and Sun (2009) develop dynamic policies in a slightly different context (with single-sided errors and discrete questions). Although we are not aware of such implementations in nurse line practice, we think this is an interesting research direction: For example, in computerized protocols, the protocol depth could be dynamically adjusted as a function of the congestion level. LVM Systems’ E-Centaurus software implementation of Barton Schmitt’s protocols provides flexibility for queue management for calls on hold, which could in principle be linked back to the protocols. See: http://www.lvmsystems.com/download/telehealth.pdf.”


authors. 2010. Managing nurse lines — practical challenges and the developing theory. Working paper.


