1-2012

Imaging Room and Beyond: The Underlying Economics Behind Physicians' Test-Ordering Behavior in Outpatient Services

Mustafa Akan
Carnegie Mellon University, akan@cmu.edu

Tinglong Dai
Carnegie, dai@cmu.edu

Sridhar Tayur
Carnegie Mellon University, stayur@andrew.cmu.edu

Imaging Room and Beyond: The Underlying Economics Behind Physicians' Test-Ordering Behavior in Outpatient Services

Tinglong Dai Mustafa Akan Sridhar Tayur
Tepper School of Business
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
{dai, akan, stayur}@andrew.cmu.edu

Excessive diagnostic tests have long been viewed as one major aspect of the inefficiency in the healthcare system and are often attributed to the fee-for-service payment model. In this study, we investigate the underlying operational and economic drives behind physicians' test-ordering behavior in an outpatient setting. We model and analyze the strategic interaction between a single physician and a group of patients with health insurance coverage. We then investigate the effect of different service settings. First, setting a low reimbursement ceiling alone cannot eliminate overtesting. Second, the joint effect of misdiagnosis concerns and insurance coverage can lead to both overtesting and undertesting, which differs from popular beliefs about the effect of misdiagnosis concerns. Third, patient heterogeneity can further encourage physicians to overtest in order to cherry-pick patients. Fourth, we show that our main insights carry over when the physician can increase the service quality through reading and analyzing the results from diagnostic tests. Last, we consider asymmetric information in physician type and find that physicians' signaling efforts can lead to more salient overtesting behavior, especially when technology advancements flatten out differentiation among physicians.

Key words: Healthcare operations, test-ordering behavior, physician-patient interaction

1. Introduction

Over the last few decades, a strong consensus has emerged among patients, physicians, and policy makers that health care is not delivered efficiently in the United States. One major aspect of the inefficiency in the healthcare system is the prescription of unnecessary diagnostic tests and medical procedures by physicians (hereafter referred to as "overtesting"). The Congressional Budget Office estimated that around $700 billion per year, or 5 percent of the nation's GDP, is spent on tests, treatments, and care that do not actually improve health outcomes (Orszag 2008).

Various explanations have been suggested for overtesting. The most commonly cited one is misaligned monetary incentives (Jauhar 2009). This is manifested in President Barack Obama's description of the health care industry as "a system of incentives where the more tests and services

are provided, the more money we pay, ... a model that rewards the quantity of care rather than the quality of care; [a model] that pushes you, the doctor, to see more and more patients even if you can't spend much time with each; and gives you every incentive to order that extra MRI or EKG, even if it's not truly necessary, ... a model that has taken the pursuit of medicine from a profession—a calling—to a business" (White House 2009).

While overtesting is often believed to result from physicians' desire to collect more revenue as they order more tests (i.e., the fee-for-service model), our collaborative study with the University of Pittsburgh Medical Center (UPMC) Eye Center, one of the top ophthalmology programs in the U.S., revealed a strikingly different picture. In the existing payment model at UPMC, insurance plans only approve payment for one test per day/per patient. Moreover, depending on the type of test and disease, insurance firms limit the number of reimbursable tests per year. For instance, when a physician orders three tests for a patient, the physician understands that only one test will be reimbursed by the insurance firm, and that the other two will not generate additional revenue. To further complicate the issue, it has been observed that different physicians can charge different service fees for the same or similar procedures, and this difference is especially salient across hospitals (Economist 2010).

Based on our interviews with physicians and patients at UPMC Eye Center, we identified three crucial factors behind patients' decisions to visit doctors' offices: out-of-pocket expense, waiting time, and service quality. First, in the U.S. healthcare market, the majority of patients are insured and pay less than the actual service charge. Second, long service queues influence patients' experiences to such an extent that patients desire monetary compensation for long waiting times (Alderman 2011), and waiting-time-tracking websites like `www.medwaittime.com` have emerged. Third, patients are concerned about the service quality, which is closely tied to the quantity of diagnostic tests, though the marginal return from ordering additional tests is diminishing (Mold et al. 2010).

We aim to examine the driving forces behind physicians' test-ordering behavior, and our model captures key financial, operational and clinical incentives that govern the strategic interactions between the physician and patients. While the physician strikes a balance between economic gains and diagnosis certainty, patients optimally trade off between waiting time, out-of-pocket expense, and service quality. We characterize the physician's optimal service decision and patients' queue-joining behavior, which we refer to as the *market equilibrium*, as opposed to the *social optimum* in which the social welfare is maximized. The measure of inefficiency in overtesting is the loss of social welfare with respect to the socially efficient administration of diagnostic tests. We focus attention on the following service settings that affect physicians' test-ordering patterns:

• **Insurance structure**. Health insurance distorts the cost of services to insured patients and has thus been documented as one reason for excessive utilization of health care services. While existing studies hold that lower out-of-pocket expenses lead to higher consumption levels, we refine this statement by showing that the copayment and the coinsurance rate affect physician's prescription decisions differently.

• **Misdiagnosis concerns**. Physicians bear the risk of misdiagnosis that can be attributed to either misinterpretation of results from diagnostic tests, or failure to order adequate diagnostic tests. We show that, both overtesting and undertesting are possible outcomes with the introduction of misdiagnosis concerns. The underlying intuition is that physicians' misdiagnosis concerns push up the socially efficient consumption level.

• **Patient heterogeneity**. Patients with comparable medical conditions can have different insurance coverage and waiting costs. We model the inherent patient heterogeneity and characterize the optimal service rate. We show that patient heterogeneity justifies ordering more tests under both the market equilibrium and the social optimum.

• **Diagnosis time**. Physicians need to spend time in reading and analyzing test results, which essentially shorten their effective service time. The tradeoff here is between maximizing service throughput and increasing service quality. We show that the impact of insurance structure carries over when considering the physician's diagnosis time.

• **Information asymmetry**. Physicians possess heterogenous diagnostic skill levels that are unobservable to patients. We model the physician-patient interaction under information asymmetry as a signaling game, and characterize the perfect Bayesian equilibrium that can be either a *costless* or *costly* separating equilibrium. Our analysis reveals that price transparency—as opposed to the dominating practice of opaque pricing—can encourage physicians to overtest, especially when technological advancements diminish differentiation among physicians.

1.1 Literature Review

Our research continues the theme of expert services literature for which Dulleck and Kerschbamer (2006) provide an extensive review. A recent paper related to ours is by Debo et al. (2008), who consider a monopolist expert offering a service for which consumers cannot verify its necessary service time even after purchase; the expert hence has the incentive to prolong the service duration. They characterize the expert's and customers' equilibrium behavior. While embedding asymmetric information, their model does not address the differences in service quality. Veeraraghavan and Debo (2009) consider consumers who cannot determine whether a low service rate or a high arrival rate contributes to a long queue. Consumers rely on their private information to make queue-joining

decisions. Anand et al. (2011) study a service provider's optimal tradeoff between service quality and speed in the presence of strategic customers. They show that one major driving force behind the service provider's decision is the customer intensity of the industry. Furthermore, they extend their analysis to the competition among multiple service providers and show that higher prices and service quality levels can emerge as more providers join in the competition. Our paper differs from Anand et al. (2011) in several ways. First, we consider the insurance coverage that distorts actual prices to insured patients, and emphasize the profound impact of insurance structure on the service consumption under various service environments. Second, one key aspect of our analysis is to compare the actual and the socially efficient service consumption levels; this comparison is closely tied to our research motivation but is not considered in Anand et al. Third, we consider the impact of asymmetric information on the physicians' test-ordering behavior. Kostami and Rajagopalan (2009) analyze the intertemporal tradeoff between speed and quality in a general service setting. Our model of physician type uncertainty in §3.5 is similar to theirs except that we allow strategic consumers (patients) and asymmetric information. Wang et al. (2010) develop a multi-server queueing model of a diagnostic service center that advises patients over phone about appropriate course of action. The service manager needs to strike a balance between accuracy of advice, callers' waiting time, and staffing costs. Our paper also addresses the tradeoff between accuracy of diagnosis and waiting time but focuses on the economic side of physicians' test-ordering behavior.

Another strand of literature contends that doctors, as service providers, can directly influence patients' service consumption decisions. Patients seek advice from doctors largely because they do not know the procedures and tests necessary to reach informed medical decisions. Such an intuitive argument leads to the fundamental assumption in health economics, namely the supplier-induced demand (SID) hypothesis (Evan 1974). While early SID models often view patients as perfectly informed but passively sovereign consumers, later studies begin to model patients as Bayesian decision-makers whose information-acquisition mechanism serves as a constraint on physicians' demand inducement. Three key features separate our paper from the SID literature. First, prior SID models in general assume that physicians can costlessly observe patients' private information. Second, while waiting time negatively affects patients' experience and reduces their access to healthcare, it has been regarded as a mechanism to control utilization and hence reduce the cost of *ex post* moral hazard (Gravelle and Siciliani 2008). The extant health economics literature, however, treats the waiting time as the healthcare provider's unilateral, self-concocted decision variable

rather than an output variable formed during the physician-patient interaction. Third, the fee-for-service payment model is generally assumed in SID models. Sorensen and Gyrtten (1999), for example, build their models on the premise that only contract physicians in Norway, whose incomes come exclusively from patient visits or laboratory tests, have the incentive to induce demand. Our paper, by considering the information acquisition costs, insurance coverage and waiting time, reveals physicians' incentives to overtest even when more services do not necessarily bring about additional revenue.

2. Modeling Physician-Patient Strategic Interaction

In this section, we develop a baseline model of the strategic encounters between a physician and a group of exogenous arriving patients under perfect information. We start by modeling the relationship between diagnostic tests and the service quality, followed by incorporating various factors affecting patients' and the physician's decision-making. Then we characterize the market equilibrium and the social optimum, and identify the condition under which the physician would overtest.

2.1 Diagnostic Tests and Service Quality

Modeling the relationship between diagnostic tests and service quality can be a daunting task, especially when different types of diagnostic tests are designated to produce different areas of diagnostic information. Our collaborative experience in the outpatient setting helps us simplify the modeling in two ways. On the one hand, although there exists a large pool of available diagnostic tests and numerous possible combinations of tests, physicians typically determine the combination of tests in accordance with standard guidelines that specify the priorities of various diagnostic tests. When physicians choose a larger number of tests, they are essentially choosing a wider set of diagnostic tests. In other words, there exist pre-determined sequences that relate the number of tests to the specific combination of tests. On the other hand, while it is hard to determine the total number of tests for a specific patient *ex ante*, physicians typically pre-allocate "appointment intervals" that specify the duration reserved for each patient; the chosen appointment interval includes the service time for consultation and diagnostic tests, although the actual service time is a random variable. To consider it in the framework of queueing theory, when physicians choose their appointment intervals that are reserved for both medical consultation and diagnostic tests, they are essentially determining an aggregate *service rate*, or how fast they serve patients. A slower service rate corresponds to more diagnostic tests.[1] These two observations lead to a simplified

[1] In practice, a physician time is divided into "service sessions" for consultation and diagnosis, which typically operate during certain time intervals. This means that only a proportion of the total available time could be used each process. We will address this issue in the next section.

model in which the physician uses the service rate as a decision variable. Once the service rate is determined, so is the number of tests, which determines the depth of diagnostic tests based on the existing guidelines.

Extant medical literature, for example, Kassirer (1989) and Connell and Koepsell (1985), supports the observation that, the longer service time physicians allocate for each patient visit (or the more diagnostic tests the physician orders), the more likely it will be to reach an accurate diagnosis. Although a slow service rate is often clinically beneficial to each individual patient, it comes at a negative externality of putting other patients in a long service queue.

Inspired by our collaborative study, we capture the relationship between diagnostic tests and service quality in the following way: we relate a lower service rate to more tests, or a higher service rate to fewer tests. Furthermore, we assume that the service rate can take values from a continuous set. The service quality (or the diagnostic certainty) is determined by the consultation session, as well as the physician's analysis of results from various diagnostic tests. The service quality from the consultation alone is denoted as Q_c. The service quality from both the consultation and diagnostic tests, given service rate μ, is defined as

$$Q(\mu) := Q_c + \alpha(\mu_c - \mu), \tag{1}$$

where μ_c, referred to as the baseline service rate, is the service rate at which the service quality is Q_c, that is, $Q(\mu_c) = Q_c$; α denotes the sensitivity of the service quality to service speed, and describes the rate in which the service quality improves when the service rate decreases. α can be viewed as a parameter measuring the physician's skill level.[2] We assume $\mu \leq \mu_c$ because it is a legal requirement that the physician cannot skip the consultation stage; $\mu = \mu_c$ corresponds to the case where the physician does not order any diagnostic tests. It follows from (1) that $Q(\mu)$ decreases in μ, meaning that a slower service rate leads to higher service quality. In addition, $Q(\mu)$ increases in α, which is aligned with the intuition that a more skillful physician can provide higher-quality service with the same amount of information. Figure 1 illustrates our modeling of the service process.

Empirically, μ_c can be estimated as the maximum number of patients that the physician serves in a given time unit, assuming that no additional diagnostic tests are prescribed. Although α cannot be directly observed, it can nonetheless be estimated in two ways: 1) α can be inferred

[2] This assumption is similar to Anand et al. (2011) but the interpretation for α is different: instead of viewing α as the "customer-intensity," we interpret the differences in α as physicians' differences in skill levels for the same type of job, that is, analyzing data from diagnostic tests.

from the data of the service queue since the pricing, insurance, arrival rate, and congestion data can be observed; 2) α can be approximated after the accuracy levels of different combinations of diagnostic tests are calibrated using logistic regression (Habbema et al. 2002). For simplicity of analysis, $Q(\mu)$ is assumed to be an affine function of μ; however, we show in Appendix B.3 that our major results extend to a general relationship between the service rate and the service quality.

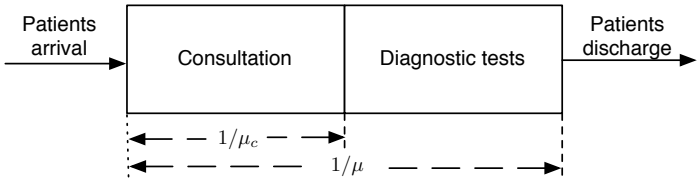

Figure 1 Modeling of the service process.

2.2 Patient Utility

Patients' utility from the service depends on three factors: service quality, waiting time, and *out-of-pocket* payment. Patients are insured and hence pay less than the nominal service charge. All patients are assumed to have the same insurance coverage with zero deductible, a copayment of π, and a coinsurance rate of β; the deductible is the accumulative amount that the patient needs to pay out-of-pocket before enjoying insurance coverage from the insurance firm; the copayment is the fixed per-visit charge that the patient must pay out of pocket; the coinsurance rate is the percentage of service fee, after accounting for the copayment, that the patient must pay out-of-pocket. We assume homogeneous patients but will extend our model to incorporate patient heterogeneity in insurance coverage in §3.3. The premium is viewed as a sunk cost and ignored. The deductible is also ignored to avoid the difficulty of defining the service fee below the deductible (cf. Newhouse 1978). Letting p denote the nominal service fee, the patient's out-of-pocket payment is hence $\pi+\beta(p-\pi)^+$, where $(p-\pi)^+$ is equivalent to $p-\pi$ as we focus solely on the interesting case where $p \geq \pi$.

Patients arrive at an exogenous rate Λ, which is referred to as the potential demand for the service.[3] Upon observing the physician's chosen service rate μ and service fee p, patients make queue-joining decisions by adopting the following mixed strategies: each patient joins the queue with probability $\rho(\mu,p)$, and balks and resorts to an outside option with probability $1-\rho(\mu,p)$.

[3] This assumption is also adopted in Anand et al. (2011). Our analysis is similar to theirs but we focus on the most realistic case that Λ is sufficiently large so that neither full coverage (i.e., $\rho(\mu,p)=1$) nor no coverage (i.e., $\rho(\mu,p)=0$) would occur.

Each patient's reservation utility is assumed to be zero without loss of generality. The induced arrival rate can be denoted as a function of μ and p such that $\lambda(\mu,p)=\rho(\mu,p)\cdot\Lambda$. The above setting of patients' decision-making is consistent with the literature on equilibrium behavior of customers and servers in queueing systems (Hassin and Haviv 2003).

The potential demand for the service is assumed to follow a Poisson process, a reasonable representation for arrival processes in healthcare applications (Green 2006); the induced arrival process resulting from patients' joint randomized decisions, therefore, also follows a Poisson process. For simplicity of analysis, we assume that all the service times are exponential, and model the service system as an $M/M/1$ queue; we show in Appendix B.2 that our major results carry over to a general service time distribution. To be consistent with *money price* models (e.g., Coffey 1983), we define each patient's waiting time as the sum of the queueing time and the service time; an alternative definition of waiting time is considered in Appendix B.1. The expected waiting time in the $M/M/1$ queue is given by

$$\mathbb{E}\left[W(\mu,\lambda(\mu,p))\right]=\frac{1}{\mu-\lambda(\mu,p)}. \tag{2}$$

Let ω denote patients' unit waiting cost. In practice, ω can be estimated as the value of lost productivity while waiting in the service queue (Phelps and Newhouse 1973; Coffey 1983). The market clearing condition, that is, $Q(\mu)-\omega\mathbb{E}\left[W(\mu,\lambda(\mu,p))\right]-\pi-\beta(p-\pi)^+=0$, together with (2), gives the induced arrival rate

$$\lambda(\mu,p)=\mu-\frac{\omega}{Q(\mu)-\pi-\beta(p-\pi)^+}.$$

Throughout the paper, we assume that different patient visits are independent of each other; in Appendix B.5, we consider the case where patients might be advised by the physician to make follow-up visits.

2.3 Physician Behavior

We model the physician as a price-setter such that "the physician is assumed to have some control over the price he can charge and still obtain business" (Pauly 1980). We assume that the physician charges a fixed service fee p regardless of the realization of the actual service time; in Appendix B.7, we show that our analysis extends to the case where the service fee depends on the actual service duration. The physician's decision consists of choosing the service rate μ and the service fee p to maximize the revenue rate, that is,

$$g(\mu,p)=p\cdot\lambda(\mu,p). \tag{3}$$

The technical labor costs associated with the diagnostic tests, usually considered to be fixed, are ignored here, but our results extend to the case where the technical labor costs are variable (cf. Appendix B.6).

In addition, we make the assumption that $Q_c < \alpha\mu_c + (1-\beta)\pi$ to ensure that, as implied by the next proposition, the trivial case $\mu^* \geq \mu_c$ will never occur. The assumption requires that the baseline service quality Q_c is lower than the sum of 1) $\alpha\mu_c = \lim_{\mu\to 0} Q(\mu) - Q_c$, the unattainable maximum service quality improvement, and 2) $(1-\beta)\pi$, each patient's net copayment since β of the copayment is covered by the insurance. Below we characterize the equilibrium.

PROPOSITION 1. *With symmetric information, there exists a unique equilibrium as follows.*

i) The physician chooses the service rate $\mu^* = \frac{Q_c+\alpha\mu_c-(1-\beta)\pi}{2\alpha}$, *and the service fee* $p^* = \frac{1}{\beta}\left[\frac{Q_c+\alpha\mu_c-(1-\beta)\pi}{2} - \sqrt{\alpha\omega}\right]$.

ii) The induced arrival rate is $\lambda(\mu^*, p^*) = \frac{Q_c+\alpha\mu_c-(1-\beta)\pi}{2\alpha} - \sqrt{\frac{\omega}{\alpha}}$.

iii) The average waiting time is $\mathbb{E}[W(\mu^*, \lambda(\mu^*, p^*))] = \sqrt{\frac{\alpha}{\omega}}$.

The above proposition immediately gives the following result:

COROLLARY 1. *i) The physician's optimal service rate* μ^* *decreases in the copayment* π.

ii) The physician's optimal service rate μ^* *increases in the coinsurance rate* β.

iii) The physician's optimal service fee p^* *decreases in both the copayment* π *and the coinsurance rate* β.

The extant literature often suggests that increasing the patients' out-of-pocket expenses leads to reduced consumption of medical resources. The above corollary, by contrast, reveals that the copayment and the coinsurance rate drive the consumption of diagnostic tests toward reverse directions. In particular, the number of tests increases in the copayment π but decreases in the coinsurance rate β. To understand the underlying intuition for this result, we examine each patient's out-of-pocket expense $\pi + \beta(p^* - \pi) = [Q_c + \alpha\mu_c + (1-\beta)\pi]/2 - \sqrt{\alpha\omega}$, which increases in π but decreases in β. As the copayment goes up, the physician needs to cut the service fee to ease the patients' monetary burden. Nevertheless, each patient's out-of-pocket expense still goes up because cutting the service fee by one dollar only reduces each patient's out-of-pocket expenses by $\beta < 1$ dollar, calling for more tests to be ordered to match the patients' increased monetary burden. With a higher coinsurance rate, however, the physician will charge a lower service fee, which leads to a reduced out-of-pocket expense for each patient, and justifies fewer tests ordered by the physician.

To the best of our knowledge, this is the first analytical finding about the impact of per-visit copayment on physicians' test-ordering behavior. There exist supporting empirical evidences for the

result. Newhouse (1978) cites empirical inpatient studies to show that increased daily copayment leads to reduced patient stay but increased intensity of care per case. Jung (1998) shows under an outpatient setting that increasing the per-visit copayment significantly reduces the number of office visits but increases the intensity of medical resource consumption per episode.

2.4 Social Optimum and Overtesting Condition

The benchmark that we will consider to measure overtesting is with respect to the social optimum in which the social planner determines the admission policy and the service rate to maximize the social welfare. Each physician-patient interaction generates the social surplus that is equal to the service quality, less patients' disutility from waiting. The expected social welfare rate is formulated as follows:

$$U(\mu,\lambda)=\lambda\cdot\{Q(\mu)-\omega\mathbb{E}[W(\mu,\lambda)]\}. \tag{4}$$

The next proposition gives the socially efficient service rate and arrival rate, denoted by μ^{SE} and λ^{SE}, respectively.

PROPOSITION 2. *In the social optimum,*

i) the optimal service rate is $\mu^{SE}=\frac{Q_c+\alpha\mu_c}{2\alpha}$;

ii) the optimal arrival rate is $\lambda^{SE}=\frac{Q_c+\alpha\mu_c}{2\alpha}-\sqrt{\frac{\omega}{\alpha}}$;

iii) the expected waiting time is $\mathbb{E}\left[W(\mu^{SE},\lambda^{SE})\right]=\sqrt{\frac{\alpha}{\omega}}$.

The following corollary compares the social optimum with the market equilibrium.

COROLLARY 2. *i) In the market equilibrium, the physician orders no fewer tests than in the social optimum, that is,* $\mu^*\le\mu^{SE}$.

ii) In the market equilibrium, the arrival rate is always no greater than in the social optimum, that is, $\lambda(\mu^*,p^*)\le\lambda^{SE}$.

iii) The average waiting time in the social optimum is the same as in the market equilibrium, that is, $\mathbb{E}[W(\mu^{SE},\lambda^{SE})]=\mathbb{E}[W(\mu^*,\lambda^*(\mu^*,p^*))]=\sqrt{\alpha/\omega}$.

In the market equilibrium, the physician tends to overtest due to the price distortions introduced by insurance coverage. This result is aligned with Feldstein's (1973) empirical finding that raising the coinsurance rate leads to increased social welfare. In fact, when $\pi=0,\beta=1$, patients are responsible for the full payment, and the physician will set the service rate at the socially efficient level.

We conclude the section with a corollary about the social welfare gap between the market equilibrium and the social optimum.

COROLLARY 3. *The social welfare gap is convex decreasing in the coinsurance rate β, and convex increasing in the copayment π.*

As the copayment increases, the physician tends to order more diagnostic tests for each patient. In the meantime, the equilibrium arrival rate decreases. Combining the decreased arrival rate with the increased number of tests per patient visit, we recognize the phenomenon that more and more resources are consumed by fewer and fewer individuals at any given time, leading the social welfare gap to widen at a faster pace. This phenomenon explains why the social welfare gap is convex increasing in the copayment π. As the coinsurance rate increases, the physician's test-ordering pattern converges to the socially efficient one, which is social-welfare-improving. Moreover, a higher coinsurance rate effectively brings the equilibrium arrival rate closer to the socially efficient arrival rate.

3. Impact of Service Environments

This section considers several service environments, including the reimbursement ceiling, physicians' misdiagnosis concerns, patient heterogeneity, physicians' time spent in reading and analyzing test results, and patients' *ex ante* uncertainty about physician type. We are interested in analyzing the effect of various service environments on the physician-patient interaction, and hence physicians' test-ordering behavior.

3.1 Reimbursement Ceiling

Observing that insurance coverage distorts the demand curve for diagnostic services, one natural countermeasure is to introduce a reimbursement ceiling, that is, the maximum reimbursable amount for each service session. We use $p_{\max}$ to denote the reimbursement ceiling. The physician's decision consists of choosing the service rate μ and the service fee p that maximize her revenue rate. Defining $q_{\max} = \pi + \beta(p_{\max} - \pi)$, the equilibrium is characterized in the proposition that follows.

PROPOSITION 3. *Depending on the size of $p_{\max}$, two possible equilibrium outcomes can arise:*

i) If $p_{\max} > \frac{Q_c+\alpha\mu_c-2\sqrt{\omega\alpha}-(1-\beta)\pi}{2\beta}$, then the equilibrium is the same as in Proposition 1.

ii) If $p_{\max} \le \frac{Q_c+\alpha\mu_c-2\sqrt{\omega\alpha}-(1-\beta)\pi}{2\beta}$. In equilibrium,

a) the physician chooses the service fee $p^* = p_{\max}$ and the service rate $\mu^* = \frac{Q_c+\alpha\mu_c-q_{\max}}{\alpha} - \sqrt{\frac{\omega}{\alpha}}$;

b) the induced arrival rate $\lambda(\mu^*, p^*) = \frac{Q_c+\alpha\mu_c-q_{\max}}{\alpha} - 2\sqrt{\frac{\omega}{\alpha}}$;

c) the average waiting time $\mathbb{E}[W(\mu^*, \lambda(\mu^*, p^*))] = \sqrt{\frac{\alpha}{\omega}}$.

The following corollary illustrates how physicians' test-ordering behavior varies with different copayments and coinsurance rates in the presence of the reimbursement ceiling.

COROLLARY 4. *i) The physician's optimal service rate μ^* decreases in the copayment π.*

ii) The physician's optimal service rate μ^* increases in the coinsurance rate β if the reimbursement ceiling $p_{\max} > [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} - (1-\beta)\pi]/(2\beta)$; the physician's optimal service rate μ^* decreases in the coinsurance rate β otherwise.

The intuitions behind Corollary 4 are three-fold. First, *ceteris paribus*, when the copayment increases, the physician compensates patients' utility loss by ordering more tests. Second, when the reimbursement ceiling $p_{\max}$ is high enough, greater insurance coverage encourages overtesting, as patients are less sensitive to the service fee, i.e., the physician responds to a decrease in the coinsurance rate β by ordering more tests. Third, when the insurance firm sets a low reimbursement ceiling $p_{\max}$, the physician will set the service fee at exactly $p_{\max}$. A lower coinsurance rate β, similar to a lower copayment π, reduces patients' fixed out-of-pocket payment, and the physician can order fewer tests without sacrificing patients' net surplus.

We briefly discuss the social welfare gap based on Corollary 4. As in the baseline model, the social welfare gap is convex increasing in the copayment π since both the service rate μ^* and the equilibrium arrival rate $\lambda(\mu^*, p^*)$ decrease in π. The social welfare gap is convex decreasing in β when the reimbursement ceiling $p_{\max}$ is high, as in the baseline model. With a low reimbursement ceiling, however, both the arrival rate and the service rate decrease in β, meaning that the social welfare gap is convex increasing in β.

The following corollary compares the market equilibrium with the social optimum.

COROLLARY 5. *i) If the reimbursement ceiling $p_{\max} > [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} - (1-\beta)\pi]/(2\beta)$, then the physician always orders more tests than the socially efficient level, that is, $\mu^* < \mu^{SE}$.*

ii) If the reimbursement ceiling $p_{\max} \le [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} - (1-\beta)\pi]/(2\beta)$, then the physician can order more or fewer tests than the socially efficient level, that is, both $\mu^* \le \mu^{SE}$ and $\mu^* > \mu^{SE}$ are possible.

The above corollary shows the conditions under which overtesting occurs. When the reimbursement ceiling is sufficiently high, the physician always overtests. With a low reimbursement ceiling, however, the physician can either overtest or undertest, depending on whether each patient' out-of-pocket expense $q_{\max}$ is over $(Q_c + \alpha\mu_c - 2\sqrt{\alpha\omega})/2$. The effect comes from the intuition that a higher net payment is compensated by more diagnostic tests, and vice versa.

Corollary 5 also helps uncover the puzzle that motivates our research. Recall from §1 that, overtesting occurs even under the *capitation payment* scenario, that is, physician receives the same income per patient visit regardless of the number of tests ordered. Consider a setting in which the

physician's compensation per patient visit is fixed at $\bar{p}$. The service rate becomes the physician's sole decision. This problem is equivalent to the case where the reimbursement ceiling is set low enough, and the physician always receives the upper bound of the service fee. In equilibrium, the physician chooses a service rate of $\mu^* = [Q_c + \alpha\mu_c - \pi - \beta(\bar{p} - \pi)]/\alpha - \sqrt{\omega/\alpha}$, which can be either higher or lower than the socially efficient service rate μ^{SE}. In other words, overtesting is still possible even under a capitation payment system, and is more likely to occur under a low coinsurance rate or a high copayment.

3.2 Misdiagnosis Concerns

The physician bears the risk of misdiagnosis. In some cases, the physician is subject to a penalty if there exists substantial proof that a patient's condition worsens in the face of inaction because the physician fails to interpret the testing results accurately. In some other cases, an inadequate amount of tests can indicate that a normal patient is abnormal, exposing patients to unnecessary tests and treatments. Prior medical literature validates the significance of misdiagnosis concerns in their scope and impacts. Studdert et al. (2006) find that 37% of malpractice claims do not involve any *real* medical errors but account for 13–16% of the system's total costs. In a study aiming at revealing physicians' perceived risk of misdiagnosis, Carrier et al. (2010) confirm high malpractice concerns among physicians in all levels even when malpractice risks are sufficiently low by objective measures. They also find that such concerns are not eased by common tort reforms. Baicker et al. (2007) show that increased malpractice risk drives higher consumption levels of healthcare services, especially in discretionary services.

We model the physician's misdiagnosis concerns as a simple misdiagnosis cost function of the service rate: $\theta(\mu) := d \cdot \mu$, where d is a constant denoting the marginal misdiagnosis cost in the service rate μ and can be empirically estimated as a latent variable (Skrondal and Rabe-Hesketh 2004) that relates the physician's service rate to her malpractice concerns; in Appendix B.4, we extend $\theta(\mu)$ to a general function form. The misdiagnosis cost is increasing in μ, aligning with the observation that fewer diagnostic tests make the physician more concerned about reaching inaccurate diagnosis. When μ is very small, indicating that the physician orders a sufficiently large number of tests, the misdiagnosis cost approaches zero.

The physician's decision consists of choosing the service rate $\mu \in (0, \mu_c)$ and the service fee p to maximize the utility rate $g^\theta(\mu, p) = [p - \theta(\mu)] \cdot \lambda(\mu, p)$. We characterize the equilibrium in the following proposition:

PROPOSITION 4. *In the case with misdiagnosis concerns,*

i) the physician chooses the service rate $\mu^* = \frac{Q_c+\alpha\mu_c-(1-\beta)\pi}{2(\alpha+\beta d)}$, *and the service fee* $p^* = \frac{1}{\beta}\left[\frac{(\alpha+2\beta d)[Q_c+\alpha\mu_c-(1-\beta)\pi]}{2(\alpha+\beta d)} - \sqrt{\omega(\alpha+\beta d)}\right]$;

ii) the induced arrival rate is $\lambda(\mu^*, p^*) = \frac{Q_c+\alpha\mu_c-(1-\beta)\pi}{2(\alpha+\beta d)} - \sqrt{\frac{\omega}{\alpha+\beta d}}$;

iii) the average waiting time is $\sqrt{\frac{\alpha+\beta d}{\omega}}$.

The corollary below is immediate from Part i) of Proposition 4.

COROLLARY 6. *i) With misdiagnosis concerns, the physician's optimal service rate* μ^* *is decreasing in the copayment* π.

ii) If the copayment π *is lower than* $(Q_c + \alpha\mu_c)/(1 + \alpha/d)$, *then the physician's optimal service rate* μ^* *decreases in the coinsurance rate* β; *otherwise, the physician's optimal service rate* μ^* *increases in the coinsurance rate* β.

An increase in the fixed per visit charge leads to a higher expectation in service quality, justifying more tests ordered by the physician. An increase in the coinsurance rate β, however, can lead to either an increase or decrease in the optimal service rate μ^* depending on the copayment π. The underlying intuition is that when the copayment π is low, the variable part of the out-of-pocket expense accounts for a larger role due to a high residual payment $(p - \pi)$; an increase in the coinsurance rate, therefore, should be compensated by increasing the service quality. When the copayment is high, however, the variable part of the out-of-pocket expense gains less importance, a higher coinsurance rate would make it imperative for the physician to prescribe fewer tests and charge a lower service fee to ease patients' economic burden.[4]

The threshold $(Q_c + \alpha\mu_c)/(1 + \alpha/d)$ in Part ii) of Corollary 6 has an intuitive interpretation. The numerator $Q_c + \alpha\mu_c = \lim_{\mu\to 0} Q(\mu)$ is the unattainable upper bound of the service quality, and the denominator $1 + \alpha/d$ is the relative value of the physician's skill level α to her unit misdiagnosis cost d. Consider two extreme cases: (1) As d approaches zero, the threshold approaches zero. In this case, the copayment π is always above the threshold, and the optimal service rate μ^*, consistent with Part (i) of Corollary 1, always increases in the coinsurance rate β; (2) As d approaches infinity, the threshold approaches $Q_c + \alpha\mu_c$, which is higher than any practical copayment π. In this case, the optimal service rate μ^* always decreases in the coinsurance rate β.

We use Corollary 6 to illustrate the impact of the insurance structure on the social welfare gap. The social welfare gap is convex increasing in π because a higher copayment level limits the access to the service but encourages the physician to overtest for individual visits. When π is low, the

[4] The authors wish to thank one of the anonymous referees for correcting this corollary and providing insights into understanding its underlying intuition.

social welfare gap is convex increasing in the coinsurance rate β since both the service rate and the arrival rate decrease in β. When π is high, however, the service rate decreases in β but the arrival rate increases in β, and the social welfare gap can be either convex or concave in β.

Next, we derive the conditions under which the physician would overtest. The social planner aims to maximize the social welfare rate that can be represented as $U^\theta(\mu,\lambda) = \lambda \cdot \{Q(\mu) - \theta(\mu) - \omega\mathbb{E}[W(\mu,\lambda)]\}$. The next proposition characterizes the social optimum.

PROPOSITION 5. *With misdiagnosis concerns, in the social optimum,*

i) the optimal service rate is $\mu^{SE} = \frac{Q_c+\alpha\mu_c}{2(\alpha+d)}$;

ii) the optimal arrival rate is $\lambda^{SE} = \frac{Q_c+\alpha\mu_c}{2(\alpha+d)} - \sqrt{\frac{\omega}{\alpha+d}}$;

iii) the expected waiting time is $\mathbb{E}W(\mu^{SE},\lambda^{SE}) = \sqrt{\frac{\alpha+d}{\omega}}$.

The following corollary is immediate from Propositions 4–5.

COROLLARY 7. *If the copayment π is higher than $(Q_c + \alpha\mu_c)/(1+\alpha/d)$, then the physician orders more tests than the socially efficient level, that is, $\mu^* < \mu^{SE}$; otherwise, the physician orders fewer tests than the socially efficient level.*

Corollary 7 provides counterintuitive insight on the impact of misdiagnosis concerns: when physicians are concerned by potential inaccurate medical judgment, they can order either more or fewer tests than the socially efficient level (the latter case is referred to as "undertesting"). It is especially surprising if we recall from Corollary 2 that the physician always overtests when they do not have misdiagnosis concerns. To understand this result, we need to examine the socially efficient level as characterized in Proposition 5, which shows that misdiagnosis concerns lead to a higher socially efficient consumption level. The insurance coverage, on the other hand, enables patients to pay less than the actual service fee. Specifically, when the copayment is lower than $(Q_c+\alpha\mu_c)/(1+\alpha/d)$, the physician can satisfy patients by ordering fewer tests than the socially efficient level. When the copayment amount is above $(Q_c+\alpha\mu_c)/(1+\alpha/d)$, however, the physician's demand-inducing efforts are supplemented by the insurance coverage. Furthermore, given Q_c and μ_c, the threshold decreases in the ratio between α and d. Consider, as a special case, that the physician's misdiagnosis concern is sufficiently low (d is small), the threshold will be close to zero, meaning that the physician will invariably overtest, which is consistent with Corollary 2.

COROLLARY 8. *The average waiting time in the social optimum is longer than in the market equilibrium, that is,* $\mathbb{E}[W(\mu^{SE},\lambda^{SE})] > \mathbb{E}[W(\mu^*,\lambda^*(\mu^*,p^*))]$.

Corollary 8 may initially seem surprising in that, even when the physician orders more tests than the socially efficient level, patients still experience a shorter expected waiting time. The underlying intuition is as follows. We first recognize that one way to implement the social optimum is to charge each patient a service fee coinciding with the patient's externality by joining the queue

$$p^{SE} = Q(\mu^{SE}) - \omega \mathbb{E}W(\mu^{SE}, \lambda^{SE}) = \frac{(\alpha+2d)(Q_c+\alpha\mu_c)}{2(\alpha+d)} - \sqrt{\omega(\alpha+d)}. \tag{5}$$

Under the market equilibrium, however, each patient's out-of-pocket expense is

$$\pi + \beta(p^* - \pi) = \frac{(\alpha+2\beta d)(Q_c+\alpha\mu_c) + \alpha\pi(1-\beta)}{2(\alpha+\beta d)} - \sqrt{\omega(\alpha+\beta d)}. \tag{6}$$

Recall from Corollary 7 that, when $\pi > (Q_c + \alpha\mu_c)/(1+\alpha/d)$, the physician overtests. In the meanwhile, comparing (5) and (6) gives that $\pi + \beta(p^* - \pi) > p^{SE}$, meaning that each patient is subject to a high out-pocket expense, which essentially induces a low arrival rate. Consequently, the gap between the induced arrival rate and the service rate is higher than under the social optimum, leading to a lower expected waiting time. This phenomenon has been observed empirically in Hong Kong's healthcare system, in which the average waiting times in public and private hospitals are significantly different: the average waiting time for a physician in public hospitals is 74.7 days, while that for a private physician is 24.3 days (Harvard Team 1999).

3.3 Heterogeneous Patients

We assume in the baseline model that patients are homogenous, which provides a benchmark for understanding physicians' test-ordering behavior. Now we consider the case where patients can have different insurance coverage and valuations for time.

There are two patient groups, those who have "good" insurance plans (denoted by the subscript g), and those who have "bad" insurance plans (denoted by the subscript b). The two groups of patients are referred to as type g and type b patients, respectively. Type i patients' health plans are specified by a copayment π_i and a coinsurance rate β_i for $i = g, b$. We assume that $\pi_g < \pi_b$ and $\beta_g < \beta_b$ so that patients with good insurance have lower out-of-pocket expenses given the same nominal service fee. The potential arrival rates of the two groups are Λ_g and Λ_b, both of which are assumed to be sufficiently large such that full coverage is not a possible outcome. Within each patient group, patients differ in their sensitivity to delay, that is, their waiting costs. Furthermore, we assume that each type g patient has a waiting cost, denoted by ω_g, that is uniformly distributed in $[\underline{\omega}_g, \bar{\omega}_g]$, while each type b patient has a waiting cost ω_b that is uniformly distributed in $[\underline{\omega}_b, \bar{\omega}_b]$. We focus our attention on the case that $\bar{\omega}_g > \bar{\omega}_b$ and $\underline{\omega}_g > \underline{\omega}_b$, to be consistent with the observation that, more often than not, patients with good health plans have higher time prices.

We assume that the physician cannot discriminate patients by adopting varying appointment intervals (and hence service rates) or service fees based on patients' money price (dictated by insurance coverage) or time price. This reflects the consideration that the physician does not possess (or does not take into consideration) each patient's specific insurance information and thus chooses to base her test-ordering decisions on the profile of the "average" patient who seeks service. When the physician chooses the service rate μ and the service fee p, there exist critical levels ω_i^* such that only type i patients with $\omega_i \le \omega_i^*$ for $i = g, b$ join the queue; the other patients opt out of the queue and seek for outside options. Assuming zero reservation utilities for both types of patients, ω_g^* and ω_b^* are determined by solving the following two equations in which λ_g and λ_b correspond to the equilibrium arrival rates of the two types:

$$Q(\mu) - P_i(p) - \frac{\omega_i^*}{\mu - \lambda_g - \lambda_b} = 0 \text{ for } i = g, b, \tag{7}$$

$$\lambda_i = \left(\frac{\omega_i^* - \underline{\omega}_i}{\Delta\omega_i} \right) \Lambda_i \text{ for } i = g, b. \tag{8}$$

Jointly solving (7)–(8) gives the equilibrium arrival rates when the physician chooses μ and p:

$$\lambda_i(\mu, p) = \frac{\Lambda_i \left\{ \Delta\omega_j \left[\mu P_j(p) - \mu Q(\mu) + \underline{\omega}_i \right] + \Lambda_j \left[Q(\mu) \left(\bar{\omega}_i - \underline{\omega}_i \right) + P_j(p)\underline{\omega}_j - P_i(p)\underline{\omega}_i \right] \right\}}{P_j(p)\Delta\omega_j\Lambda_i + P_i(p)\Delta\omega_i\Lambda_j - Q(\mu)\left(\Delta\omega_j\Lambda_i + \Delta\omega_i\Lambda_j \right) - \Delta\omega_i\Delta\omega_j}, \tag{9}$$

where $P_i(p) := \pi_i + \beta_i(p - \pi_i)$ and $\Delta\omega_i := \bar{\omega}_i - \underline{\omega}_i$ for $i = g, b$, $j \neq i$.

After substituting (9) into the physician's objective function $\pi_p(\mu, p) = p \cdot [\lambda_b(\mu, p) + \lambda_g(\mu, p)]$, we show that $\pi_p(\mu, p)$ is concave in p, and we can subsequently verify that $\pi_p(\mu, p^*(\mu))$ is unimodal in μ. To facilitate our analysis, we define the quantity $\mu_i^* := [Q_c + \alpha\mu_c - (1 - \beta_i)\pi_i]/(2\alpha)$ for $i = g, b$, which corresponds to—recall from Proposition 1—the optimal service rate when there exist only type i patients with homogeneous waiting costs. Using similar procedures as in the proof of Proposition 1, we obtain the optimal service rate for the heterogeneous-patients system as follows:

$$\mu^* = \frac{\rho_g\mu_g^* + \rho_b\mu_b^*}{\rho_g + \rho_b} - \sqrt{\left(\frac{\rho_g\mu_g^* + \rho_b\mu_b^*}{\rho_g + \rho_b} \right)^2 - \frac{\rho_g\underline{\omega}_b + \rho_b\underline{\omega}_g}{\alpha\left(\rho_g + \rho_b \right)}}, \tag{10}$$

where $\rho_b := \Delta\omega_b/\Lambda_b$, and $\rho_g := \Delta\omega_g/\Lambda_g$.

The following corollary is straightforward from (10) and means that introducing patient heterogeneity in insurance coverage and waiting costs does not induce the physician to order fewer tests compared to the homogeneous case.

COROLLARY 9. $\mu^* < \max\{\mu_b^*, \mu_g^*\}$.

Social Optimum. Next, we analyze the social optimum under patient heterogeneity. Since both Λ_b and Λ_g are sufficiently large, the optimal admission control policy is dictated by a parameter $\hat{\omega}$ such that only patients with waiting costs lower than $\hat{\omega}$ are admitted into the queue; insurance coverage no longer plays a role. Depending on the relative quantity of $\underline{\omega}_b$, $\underline{\omega}_g$, and $\hat{\omega}$, two possible cases can arise:

Case 1. $\underline{\omega}_g > \hat{\omega}$, that is, no type g patients' waiting costs are lower than the threshold. In this case, only type b patients with waiting costs lower than $\hat{\omega}$ are admitted into the queue. The arrival rate is $\lambda_b = \frac{\hat{\omega}-\underline{\omega}_b}{\Delta\omega_b} \cdot \Lambda_b$. The social planner's problem is to choose the service rate μ and the admission control parameter $\hat{\omega}$ to maximize the social welfare:

$$SW(\mu, \hat{\omega}) = \lambda_b \cdot \left[Q(\mu) - \frac{\underline{\omega}_b + \hat{\omega}}{2} \cdot \frac{1}{\mu - \lambda_b} \right], \tag{11}$$

where

$$\lambda_b = \frac{\hat{\omega} - \underline{\omega}_b}{\Delta\omega_b} \cdot \Lambda_b. \tag{12}$$

In this case, we can show that $SW(\mu, \hat{\omega})$ is concave in μ, and the optimal service rate $\mu^*(\hat{\omega}) = \lambda_b + \sqrt{(\underline{\omega}_b + \hat{\omega})/(2\alpha)}$. Substituting this intermediate result into (11), we can write the social welfare as a function of λ_b and $\hat{\omega}$:

$$SW(\lambda_b, \hat{\omega}) = \lambda_b \cdot \left[Q_c + \alpha\mu_c - \alpha\lambda_b - \sqrt{2\alpha(\underline{\omega}_b + \hat{\omega})} \right]. \tag{13}$$

Note that (13) is in essence a function of a single variable λ_b, since (12) gives $\hat{\omega} = \Delta\omega_b/\Lambda_b \cdot \lambda_b + \underline{\omega}_b$. Solving the first-order condition gives

$$Q_c + \alpha\mu_c - \sqrt{2\alpha(\underline{\omega}_b + \hat{\omega})} - 2\alpha\lambda_b - \frac{\Delta\omega_b}{\Lambda_b} \cdot \sqrt{\frac{\alpha}{2(\underline{\omega}_b + \hat{\omega})}} = 0, \tag{14}$$

Since

$$0 < \frac{\Delta\omega_b}{\Lambda_b} \cdot \sqrt{\frac{\alpha}{2(\underline{\omega}_b + \hat{\omega})}} < \frac{\Delta\omega_b}{2\Lambda_b} \cdot \sqrt{\frac{\alpha}{\underline{\omega}_b}},$$

we see from (14) that

$$\frac{Q_c + \alpha\mu_c}{2\alpha} - \frac{\Delta\omega_b}{4\Lambda_b\sqrt{\alpha\underline{\omega}_b}} - \sqrt{\frac{\underline{\omega}_b + \hat{\omega}}{2\alpha}} \le \lambda_b^* \le \frac{Q_c + \alpha\mu_c}{2\alpha} - \sqrt{\frac{\underline{\omega}_b + \hat{\omega}}{2\alpha}},$$

and hence

$$\frac{Q_c + \alpha\mu_c}{2\alpha} - \frac{\Delta\omega_b}{4\Lambda_b\sqrt{\alpha\underline{\omega}_b}} \le \mu^* = \lambda_b^* + \sqrt{\frac{\underline{\omega}_b + \hat{\omega}}{2\alpha}} \le \frac{Q_c + \alpha\mu_c}{2\alpha},$$

which shows that the existence of patient heterogeneity essentially reduces the socially efficient service rate. The underlying explanation is that, as the arrival rate increases, the average waiting

cost also increases. The social planner, therefore, admits fewer patients at any given time, and provides slower service for each patient accordingly.

Case 2. $\underline{\omega}_g \leq \hat{\omega}$, that is, some type g patients' waiting costs are lower than the threshold. In this case, both types of patients with waiting costs lower than $\hat{\omega}$ are admitted into the queue. The choice of the admission control parameter $\hat{\omega}$ leads to arrival rates of $\lambda_i^{SE} = (\hat{\omega} - \underline{\omega}_i)/\Delta\omega_i \cdot \Lambda_i, i = g, b$. The social planner chooses the service rate μ and the admission control parameter $\hat{\omega}$ to maximize the social welfare:

$$SW(\mu, \hat{\omega}) = \sum_{i=g,b} \lambda_i \cdot \left[Q(\mu) - \frac{\underline{\omega}_i + \hat{\omega}}{2} \cdot \frac{1}{\mu - \lambda_g - \lambda_b} \right],$$

where $\lambda_i = \Lambda_i \left(\hat{\omega} - \underline{\omega}_i\right)/\Delta\omega_i$ for $i = g, b$. We see from the above equation that patient heterogeneity, again, makes it more desirable to admit fewer patients at any given time because of the increased average waiting cost as a result of increased access. We use μ_h^{SE} to denote the socially efficient service rate in the heterogeneous system. The socially efficient service rate in a homogeneous system μ^{SE}, recall from Proposition 2, is independent of the waiting cost. The following corollary summarizes the effect of patient heterogeneity on the socially efficient service rate:

Corollary 10. *The socially efficient service rate in the heterogeneous system is always lower than in the homogeneous system, i.e.,* $\mu_h^{SE} < \mu^{SE}$.

We now compare the market equilibrium with the social optimum. From the optimality conditions for the market equilibrium, we derive that $\omega_b^* > \omega_g^*$, that is, the marginal type g patient has a higher delay cost rate than the marginal type b patient. Moreover, when translated into the equilibrium arrival rates, this implies that $\lambda_g/\lambda_b > \Lambda_g/\Lambda_b$. That is, the physician distorts the fraction of type g patients that she sees compared to the population average. Moreover, notice that when $\Delta\omega_g \geq \Delta\omega_b$, that is, the type g patients' waiting costs are also more variable, we have $\omega_g^* > \omega_b^* + \left(\underline{\omega}_g - \underline{\omega}_b\right)$. In other words, not only the marginal type g patient has a higher delay cost rate than the marginal type b patient, the physician also finds it optimal to see a disproportionate fraction of type g patients compared to the population average since $\lambda_g/\lambda_b > \Lambda_g/\Lambda_b$. In contrast, in the social optimum, the waiting cost rate of the marginal patients for both types is equal to the common threshold $\hat{\omega}$. Therefore, one might expect the average waiting time in the market equilibrium to be lower than in the social optimum since the additional amount that can be charged to type g patients as a result of the drop in waiting times is higher than that of type b patients: type g patients are not only more delay-sensitive, they can also absorb a higher price increase because of their better insurance coverage. While in both the market equilibrium and the social optimum the waiting cost serves as

an incentive to increase the service rate, in the market equilibrium this decrease in the waiting cost is experienced mainly by the type g patients through the increase in fees collected per unit time. This is consistent with the general view from welfare economics that market equilibrium leads to under-utilization of a system than is optimal from the social planner's point of view. We expect a similar phenomenon to hold in the present context.

We close this section by briefly discussing the impact of patient heterogeneity on the social welfare gap. The social planner's objective is to maximize the social welfare, and therefore does not place any weight on the individual patient's insurance coverage when determining the admission policy and the service rate. The physician, however, has the incentive to choose the service rate and the service fee to cherry-pick a mix of patients who are less price-sensitive. The difference is clearly reflected in the fact that there exist two cut-off waiting costs, namely, $\omega_i^*, i = g, b$, under the market equilibrium, but only a single cut-off waiting cost $\hat{\omega}$ in the social optimum. The higher the difference between ω_g^* and ω_b^*, the wider the social welfare gap extends between the market equilibrium and the social optimum.

3.4 Diagnosis Time

In this section, we consider the physician's diagnosis time, that is, the time spent by the physician in reading and analyzing the results from diagnostic tests, which might not be perfectly substituted by ordering more tests. However, the physician's time dedicated to reading and analyzing results of tests alters her effective service rate due to her absence for consultation process. Incorporating the diagnosis time provides new managerial insights into understanding the physician's test-ordering behavior.

Our model setup resembles the practical setting that we observed during our collaborative study: After each patient finishes the imaging tests, all the test results will be electronically transmitted to a terminal that the physician in charge can access at a later point in time (usually on the same day). In this service environment, two service rates capture the physician's behavior: (1) the service rate for consultation and diagnostic tests, denoted by μ, and (2) the service rate for reading and analysis of the data from diagnostic tests, denoted by μ_r. Once μ and μ_r are determined, the physician effectively breaks down her time into two parts: service period and diagnosis period. The physician provides service to patients during the former but not the latter. To be more specific, the physician is available for outpatient service only during $\mu_r/(\mu+\mu_r)$ of the time. Note that the ratio $\mu_r/(\mu+\mu_r)$ is strictly increasing in μ_r, meaning that the physician would have the incentive to set $\mu_r = \infty$ if spending extra time reading and analyzing tests did not generate any additional service value. We write $\rho := \mu_r/\mu$ (referred to as "processing rate"), and rewrite this ratio as $\rho/(\rho+1)$.

We also let μ_{rc} denote the maximum service rate for reading/analyzing the test results such that $Q_r(\mu_c, \mu_{rc}) = Q_c$.

We redefine the service quality as

$$Q_r(\mu, \mu_r) = Q_c + \alpha(\mu_c - \mu) + \alpha'(\mu_{rc} - \mu_r),$$

where α is the marginal quality improvement resulting from a decrease in the outpatient service rate and α' is the marginal quality improvement resulting from a decrease in the rate of reading/analyzing test results. In such a setting, there are two ways that the physician can improve the service quality: by ordering additional diagnostic tests (i.e., decreasing μ), and by spending more time reading/analyzing test results (i.e., decreasing μ_r).

As in the baseline model, patients use a randomized strategy to form the aggregate arrival rate. Specifically, patients use the same strategy as in the baseline model to form the aggregate arrival rate during the physician's service period, which account for $\rho/(1+\rho)$ of the total time; patients do not form a queue for time slots coinciding with the physician's diagnosis period. Since the reading/review period does not directly affect patients' waiting time, the expression for the equilibrium arrival rate during operating hours is exactly the same as in the baseline model except that the service quality is now given by $Q_r(\mu, \mu_r)$ instead of $Q(\mu)$. The physician's problem then becomes one of choosing μ and p to maximize the revenue rate represented by

$$\pi_p(\mu, \mu_r, p) = p \cdot \frac{\rho}{1+\rho} \cdot \left[\mu - \frac{\omega}{Q_r(\mu, \mu_r) - \pi - \beta(p-\pi)}\right].$$

The following proposition characterizes the equilibrium when the processing rate ρ is exogenous; later, we will discuss the case wherein the physician endogenously chooses the processing rate ρ.

PROPOSITION 6. *When considering the physician's diagnosis time,*

i) the physician chooses the service rate $\mu^* = [Q_c + \alpha\mu_c + \alpha'\mu_{rc} - (1-\beta)\pi]/[2(\alpha + \rho\alpha')]$*, and the service fee* $p^* = \left[Q_c + \alpha\mu_c + \alpha'\mu_{rc} - (1-\beta)\pi - 2\sqrt{\omega(\alpha+\rho\alpha')}\right]/(2\beta)$*;*

ii) the induced arrival rate is $\lambda(\mu^*, p^*) = [Q_c + \alpha\mu_c + \alpha'\mu_{rc} - (1-\beta)\pi]/[2(\alpha + \rho\alpha')] - \sqrt{\omega/(\alpha+\rho\alpha')}$*;*

iii) the expected waiting time is $\sqrt{(\alpha+\rho\alpha')/\omega}$*.*

Comparing Proposition 6 with Proposition 1 reveals several implications of taking into account the physician's diagnosis time. First, the physician orders more tests (i.e., chooses a lower service rate for consultation and diagnostic tests) compared to the baseline scenario when $\rho > (\alpha\mu_{rc})/[Q_c + \alpha\mu_c - (1-\beta)\pi]$, and orders fewer tests otherwise. Second, patients always wait longer since the

physician's availability is restrained by the time dedicated to reading and analyzing the results of diagnostic tests. Interestingly, as the processing rate ρ increases, patients' expected waiting time increases even though the physician speeds up the process of reading and analyzing results. This is because, as the processing rate ρ increases, the physician compensates the loss of service quality by ordering a greater number of tests (i.e., μ^* decreases), which leads to more salient congestion. Third, our major results from the baseline model remain qualitatively unchanged.

Social Optimum. The social planner chooses μ and λ to maximize the expected social welfare rate that is formulated as follows:

$$U_r(\mu,\lambda) = \frac{\rho}{1+\rho}\cdot\lambda\cdot\{Q_r(\mu,\mu_r) - \omega\mathbb{E}[W(\mu,\lambda)]\}.$$

Maximizing $U_r(\mu,\lambda)$ gives the socially optimal service and arrival rates: $\mu^{SE} = (Q_c + \alpha\mu_c + \alpha'\mu_{rc})/[2(\alpha+\rho\alpha')] \geq \mu^*$ and $\lambda^{SE} = \mu^{SE} - \sqrt{\omega/(\alpha+\rho\alpha')} \geq \lambda(\mu^*, p^*)$. Hence the phenomenon of overtesting sustains, confirming the robustness of our baseline model.

Endogenous Processing Rate. Next, we consider the scenario where the physician can endogenously determine the processing rate ρ (and hence $\mu_r = \rho\cdot\mu$). In other words, the physician has three decision variables: μ, μ_r, and p. To optimize physician's revenue rate, we substitute the intermediate results in Proposition 6 to represent the physician's revenue rate as a function of ρ:

$$h(\rho) := \frac{\rho}{\beta(1+\rho)}\left(\frac{K}{2\sqrt{\alpha+\rho\alpha'}} - \sqrt{\omega}\right)^2,$$

where $K := Q_c + \alpha\mu_c + \alpha'\mu_{rc} - (1-\beta)\pi$ is a constant that denotes the maximum possible economic utility that a patient can obtain from the outpatient service less the net copayment.

While $h(\rho)$ for $\rho > 0$ is not concave, we can show that $h'(\rho)$ always crosses zero twice, where the first zero corresponds to the optimal solution, and the second zero, denoted as $\hat{\rho} := (K^2 - 4\alpha\omega)/(4\alpha'\omega)$, constitutes the global minimum. We omit the closed-form representation of the optimal solution ρ^* due to excessive length. Instead, in the following proposition, we provide the implicit equation for ρ^*.

PROPOSITION 7. *When the physician endogenously chooses the service rate for diagnosis,*

i) the physician chooses the service rate $\mu^* = \frac{Q_c+\alpha\mu_c+\alpha'\mu_{rc}-(1-\beta)\pi}{2(\alpha+\rho^*\alpha')}$, $\mu_r^* = \rho^*\mu^*$, *where* ρ^* *is the unique solution to*

$$K[2\alpha - \alpha'\rho(\rho-1)] - 4\sqrt{\omega}(\alpha+\alpha'\rho)^{3/2} = 0,\ \ 0<\rho<\hat{\rho},$$

and the service fee is given by $p^* = \frac{1}{2\beta}\left[Q_c + \alpha\mu_c + \alpha'\mu_{rc} - (1-\beta)\pi - 2\sqrt{\omega(\alpha+\rho^*\alpha')}\right]$;

ii) the induced arrival rate is $\lambda(\mu^*, p^*) = \frac{Q_c+\alpha\mu_c+\alpha'\mu_{rc}-(1-\beta)\pi}{2(\alpha+\rho\alpha')} - \sqrt{\frac{\omega}{\alpha+\rho^*\alpha'}}$;

iii) the expected waiting time is $E(W(\lambda, \mu, \mu_r)) = \sqrt{\frac{\alpha+\rho^*\alpha'}{\omega}}$.

We conduct numerical experiments to glean managerial insights from the equilibrium characterization. Figure 2 illustrates the impact of the insurance structure on the physician's choice of $\mu^*, \mu_r^* = \rho^*\mu^*$ and p^*. Figure 2 demonstrates that insurance structure has the same directional effect on physician behavior as in the baseline model (cf. Corollary 1 and its discussions) after incorporating the physician's time spent in reading and analyzing test results. Figure 3 shows that, *ceteris paribus*, as the physician's time spent on diagnosi becomes more valuable (α' increases), the physician orders fewer tests and devotes more time in reading and analyzing test results, which in turn prolongs the expected waiting time.

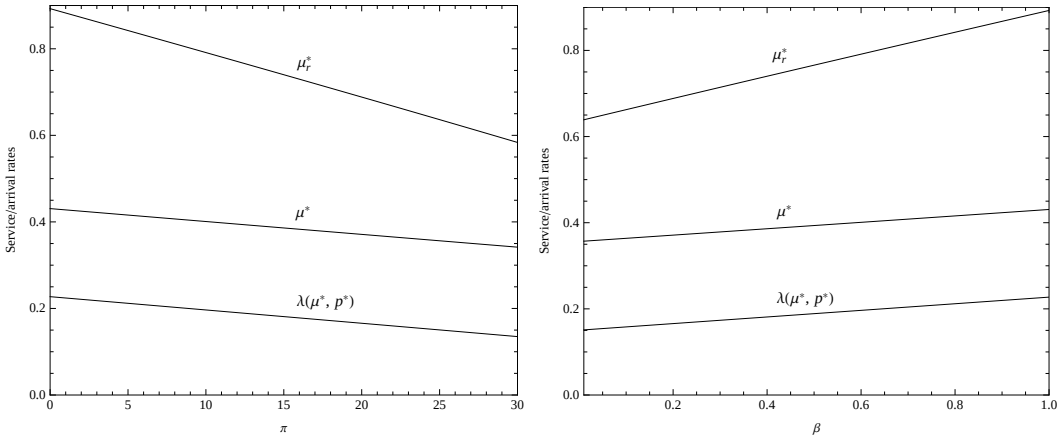

Figure 2 **Effect of insurance structure (left panel: copayment π; right panel: coinsurance rate β) on optimal service rates (μ^* and $\mu_r^* = \rho^*\mu^*$) and equilibrium arrival rate $\lambda(\mu^*, p^*)$. Service parameters are: $\mu_c = 0.8, Q_c = 8, \omega = 5, \alpha = 100, \mu_{rc} = 1.6$, and $\alpha' = 10$. In the left panel, $\beta = 0.2$; in the right panel, $\pi = 5$.**

3.5 Physician Type Uncertainty

In this section, we model the encounters between patients and the physician under initial uncertainty of the physician's skill level. In contrast to the preceding sections, the physician's skill level, referred to as "type," is unobservable to patients. The physician's type is denoted by $s \in \{h, l\}$, and a type s physician's skill level is α_s. We assume that $\alpha_h > \alpha_l$, indicating that, given the amount of diagnostic tests, the service provided by a type h physician yields higher diagnostic certainty. Unaware of the physician's type, patients are provided with access to the physician's pricing information before choosing a physician.

The interaction between the physician and patients lasts for two service periods and proceeds as follows. Figure 4 provides a time line depicting the sequence of events. At $t = -1$, the physician

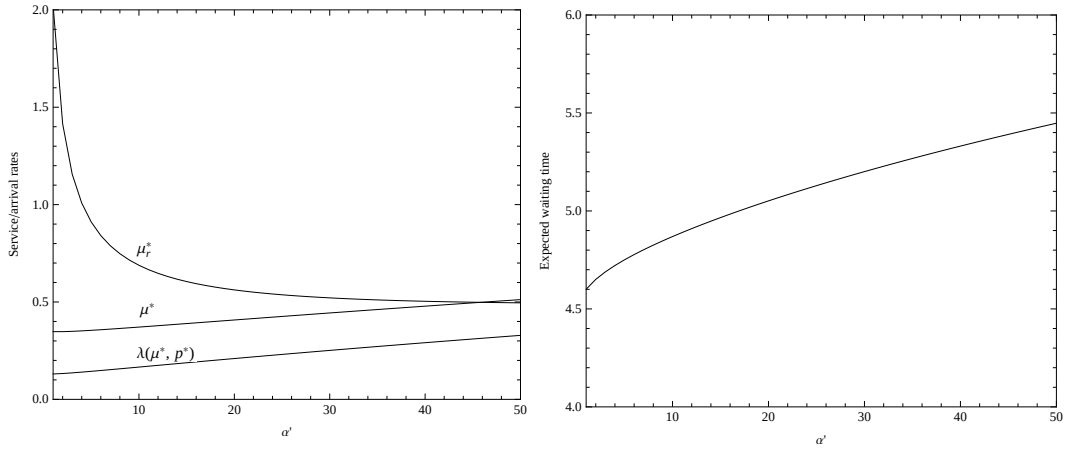

Figure 3 **Effect of α' on arrival/service rates (optimal service rates μ^* and $\mu_r^* = \rho^*\mu^*$, and equilibrium arrival rate $\lambda(\mu^*, p^*)$) and expected waiting time. Service parameters are: $\mu_c = 0.8, Q_c = 8, \omega = 5, \alpha = 100, \mu_{rc} = 1.6, \beta = 0.2$, and $\pi = 5$.**

discovers her own type s, which can be either h (with prior probability ψ) or l (with prior probability $(1-\psi)$). Patients are perfectly informed of the distribution of the physician's type but not its realization. At $t=0$, the physician sets her service fee p_s which remains unchanged thereafter. After observing the posted service fee p_s, patients form their posterior beliefs such that the probability is $\Psi(p_s)$ that the physician is of type h, $(1-\Psi(p_s))$ that the physician is of type l. Anticipating patients' beliefs of her type, the physician chooses the service rate in the first period, denoted by $\mu_{s1}, s \in \{l, h\}$. The equilibrium arrival rate during the first service period, denoted by $\lambda_1(\mu_{s1}|p_s)$, is affected by both patients' posterior beliefs as well as the service rate chosen by the physician, and can be solved from the following equation:

$$Q_h(\mu_{s1}) \cdot \Psi(p_s) + Q_l(\mu_{s1}) \cdot (1 - \Psi(p_s)) - \omega \mathbb{E}[W(\mu_{s1}, \lambda_1(\mu_{s1}|p_s))] - p_s = 0, s = l, h,$$

where $Q_s(\mu) = Q_c + \alpha_s(\mu_c - \mu)$ for $s \in \{h, l\}$ represents the service quality given the service rate μ when the physician type is h and l, respectively. At $t=1$, the physician chooses the service rate μ_{s1} to maximize her total expected utility during the two service periods; patients make their queue-joining decisions based on their belief structure $\Psi(p_s)$. Let $\lambda_1(\mu_{s1}|p_s)$ denote the equilibrium arrival rate in the first service period when the type s physician chooses the service rate μ_{s1}. At $t=0$, the physician chooses p_s and μ_{s1} that maximize

$$p_s \left\{ \lambda_1(\mu_{s1}|p_s) + \max_{\mu_{s2}} [\lambda_2(\mu_{s2})] \right\}, s = h, l.$$

At $t=2$, the physician type is revealed to patients. The queue-joining decisions faced by patients, as well the service rate decision faced by the physician, are similar to those in the baseline model.

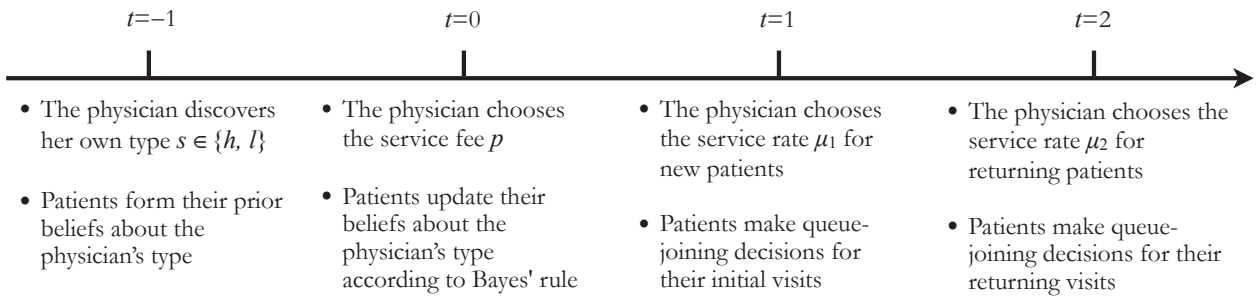

Figure 4 Time line of the model with physician type uncertainty

Applying the standard methodology of solving similar problems (e.g., Debo and Veeraraghavan 2010), we model the physician-patient interaction as a sequential game of incomplete information and establish a Perfect Bayesian Equilibrium. We restrict attention to pure strategy equilibria satisfying the intuitive criterion of Cho and Kreps (1987). The intuitive criterion is an equilibrium refinement which restricts beliefs off the equilibrium path. In particular, it requires that the updating of beliefs should not assign positive probability to a player taking an action that is equilibrium dominated. Essentially, the intuitive criterion allows us to eliminate any perfect Bayesian equilibrium from which some type of physician would want to deviate even if she were not sure what exact belief the patients would have as long as she knows that the patients would not think she is a type who would find the deviation equilibrium dominated.

Full Information Benchmark. We first consider the case in which patients can observe the physician's optimal service choices under full information. The physician's service decision, together with patients' corresponding queue-joining strategy, constitutes the full-information equilibrium. By choosing the service rate μ and the service fee p, the per-period revenue that the type s physician collects from patients is $g_s(\mu,p)$ for $s\in\{h,l\}$. Let the pair (μ_s^*,p_s^*) denote the physician's optimal decision. The following lemma is immediate from Proposition 1.

LEMMA 1. *The full-information equilibrium is unique and characterized as follows:*

i) The physician chooses the service rate $\mu_s^*=\frac{Q_c+\alpha_s\mu_c-(1-\beta)\pi}{2\alpha_s}$, *and the service fee* $p_s^*=\frac{Q_c+\alpha_s\mu_c-2\sqrt{\omega\alpha_s}-(1-\beta)\pi}{2\beta}$ *for* $s\in\{h,l\}$.

ii) Patients choose their queue-joining strategy so that the induced arrival rate is $\lambda_s(\mu_s^*,p_s^*)=\frac{Q_c+\alpha_s\mu_c-(1-\beta)\pi}{2\alpha_s}-\sqrt{\frac{\omega}{\alpha_s}}$ *for* $s\in\{h,l\}$.

iii) The average waiting time is $\mathbb{E}[W(\mu_s^*,\lambda_s(\mu_s^*,p_s^*))]=\sqrt{\frac{\alpha_s}{\omega}}$ *for* $s\in\{h,l\}$.

It is straightforward to see from Lemma 1 that both types of physicians overtest in the full-information equilibrium. This serves as a foundation for us to understand whether asymmetric

information about the physician type would exacerbate or alleviate overtesting. We further define the following two functions to facilitate the subsequent analysis:

$$\hat{g}_s(p) := \max_{\mu} g_s(\mu, p) = p \cdot \left(\frac{Q_c + \alpha_s \mu_c - \pi - \beta(p-\pi)}{\alpha_s} - 2\sqrt{\frac{\omega}{\alpha_s}} \right), s \in \{h, l\}, \text{ and}$$
$$\hat{\mu}_s(p) := \arg\max_{\mu} g_s(\mu, p) = \frac{Q_c + \alpha_s \mu_c - \pi - \beta(p-\pi)}{\alpha_s} - \sqrt{\frac{\omega}{\alpha_s}}, s \in \{h, l\},$$

which are the type s physician's maximum total revenue and optimal service rate, respectively, when the service fee is fixed at p. In the next corollary, we compare the two types of physicians' revenue rates under the full-information equilibrium.

COROLLARY 11. *If patients can reliably distinguish the type h physician from the type l physician ex ante, then the type h physician's expected revenue rate is always higher than the type l physician's, that is,* $g_h(\mu_h^*, p_h^*) > g_l(\mu_l^*, p_l^*)$.

Corollary 11 suggests that, when the physician type is unobservable to patients, the type h physician prefers to be separated from the type l physician, while the type l physician prefers not to be separated from the type h physician. In other words, the type l physician might have the incentive to *mimic* the type h physician.

Physicians' Test-Ordering Behavior Under Physician Type Uncertainty. We now characterize the equilibrium for the asymmetric-information game and discuss its implication for the physician's test-ordering behavior. In the following proposition, we show that the full-information equilibrium might arise as an outcome of the game. Another situation is that the type h physician manages to separate from the type l physician by deviating from the full-information equilibrium. We make the following two assumptions:

ASSUMPTION 1. $\alpha_h > \alpha_l \geq \omega/\mu_c^2$

ASSUMPTION 2. $\pi(1-\beta) + \frac{\omega}{\mu_c + 2\sqrt{\omega/\alpha_h}} \leq Q_c \leq \pi + \beta(p-\pi) + \sqrt{\alpha_l \omega}$

These two assumptions specifies the boundaries of service parameters to facilitate the characterization of the equilibrium; see the appendix for their intuitive explanations.

PROPOSITION 8. *Given* α_h*, there always exists* $\Delta\alpha^* > 0$ *satisfying* $g_h(\mu_h^*, p_h^*) + \hat{g}_l(p_h^*) - 2g_l(\mu_l^*, p_l^*) = 0$ *at* $\alpha_l = \alpha_h - \Delta\alpha^*$ *such that the resultant separating equilibrium has two possible cases:*

a. ***Costless separating equilibrium.*** *When* $\Delta\alpha \geq \Delta\alpha^*$*, there exists a unique separating equilibrium in which*

(i) the type h physician charges $p_h = p_h^*$*;*

(ii) the type l physician charges $p_l = p_l^*$*;*

(iii) patients' beliefs are: $\Psi(p) = 1$ *if* $p = p_h^*$*;* $\Psi(p) = 0$ *otherwise.*

b. ***Costly separating equilibrium.*** *When* $\Delta\alpha < \Delta\alpha^*$*, and there exists* $p_h' > p_h^*$ *that maximizes the type h physician's total expected utility rate subject to the following two constraints:*

$$\hat{g}_h(p_h') + \hat{g}_l(p_h') \le 2g_l(\mu_l^*, p_l^*), \text{ and } g_l(\mu_l^*, p_l^*) + \hat{g}_h(p_l^*) \le 2\hat{g}_h(p_h'), \tag{15}$$

and a separating equilibrium sustains in which

(i) the type h physician charges $p_h = p_h' > p_h^*$*;*

(ii) the type l physician charges $p_l = p_l^*$*;*

(iii) the patient's beliefs are: $\Psi(p) = 1$ *if* $p = p_h'$*;* $\Psi(p) = 0$ *otherwise.*

Under the costless separating equilibrium, both types of physicians behave as if in the full-information equilibrium. When the physicians' skill level difference $\Delta\alpha$ is low enough, the type l physician has the incentive to mimic the type h physician. In this case, a separating equilibrium prevails if the type h physician manages to signal her type by deviating from the full-information equilibrium. The signal is said to be *costly* because the type h physician sacrifices a proportion of her revenue to deter the type l physician from mimicking. Under the costly separating equilibrium, the type h physician chooses a service fee higher than p_h^* to signal her type. In the meantime, the type h physician orders more tests than in the full information equilibrium to compensate for patients' utility loss. In other words, this costly signaling effort essentially encourages the type h physician's overtesting behavior. The first half of (15) ensures that the type l physician does not have the incentive to mimic the type h physician, while the second half ensures that the type h physician is better off charging a higher service fee and prescribing more tests rather than mimicking the type l physician.

Outside the separating equilibria we have discussed, one might expect a pooling equilibrium to arise as a possible outcome of the incomplete-information game. In a pooling equilibrium, the high and low types offer the same service fee and hence patients are unable to update their beliefs. The next proposition, however, shows that a pooling equilibrium is *not* a possible outcome of the signaling game with reasonable beliefs. In other words, if $\Delta\alpha < \Delta\alpha^*$, then there always exists p' that satisfies both (**??**) and (15).

PROPOSITION 9. *There exists no pooling equilibrium that satisfies the intuitive criterion.*

The intuition behind Proposition 9 is that the high type can always exploit the economic benefits of providing higher diagnostic certainty in order to separate from the low type while such a deviation would be dominated for the low type.

Two observations may now be made.

Observation 1. *Price transparency might lead to higher service fees.*

The opacity in pricing in health care services is a well-known phenomenon that separates the health care industry from markets for most goods and services. The pending *Transparency in All Health Care Pricing Act of 2010* will require all the health care providers to post prices for various services. Notwithstanding many intuitive benefits associated with price transparency, the Congressional Budget Office (2008), by citing empirical evidences from other industries, contends that increasing transparency in the healthcare market can result in higher prices.

Proposition 9 indicates that, with price transparency, a pooling equilibrium, in which both types of physicians choose a medium service level, can never sustain; price transparency encourages the type h physician to overtest, and prevents the type l physician from mimicking the type h physician. By comparison, under pricing opacity, a pooling equilibrium sustains. This gives an implication similar to the finding by the Congressional Budget Office (2008) albeit from a different angle: price transparency leads to higher prices and encourages the prescription of unnecessary tests.

Observation 2. *Improved diagnostic technology can either exacerbate or alleviate the phenomenon of overtesting.*

The medical community has divided views regarding whether improved technology will increase or reduce healthcare expenditure and the social welfare (Newhouse 1992). We conduct numerical experiments to understand the impact of improved diagnostic technology. We maintain the type h physician's skill level and increase the type l physician's skill level gradually to reflect the notion that improved technology flattens out the skill level differences among physicians. The results, as shown in Figure 5, illustrate that technology advancements can either exacerbate or alleviate the phenomena of overtesting depending on the range of the skill level differences between the two types of physicians:

- Region I: the physicians' skill level difference is high (α_l is low). In this case, the costless separating equilibrium prevails, and the improvement in diagnostic technology has little impact on the physician's test-ordering behavior.
- Region II: the physicians' skill level difference is medium (α_l is medium). In this case, the type h physician uses a costly signal to separate from the type l physician. As technological advancements

lead to less differentiation between physicians in the level of skill, even though patients ahieve diminishing service quality gains by switching from the type l physician to the type h one, the type h physician has an even stronger incentive to overtest as a costly signaling effort. In other words, the improvement of diagnostic technology leads to more salient overtesting behavior.

- Region III: the physicians' skill level difference is low (α_l is comparable to α_h). The costly separating equilibrium continues to prevail, but an increased α_l makes it less rewarding for the type h physician to signal her type. As a consequence, the improvement in diagnostic technology leads to a lower incentive for overtesting.

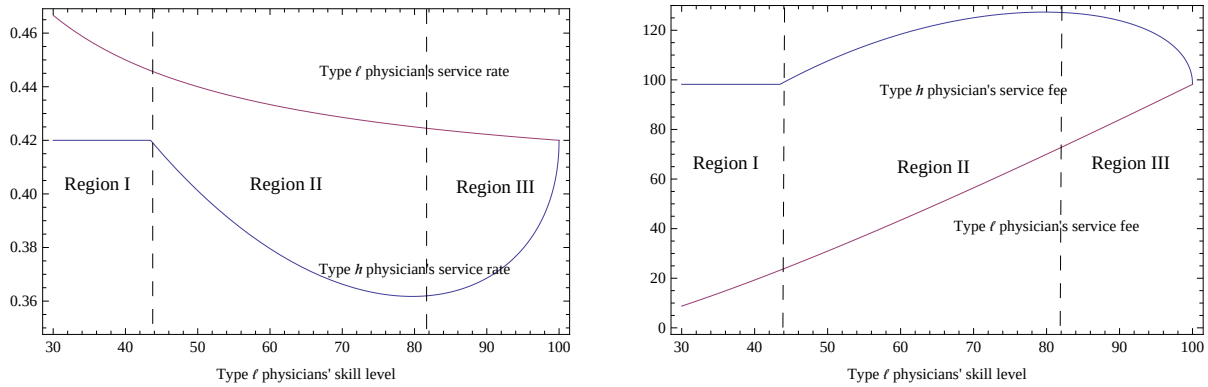

Figure 5 **The impact of the different skill level differences among physicians on the physicians' service rates and fees. Parameters are:** $\mu_c = 0.8, Q_c = 8, \omega = 5, \pi = 5, \beta = 0.2, \alpha_h = 100$.

4. Concluding Remarks

This work is—to our best knowledge—the first to analytically investigate financial, operational, and clinical incentives behind physicians' test-ordering behavior. Our baseline model reveals that overtesting always occurs due to the existence of insurance coverage, and the copayment and coinsurance play reverse roles in affecting the equilibrium service rate: with a higher copayment, the physician orders more tests; with a higher coinsurance rate, however, the physician orders fewer tests. Then we consider five different service environments: i) reimbursement ceiling, ii) misdiagnosis concerns, iii) patient heterogeneity, iv) diagnosis time, and v) uncertainty about the physician type. First, we show that setting a reimbursement ceiling alone cannot eliminate overtesting, and, surprisingly, even when the ceiling is low, either a high copayment or a low coinsurance rate could encourage the physician to order more diagnostic tests. Second, we show that, when physicians are concerned about inaccurate diagnosis, both overtesting and undertesting are possible outcomes, and the waiting time in equilibrium is shorter than is socially efficient. Third, we consider patient

heterogeneity and show that the resultant service rate becomes lower in both the market equilibrium and the social optimum. Fourth, the impact of insurance structure carries over when considering the physician's diagnosis time. Last but not least, we address the issue of information asymmetry about physicians' skill levels. We rule out the occurrence of a pooling equilibrium and show that price transparency can fuel the type h physician's costly signaling efforts. Furthermore, technology improvements have mixed effects on overtesting.

We highlight a few key operational and policy implications from our work. First, overtesting is a complex phenomenon that cannot be eliminated by simple fixes, such as imposing a reimbursement ceiling, or eliminating insurance coverage all at once. As physicians' test-ordering behavior is closely tied to patients' strategic responses, a comprehensive understanding of physicians' and patients' clinical, operational, and monetary incentives is essential before embarking on any radical changes in the public policy. Second, physicians' misdiagnosis concerns lead to overtesting only when bundled together with a certain incentive environment. To address the issue of overtesting, therefore, requires not just adjusting the relevant incentive structures, but also a legal environment that supports physicians' broader adoption of evidence-based guidelines. This aspect supports physicians' expanding implementation of evidence-based guidelines (Walshe and Rundall 2001) and contemporary political discourse (White House 2009). Third, two factors are important in evaluating the physician's test-ordering behavior: physicians' skill level difference, and the lack of publicly accessible knowledge of such information. Making professional evaluation for physicians more transparent to the public, through credible and accessible channels, indeed helps reduce overtesting associated with costly signaling efforts.

Acknowledgments

We are greatly indebted to Joshua Rheinbolt, MD and Robert J. Noecker, MBA, MD at the UPMC Eye Center for devoting countless hours to collaborating in the patient experience improvement project and providing insiders' thoughts that directly motivated this research. Thanks to the Editor, Stephen Graves, and the two special issue Co-Editors, Abraham (Avi) Seidmann and Bruce Golden, and the three anonymous reviewers for writing detailed, insightful review reports that helped considerably improve the quality of the paper. We also received valuable comments from Alan Scheller-Wolf, Laurens Debo, R. Ravi, Soo-Haeng Cho, and Katia Sycara, and seminar participants at Carnegie Mellon University, INFORMS 2010 Annual Meeting, MSOM 2011 Healthcare Operations Management SIG Meeting, and MSOM Service Management SIG at INFORMS 2011 Annual Meeting. Tinglong Dai was partially supported by an Office of Naval Research grant N0001409-10680.

References

Alderman, L. 2011. The doctor will see you ... eventually. *New York Times* (August 2) D6.

Anand, K. S., M. F. Pac, S. K. Veeraraghavan. 2011. Quality-speed conundrum: tradeoffs in customer-intensive services. *Management Sci.* **57**(1) 40–56.

Baicker, K., E. S. Fisher, A. Chandra. 2007. Malpractice liability costs and the practice of medicine in the medicare program. *Health Affair.* **26**(3) 841–852.

Carrier, E. R., J. D. Reschovsky, M. M. Mello, R. C. Mayrell, D. Katz. 2010. Physicians' fears of malpractice lawsuits are not assuaged by tort reforms. *Health Affair.* **29**(9) 1585–1592.

Cho, I.-K., D. M. Kreps. 1987. Signaling games and stable equilibria. *Quart. J. Econom.* **102**(2) 179–221.

Coffey, R. M. 1983. The effect of time price on the demand for medical-care services. *J. Human Res.* **18**(3) 407–424.

Congressional Budget Office. 2008. Increasing transparency in the pricing of health care services and pharmaceuticals. *Economic and Budget Issue Brief* (June 5).

Connell, F. A., T. D. Koepsell. 1985. Measures of gain in certainty from a diagnostic test. *Amer. J. Epidemiol.* **121**(5) 744–753.

Debo, L., S. K. Veeraraghavan. 2010. Prices and congestion as signals of quality. Working paper.

Debo, L., B. Toktay, L. V. Wassenhove. 2008. Queuing for expert services. *Management Sci.* **54**(8) 1497–1512.

Dulleck, U., R. Kerschbamer. 2006. On doctors, mechanics, and computer specialists: the economics of credence goods. *J. Economic Literature* **44**(1) 5–42.

Economist. 2010. Clear diagnosis, uncertain remedy. *The Economist* (Feb 18).

Evans, R. G. 1974. Supplier-induced demand: some empirical evidence and implications. M. Perlman ed. *The Economics of Health and Medical Care.* Macmillan, London, U.K. 162–201.

Feldstein, M. S. 1973. The welfare loss of excess health insurance. *J. Political Econom.* **81**(March–April) 251–280.

Gravelle, H., L. Siciliani. 2008. Optimal quality, waits and charges in health insurance. *J. Health Econom.* **27**(3) 663–674.

Green, L. 2006. Queueing analysis in healthcare. Hall, R.W., ed. *Patient Flow: Reducing Delay in Healthcare Delivery.* Springer, New York.

Habbema, J. D. F., R. Eijkmans, P. Krijnen, J. A. Knottnerus. 2002. Analysis of data on the accuracy of diagnostic tests. Knottnerus JA, ed. *The Evidence Base of Clinical Diagnosis.* BMJ Books, London, U.K., 117–143.

Harchol-Balter, M. 2011. *Performance Analysis and Design of Computer Systems.* Working report. Carnegie Mellon University.

Harvard Team. 1999. *Harvard Report: Improving Hong Kong's Health Care System: Why and for Whom?* Hong Kong SAR Food and Health Bureau.

Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems.* Kluwer Academic Publishers, Norwell, MA.

Jauhar, S. 2009. A doctor by choice, a businessman by necessity. *New York Times* (July 7) D5.

Jung, K.-T. 1998. Influence of a per-visit copayment on health care use and expenditures: The Korean experience. *J. Risk Ins.* **65**(1) 33–56.

Kassirer, J. P. 1989. Our stubborn quest for diagnostic certainty: a cause of excessive testing. *N. Engl. J. Med.* **320**(22) 1489–1491.

Kleinrock, L. 1975. *Queueing Systems. Vol. I: Theory.* John Wiley & Sons, New York, NY.

Kostami, V., S. Rajagopalan. 2009. Speed quality tradeoffs in a dynamic model. University of Southern California working paper.

Mold, J. W., R. M. Hamm, L. H. McCarthy. 2010. The law of diminishing returns in clinical medicine: how much risk reduction is enough? *J. Amer. Board Fam. Med.* **23** 371–375.

Newhouse, J. P. 1978. Insurance benefits, out-of-pocket payments, and the demand for medical care: a review of the literature, RAND: Santa Monica, CA.

Newhouse, J. P. 1992. Medical care costs: how much welfare loss? *J. Economic Perspectives* **6**(3) 3–21.

Orszag, P. R. 2008. Increasing the value of Federal spending on health care, testimony. Congressional Budget Office. July 16.

Pauly, M. V. 1980. *Doctors and Their Workshops: Economic Models of Physician Behavior.* The University of Chicago Press, Chicago, IL.

Phelps, C. E., J. P. Newhouse. 1974. Coinsurance, the price of time, and the demand for medical services. *Rev. Econom. Stat.* **56**(3) 334–342.

Skrondal, A., S. Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Chapman & Hall, Boca Raton, FL.

Sorensen, R., J. Grytten. 1999. Competition and supplier-induced demand in a health care system with fixed fees. *Health Econom.* **8**(6) 497–508.

Studdert, D. M., M. M. Mello, A. A. Gawande, T. K. Gandhi, A. Kachalia, C. Yoon, A. L. Puopolo, T. A. Brennan. 2006. Claims, errors, and compensation payments in medical malpractice litigation. *N. Engl. J. Med.* **354**(19): 2024–2033.

Veeraraghavan, S. K., Debo, L. 2009. Joining longer queues: information externalities in queue choice. *Manufacturing Service Oper. Management* **11**(4) 543–562.

Walshe, K., T. G. Rundall. 2001. Evidence-based management: from theory to practice in health care. *Milbank Quarterly* **79**(3) 429–457.

Wang, X., L. G. Debo, A. Scheller-Wolf, S. F. Smith. 2010. Design and analysis of diagnostic service centers. *Management Sci.* **56**(11) 1873–1890.

White House. 2009. Remarks by the President at the annual conference of the American Medical Association. `http://www.whitehouse.gov`

Appendix A: Technical Proofs

Proof of Proposition 1. We first show that the physician's objective function specified in (3) is concave in p since

$$\frac{\partial^2 g(\mu,p)}{\partial p^2} = -\frac{2p\beta^2\omega}{[Q(\mu)-\pi-\beta(p-\pi)]^3} - \frac{2\beta\omega}{[Q(\mu)-\pi-\beta(p-\pi)]^2} < 0.$$

Solving the first-order condition gives the optimal service fee p^*, conditional on the service rate μ: $p^*(\mu) = \left\{\mu Q(\mu) - \mu(1-\beta)\pi - \sqrt{\mu\omega[Q(\mu)-(1-\beta)\pi]}\right\}/(\mu\beta)$. Let $g(\mu,p)$ denote the physician's payoff under the service rate μ and the service fee p, we see that

$$g(\mu,p^*(\mu)) = \left\{\mu\left[Q_c+\alpha\ (\mu_c-\mu)-\pi(1-\beta)\right]+\omega-2\sqrt{\mu\omega\left[Q_c+\alpha\ (\mu_c-\mu)-\pi(1-\beta)\right]}\right\}/\beta.$$

Next, we show that $g(\mu,p^*(\mu))$ is unimodal in μ. Note that

$$\begin{aligned}
\mu(Q_c+\pi(\beta-1)-\alpha\mu+\alpha\mu_c) &= \mu[Q(\mu)+\pi(\beta-1)] \\
&> \mu[Q(\mu) - \underbrace{(\beta p+\pi(1-\beta))}_{\text{Out-of-pocket expense}}] \\
&\geq \mu\cdot \underbrace{\frac{\omega}{\mu-\lambda(\mu,p^*(\mu))}}_{\text{Waiting costs}} \\
&= \omega\cdot\frac{\mu}{\mu-\lambda(\mu,p^*(\mu))} > \omega,
\end{aligned}$$

which gives

$$\sqrt{\underbrace{\mu(Q_c+\pi(\beta-1)-\alpha\mu+\alpha\mu_c)}_{>\omega}\omega}-\omega>0.$$

Hence we see that the sign of

$$\frac{dg(\mu,p^*(\mu))}{d\mu} = [Q_c+\pi(\beta-1)+\alpha(-2\mu+\mu_c)]\cdot\frac{\overbrace{\sqrt{\mu(Q_c+\pi(\beta-1)-\alpha\mu+\alpha\mu_c)\omega}-\omega}^{>0}}{\beta\sqrt{\mu(Q_c+\pi(\beta-1)+\alpha(-\mu+\mu_c))\omega}}$$

is the same as that of $Q_c+\pi(\beta-1)+\alpha(-2\mu+\mu_c)$, which is positive when $\mu=0$, decreases in μ, and turns negative when μ is large enough. $g(\mu,p^*(\mu))$ is therefore unimodal in μ. Equating the first-order derivative of $g(\mu,p^*(\mu))$ in terms of μ to zero gives $\mu^*=[Q_c+\alpha\mu_c-(1-\beta)\pi]/(2\alpha)$, which in turn yields $p^*=[Q_c+\alpha\mu_c-(1-\beta)\pi-2\sqrt{\alpha\omega}]/(2\beta)$, and $\lambda(\mu^*,p^*)=[Q_c+\alpha\mu_c-(1-\beta)\pi]/(2\alpha)-\sqrt{\omega/\alpha}=[Q_c+\alpha\mu_c-(1-\beta)\pi-2\sqrt{\alpha\omega}]/(2\alpha)$. The expected waiting time can thus be determined given μ^* and $\lambda(\mu^*,p^*)$: $\mathbb{E}[W(\mu^*,\lambda(\mu^*,p^*))]=1/[\mu^*-\lambda(\mu^*,p^*)]=\sqrt{\alpha/\omega}$. $\square$

Proof of Proposition 2. We first recognize that $U(\mu,\lambda)$ is concave in μ as $\partial^2 U(\mu,\lambda)/\partial\mu^2 = -2\lambda\omega/(\mu-\lambda)^3 < 0$ for any pair of (μ,λ) that satisfies $\mu > \lambda$. By solving the first-order condition of (4) in terms of μ, we obtain the conditional expression of the optimal service rate: $\mu^{SE}(\lambda) = \lambda + \sqrt{\omega/\alpha}$, which, together with (4), simplifies the objective function as $-\alpha\lambda^2 + [\alpha\mu_c + Q_c - 2\sqrt{\alpha\omega}]\,\lambda$, a concave function of λ. The first-order condition gives $\lambda^{SE} = (Q_c + \alpha\mu_c)/(2\alpha) - \sqrt{\omega/\alpha}$, and hence $\mu^{SE} = (Q_c + \alpha\mu_c)/(2\alpha)$. The expected waiting time is thus $W(\mu^{SE},\lambda^{SE}) = 1/(\mu^{SE} - \lambda^{SE}) = \sqrt{\alpha/\omega}$. □

Proof of Corollary 3. The social welfare gap, written as a function of β and π, is $\Delta U(\pi,\beta) = U(\mu^{SE},\lambda^{SE}) - U(\mu^*,\lambda^*) = \pi^2(1-\beta)^2/(4\alpha)$, and its second-order derivatives in terms of β and π are $\partial^2\Delta U/\partial\beta^2 = \pi^2/(2\alpha) \geq 0$, and $\partial^2\Delta U/\partial\pi^2 = (1-\beta)^2/(2\alpha) \geq 0$, respectively. Hence $\Delta U(\pi,\beta)$ is convex decreasing in β, and convex increasing in π. □

Proof of Proposition 3. Similar to the proof of Proposition 1. □

Proof of Corollary 5. We have two cases to consider depending on the size of $p_{\max}$ (cf. Proposition 3). Case i): $p_{\max} > [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} - (1-\beta)\pi]/(2\beta)$. In this case, we have $\mu^* = [Q_c + \alpha\mu_c - (1-\beta)\pi]/(2\alpha) < \mu^{SE} = (Q_c + \alpha\mu_c)/(2\alpha)$. Case ii): $p_{\max} \leq [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} - (1-\beta)\pi]/(2\beta)$. Since $p_{\max} \leq [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} - (1-\beta)\pi]/(2\beta)$, we have $q_{\max} = \pi + \beta(p_{\max} - \pi) \leq [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} + (1-\beta)\pi]/2$. We can thus further divide case ii) into two sub-cases depending on the size of $q_{\max}$: (a) $(Q_c + \alpha\mu_c - 2\sqrt{\alpha\omega})/2 < q_{\max} \leq [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} + (1-\beta)\pi]/2$, in which sub-case we have $\mu^* = (Q_c + \alpha\mu_c - q_{\max} - \sqrt{\alpha\omega})/\alpha < (Q_c + \alpha\mu_c/(2\alpha) = \mu^{SE}$; (b) $q_{\max} \leq [Q_c + \alpha\mu_c - 2\sqrt{\alpha\omega}]/2$, in which sub-case we have $\mu^* = (Q_c + \alpha\mu_c - q_{\max} - \sqrt{\alpha\omega})/\alpha \geq (Q_c + \alpha\mu_c)/(2\alpha) = \mu^{SE}$. □

Proof of Propositions 4–6. Similar to the proof of Proposition 1. □

Proof of Proposition 7. We prove the result by examining the property of

$$h(\rho) = \frac{\rho}{\beta(1+\rho)}\left(\frac{K}{2\sqrt{\alpha+\rho\alpha'}} - \sqrt{\omega}\right)^2,$$

which is the product of two functions $h_1(\rho) = \frac{\rho}{\beta(1+\rho)}$, and $h_2(\rho) = \left(\frac{K}{2\sqrt{\alpha+\rho\alpha'}} - \sqrt{\omega}\right)^2$. To understand the property of $h(\cdot)$, we first note that $h_1(\cdot)$ and $h_2(\cdot)$: $h_1(\rho)$ is convex increasing in ρ, and $h_2(\rho)$ is convex decreasing in ρ if $\rho < \hat{\rho}$, and concave increasing in ρ otherwise. Moreover, we have the following observations:

(i) As $\rho \to \infty$, $h_1(\rho) \to 1, h_2(\rho) \to \omega$, and $h(\rho) = h_1(\rho)\cdot h_2(\rho) \to \omega$.

(ii) $h_2'(\hat{\rho}) = 0$ and $h_2(\hat{\rho}) = 0$, and hence $h'(\hat{\rho}) = h_1'(\hat{\rho})h_2(\hat{\rho}) + h_1(\hat{\rho})h_2'(\hat{\rho}) = 0$.

(iii) As $\rho > 0$ increases, $h(\rho)$ first increases, then decreases, and finally increases.

(iv) There exists a unique solution to $h'(\rho) = 0$ when $0 < \rho < \hat{\rho}$. This is true because (iii) means that $h'(\rho)$ cross zero twice, and the second zero corresponds to $\hat{\rho}$.

We conclude from the above observations that the maximum can only be achieved at the first zero of h', or at $\rho \to \infty$, because the latter solution dominates the former only when the diagnosis time generates zero

service value, which is an unrealistic scenario that we can safely ignore. Therefore, we conclude that ρ^* is the unique solution to $h'(\rho)=0, 0<\rho<\hat{\rho}$, which, after some algebra, can be rewritten as

$$[2\alpha-\alpha'\rho(\rho-1)]-4(\alpha+\rho\alpha')^{3/2}\sqrt{\omega}=0, 0<\rho<\hat{\rho},$$

which finishes the proof of the proposition. □

Proof of Lemma 1. The proof is similar to that of Proposition 1 except that we replace α with α_s for $s\in\{h,l\}$. □

Proof of Corollary 11. It is sufficient to show that for any combination of service parameters (μ, p), the type h physician always fares better than the type l physician. The overarching reason relates to the fact that for any pair (μ,p), the type h physician manages to attract a larger crowd, that is,

$$\lambda_h(\mu,p)=\mu-\frac{\omega}{Q_c+\alpha_h(\mu_c-\mu)-p}>\lambda_l(\mu,p)=\mu-\frac{\omega}{Q_c+\alpha_l(\mu_c-\mu)-p},$$

which is true because $\alpha_h>\alpha_l$. Therefore, we have $g_h(\mu,p)=p\lambda_h(\mu,p)>g_l(\mu,p)=p\lambda_l(\mu,p)$ for any feasible combination of (μ,p). This in turn gives $g_h(\mu_h^*,p_h^*)\geq g_h(\mu_l^*,p_l^*)>g_l(\mu_l^*,p_l^*)$. □

Explanations of Assumptions 1 and 2. Here we explain the intuition behind the two assumptions.

Assumption 1: $\pi(1-\beta)+\frac{\omega}{\mu_c+2\sqrt{\omega/\alpha_h}}\leq Q_c\leq\pi+\beta(p-\pi)+\sqrt{\alpha_l\omega}$. The right half of the inequality, $Q_c\leq\pi+\beta(p-\pi)+\sqrt{\alpha_l\omega}$, means that if the type l physician chooses the service rate at the baseline level μ_c, then the resultant service quality Q_c is insufficient to attract any demand since it is outweighed by the sum of each patient's money price ($\pi+\beta(p-\pi)$) and time price ($\sqrt{\alpha_l\omega}$). The left-hand side $\pi(1-\beta)+\frac{\omega}{\mu_c+2\sqrt{\omega/\alpha_h}}$ has two parts: the first part $\pi(1-\beta)$ means the patient's net copayment; the second part $\frac{\omega}{\mu_c+2\sqrt{\omega/\alpha_h}}$ is patients' waiting cost when the service rate is faster than μ_c and the arrival rate is zero. Hence it is reasonable to expect the left-hand side to be lower than the baseline service quality Q_c.

Assumption 2: $\alpha_h>\alpha_l\geq\omega/\mu_c^2$. Assumption 2 is made so that we can focus on the most realistic case that $p_h^*>p_l^*$. Recall from Lemma 1 that, in the full-information equilibrium, the average waiting time is $\sqrt{\alpha_s/\omega}, s=h,l.$, which includes both the queueing time and the service time. Thus, after rewriting assumption 2 as $1/\mu_c\leq\sqrt{\alpha_s/\omega}, s=h,l$, it becomes apparent that this assumption is not a strong one since it says that the expected baseline service time ($1/\mu_c$) is shorter than the total expected waiting time in equilibrium.

Proof of Proposition 8. To prove Proposition 8, we first present two intermediate results, namely, Lemmas 2 and 3, in the following:

LEMMA 2. *(i) $\hat{g}_h(p)>\hat{g}_l(p)$ and (ii) $\hat{g}_h(p)-\hat{g}_l(p)$ is increasing in p.*

Proof of Lemma 2. $\hat{g}_h(p)-\hat{g}_l(p)$ can be expanded as

$$\begin{aligned}\hat{g}_h(p)-\hat{g}_l(p)&=p\left[\frac{Q_c+\alpha_h\mu_c-\pi-\beta(p-\pi)}{\alpha_h}-2\sqrt{\frac{\omega}{\alpha_h}}\right]-p\left[\frac{Q_c+\alpha_l\mu_c-\pi-\beta(p-\pi)}{\alpha_l}-2\sqrt{\frac{\omega}{\alpha_l}}\right]\\&=p\underbrace{\left(\frac{1}{\sqrt{\alpha_l}}-\frac{1}{\sqrt{\alpha_h}}\right)}_{>0}\cdot\left\{2\sqrt{\omega}-[Q_c-\pi-\beta(p-\pi)]\left(\frac{1}{\sqrt{\alpha_l}}+\frac{1}{\sqrt{\alpha_h}}\right)\right\}.\end{aligned}$$

The term $2\sqrt{\omega}-[Q_c-\pi-\beta(p-\pi)]\left(\frac{1}{\sqrt{\alpha_l}}+\frac{1}{\sqrt{\alpha_h}}\right)$ is positive because

$$2\sqrt{\omega}-\underbrace{[Q_c-\pi-\beta(p-\pi)]}_{<\sqrt{\alpha_l\omega}\text{ (Assumption 2)}}\left(\frac{1}{\sqrt{\alpha_l}}+\frac{1}{\sqrt{\alpha_h}}\right)>2\sqrt{\omega}-\sqrt{\alpha_l\omega}\underbrace{\left(\frac{1}{\sqrt{\alpha_l}}+\frac{1}{\sqrt{\alpha_h}}\right)}_{<2/\sqrt{\alpha_l}}>0.$$

Therefore, $\hat{g}_h(p)-\hat{g}_l(p)$ is increasing in p. □

We define $\Delta\alpha$ as the two types of physicians' skill level differences, that is, $\Delta\alpha:=\alpha_h-\alpha_l$. The lemma below is also necessary to complete the proof of Proposition 8.

LEMMA 3. *(Single crossing property) Given α_h, as $\Delta\alpha$ increases, $g_h(\mu_h^*,p_h^*)+\hat{g}_l(p_h^*)-2g_l(\mu_l^*,p_l^*)$ is first positive, crosses zero once, and then remains negative. In other words, as $\Delta\alpha$ increases, the type l physician's benefit from mimicking the type h physician crosses zero only once.*

Proof of Lemma 3. The proof consists of two steps. We first show that $g_h(\mu_h^*,p_h^*)+\hat{g}_l(p_h^*)-2g_l(\mu_l^*,p_l^*)$ crosses zero at least once. Then we show that the crossing point is unique.

Step 1. When $\Delta\alpha$ is close to zero, that is, $\alpha_l\to\alpha_h$, we have $\hat{g}_l(p_h^*)\to g_h(\mu_h^*,p_h^*)$. Hence $g_h(\mu_h^*,p_h^*)+\hat{g}_l(p_h^*)-2g_l(\mu_l^*,p_l^*)\to 2\left[g_h(\mu_h^*,p_h^*)-g_l(\mu_l^*,p_l^*)\right]>0$. When $\Delta\alpha$ gets very close to α_h, α_l approaches 0, $g_h(\mu_h^*,p_h^*)+\hat{g}_l(p_h^*)-2g_l(\mu_l^*,p_l^*)$ tends to be negative. For example, when $\alpha_l=\frac{1}{100}\alpha_h$,

$$\begin{aligned}g_h(\mu_h^*,p_h^*)+\hat{g}_l(p_h^*)-2g_l(\mu_l^*,p_l^*)=\frac{9}{200\alpha_h\beta}\cdot\big\{&-550Q_c^2-550\pi^2(1-\beta)^2-539\alpha_h^2\mu_c^2-1800\alpha_h\omega\\&-200\pi(1-\beta)\sqrt{\alpha_h\omega}+1100\pi(1-\beta)Q_c+200Q_c\sqrt{\alpha_h\omega}+1960\mu_c\alpha_h\sqrt{\alpha_h\omega}\big\}<0\end{aligned}$$

because $550Q_c^2+550\pi^2(1-\beta)^2+539\alpha_h^2\mu_c^2+1800\alpha_h\omega\geq 1100\pi(1-\beta)Q_c+1989.97Q_c\sqrt{\alpha_h\omega}+1969.97\mu_c\alpha_h\sqrt{\alpha_h\omega}>1100\pi(1-\beta)Q_c+200Q_c\sqrt{\alpha_h\omega}+1960\mu_c\alpha_h\sqrt{\alpha_h\omega}$. Notice that $g_h(\mu_h^*,p_h^*)+\hat{g}_l(p_h^*)-2g_l(\mu_l^*,p_l^*)$ is a continuous function. Applying intermediate value theorem, we can show that there must exist $\Delta\alpha^*\in(0,\alpha_h)$ such that $g_h(\mu_h^*,p_h^*)+\hat{g}_l(p_h^*)-2g_l(\mu_l^*,p_l^*)=0$ at $\mu_l=\mu_h-\Delta\alpha^*$.

Step 2. Now we need to show that, given α_h, $\Delta\alpha^*$ is unique. It is sufficient to prove that $g_h(\mu_h^*,p_h^*)+\hat{g}_l(p_h^*)-2g_l(\mu_l^*,p_l^*)$ is concave in α_l. This is true because its second-order derivative in terms of α_l is

$$\begin{aligned}\frac{1}{4\alpha_l^3\beta}\cdot\big\{&-2[Q_c-(1-\beta)\pi]^2-2\alpha_l\mu_c\sqrt{\alpha_l\omega}-2\alpha_h^2\mu_c^2-8\alpha_h\omega-3[\alpha_h\mu_c+\pi(1-\beta)]\sqrt{\alpha_l\omega}\\&+6\sqrt{\alpha_h\alpha_l}\omega+8\alpha_h\mu_c\sqrt{\alpha_h\omega}+3Q_c\sqrt{\alpha_l\omega}\big\}<0\end{aligned}$$

since the last three positive terms are outweighed by the other negative parts. To see this, we verify the following two inequalities: a) $2[Q_c-(1-\beta)\pi]^2+2\alpha_h^2\mu_c^2+8\alpha_h\omega+2\alpha_h\mu_c\sqrt{\alpha_l\omega}\geq 6\sqrt{\alpha_h\alpha_l}\omega+8\alpha_h\mu_c\sqrt{\alpha_h\omega}$, and b) $2\alpha_l\mu_c\sqrt{\alpha_l\omega}+\alpha_h\mu_c\sqrt{\alpha_l\omega}+3\pi(1-\beta)\sqrt{\alpha_l\omega}\geq 3Q_c\sqrt{\alpha_l\omega}$.

To verify a), we first notice that

$$\begin{aligned}&2[Q_c-(1-\beta)\pi]^2+2\alpha_h^2\mu_c^2+8\alpha_h\omega\geq 4\alpha_h\mu_c[Q_c-(1-\beta)\pi]+8\sqrt{\alpha_h\omega}[Q_c-(1-\beta)\pi]+8\alpha_h\mu_c\sqrt{\alpha_h\omega} &(16)\\&=4\alpha_h\left(\mu_c+2\sqrt{\frac{\omega}{\alpha_h}}\right)[Q_c-(1-\beta)\pi]+8\alpha_h\mu_c\sqrt{\alpha_h\omega}\geq 4\alpha_h\omega+8\alpha_h\mu_c\sqrt{\alpha_h\omega} &(17)\\&\geq 4\sqrt{\alpha_h\alpha_l}\omega+8\alpha_h\mu_c\sqrt{\alpha_h\omega}, &(18)\end{aligned}$$

where (16) is a result of applying the classical inequality that $a+b+c\geq 2\sqrt{ab}+2\sqrt{bc}+2\sqrt{ca}$ for any $a,b,c>0$, (17) follows from Assumption **??**, and (18) is true because $\alpha_h>\alpha_l$. Then we see from Assumption **??** that $2\alpha_h\mu_c\sqrt{\alpha_l\omega}\geq\sqrt{\alpha_h\alpha_l}\omega$, which, combined with (18), proves part a).

Then we verify b) by noticing that

$$2\alpha_l\mu_c\sqrt{\alpha_l\omega}+\alpha_h\mu_c\sqrt{\alpha_l\omega}+3\pi(1-\beta)\sqrt{\alpha_l\omega}\geq 3\alpha_l\mu_c\sqrt{\alpha_l\omega}+3\pi(1-\beta)\sqrt{\alpha_l\omega} \tag{19}$$

$$\geq 3Q_c\sqrt{\alpha_l\omega}, \tag{20}$$

where (19) is true since $\alpha_h>\alpha_l$, and (20) comes from Assumption **??**.

Now that $g_h(\mu_h^*,p_h^*)+\hat{g}_l(p_h^*)-2g_l(\mu_l^*,p_l^*)$ is concave in α_l, there cannot exist other zeros outside $\alpha_l=\alpha_h-\Delta\alpha^*$ and $\alpha_l=\alpha_h$.

Therefore, we have established the single-crossing property from the above two steps. □

Now that we have established Lemmas 2 and 3, we proceed to prove Proposition 8. We first consider Part i). When $\Delta\alpha\geq\Delta\alpha^*$, the single-crossing condition, specified by Lemma 3, gives

$$g_h(\mu_h^*,p_h^*)+\hat{g}_l(p_h^*)\leq 2g_l(\mu_l^*,p_l^*). \tag{21}$$

To understand the condition specified by (21), we note that the left-hand side is the type l physician's payoff if the type l physician mimics the type h physician: the revenue rate in the first period cannot exceed $g_h(\mu_h^*,p_h^*)$, and the revenue rate in the second period is $\hat{g}_l(p_h^*)$. Now that $g_h(\mu_h^*,p_h^*)+\hat{g}_l(p_h^*)<2g_l(\mu_l^*,p_l^*)$, meaning that the type l physician is better off choosing a service fee of p_l^*. The costless separating equilibrium thus sustains. Then we move to Part ii). When $\Delta\alpha<\Delta\alpha^*$, the costless separating equilibrium no longer prevails. If there exists p' that maximizes the type h physician's objective function

$$\text{maximize } p_h\left\{\lambda_1(\mu_{h1}|p_h)+\max_{\mu_{h2}}[\lambda_2(\mu_{h2})]\right\}$$

subject to (15), it would guarantee that: (1) the type l physician does not have the incentive to charge the same price as the type h physician does, and (2) the type h physician is better off charging a higher service fee and prescribing more tests rather than mimicking the type l physician by charging a low service fee. The type h physician's optimal strategy, therefore, is to choose a costly price signal p_h' to deter the type l physician from mimicking her. Accordingly, the type l physician's optimal strategy is to behave as if in the full-information equilibrium. □

Proof of Proposition 9. Instead of proving the nonexistence of a pooling equilibrium, we show that a costly separating equilibrium always prevails whenever a costless separating equilibrium does not.

Firstly, we note that, because $\hat{g}_l(p_l^*)=g_l(\mu_l^*,p_l^*)<\hat{g}_h(p_l^*)$ and both $\hat{g}_h(\cdot)$ and $\hat{g}_l(\cdot)$ are unimodular, there always exist two different roots, denoted by $\underline{p}_1$ and $\bar{p}_1$, respectively, for

$$\hat{g}_h(p)+\hat{g}_l(p)=2g_l(\mu_l^*,p_l^*), \tag{22}$$

Similarly, there always exist two different roots, denoted by $\underline{p}_2$ and $\bar{p}_2$, respectively, for

$$g_l(\mu_l^*,p_l^*)+\hat{g}_h(p_l^*)=2\hat{g}_h(p). \tag{23}$$

Furthermore, we note that $\underline{p}_1<p_l^*<p_h^*<\bar{p}_1$ and $\underline{p}_2<p_l^*<p_h^*<\bar{p}_2$.

Secondly, we show in the following that $\underline{p}_1 < \underline{p}_2 < \bar{p}_1 < \bar{p}_2$. The proof consists of two parts: 1) $\bar{p}_1 < \bar{p}_2$, and 2) $\underline{p}_2 > \underline{p}_1$. We prove part 1) first. Using (22) and (23), we have the following equation:

$$2\hat{g}_h(\bar{p}_2) = \hat{g}_h(\bar{p}_1) + \hat{g}_l(\bar{p}_1) + \hat{g}_h(p_l^*) - \hat{g}_l(p_l^*). \tag{24}$$

Suppose by contradiction that $\bar{p}_1 > \bar{p}_2$. Since $\bar{p}_1 > p_h^*$, $\bar{p}_2 > p_h^*$, we see that $\hat{g}_h(\bar{p}_2) > \hat{g}_h(\bar{p}_1)$, which, together with (24), gives

$$2\hat{g}_h(\bar{p}_2) = \hat{g}_h(\bar{p}_1) + \hat{g}_l(\bar{p}_1) + \hat{g}_h(p_l^*) - \hat{g}_l(p_l^*) > 2\hat{g}_h(\bar{p}_1). \tag{25}$$

Equation (25), after a little bit of algebra, becomes $\hat{g}_h(p_l^*) - \hat{g}_l(p_l^*) > \hat{g}_h(\bar{p}_1) - \hat{g}_h(\bar{p}_1)$, which is a contradiction by Lemma 2. Then we examine part 2). Equations (22) and (23), after some algebra, jointly give

$$2\hat{g}_h(\underline{p}_2) = \hat{g}_h(\underline{p}_1) + \hat{g}_l(\underline{p}_1) + \hat{g}_h(p_l^*) - \hat{g}_l(p_l^*). \tag{26}$$

Suppose by contradiction that $\underline{p}_1 > \underline{p}_2$. By noticing that $\underline{p}_1 < p_l^*$, $\underline{p}_2 < p_l^*$, we have $\hat{g}_h(\underline{p}_1) > \hat{g}_h(\bar{p}_2)$, which, together with (26), gives

$$2\hat{g}_h(\underline{p}_2) = \hat{g}_h(\underline{p}_1) + \hat{g}_l(\underline{p}_1) + \hat{g}_h(p_l^*) - \hat{g}_l(p_l^*) < 2\hat{g}_h(\underline{p}_1). \tag{27}$$

Equation (27), after a little bit of algebra, becomes $\hat{g}_h(p_l^*) - \hat{g}_l(p_l^*) < \hat{g}_h(\underline{p}_1) - \hat{g}_h(\underline{p}_1)$, which is a contradiction by Lemma 2.

Thirdly, we recognize that the set of service fees that satisfies both of the two conditions in Proposition 7 is $\left\{p : p \le \underline{p}_1 \text{ or } p \ge \bar{p}_1\right\} \cap \left\{p : \underline{p}_2 \le p \le \bar{p}_2\right\} = \{p : \bar{p}_1 \le p \le \bar{p}_2\} \neq \emptyset$ since $\underline{p}_1 < \underline{p}_2 < \bar{p}_1 < \bar{p}_2$. We conclude from Proposition 8 that a costly separating equilibrium always prevails whenever a costless equilibrium does not exist. □

Appendix B: Relaxation of Major Assumptions

Our analysis so far has been based on a few assumptions regarding (1) the distribution of the service time, (2) the definition of the waiting time, (3) the relationship between the service rate and the service quality, and (4) the misdiagnosis cost, which were made to ensure the conciseness of our analysis. In this section, we relax these assumptions to explore the boundaries of our models. Furthermore, we consider two cases that involve more sophisticated service settings: (1) the physician is effort-averse, and (2) the physician's service fee depends on the actual service time.

B.1 An Alternative Definition of Waiting Time

We have thus far used the total time in system as the definition of the waiting time since the benefits from the service have been captured by the function $Q(\mu)$. To demonstrate the robustness of our model, we consider the case in which the waiting time is defined as the time in the queue only. Below we briefly discuss the analytical results for the market equilibrium, the social optimum, and the case considering misdiagnosis concerns.

Market Equilibrium. In the baseline model, if we replace $\mathbb{E}[W(\mu,\lambda)] = \frac{1}{\mu-\lambda}$ (time in the system) with $\mathbb{E}[W'(\mu,\lambda)] = \frac{1}{\mu-\lambda} - \frac{1}{\mu}$ (queueing time), then solving $Q(\mu) - \pi - \beta(p-\pi) - \omega\mathbb{E}[W'(\mu,\lambda)] = 0$ gives the

equilibrium arrival rate $\lambda(\mu,p)=\mu-\mu\omega/\{\omega-\mu[Q(\mu)-\pi-\beta(p-\pi)]\}$. Following similar procedures as in the proof of Proposition 1, we can show that the physician's revenue rate $p\lambda(\mu,p)$ is concave in p. Letting $p^*(\mu)$ denote the optimal service fee given the service rate μ, we can also show that $p^*(\mu)\cdot\lambda(\mu,p^*(\mu))$ is unimodal in μ. Consequently, we obtain the optimal service rate and service fee as follows:

$$\mu^*=\frac{Q_c+\alpha\mu_c-(1-\beta)\pi}{2\alpha} \text{ and } p^*=\frac{C_q}{2\beta}+\frac{2\alpha\omega}{\beta C_q}-\frac{1}{\beta}\sqrt{\alpha\omega\left(1+\frac{4\omega\alpha}{C_q^2}\right)}.$$

Thus the induced equilibrium arrival rate is $\lambda(\mu^*,p^{ast})=\mu^*-C_q\sqrt{\frac{\omega}{\alpha(C_q^2+4\alpha\omega)}}$, where $C_q:=Q_c+\alpha_c-(1-\beta)\pi$.

The above result shows that changing the definition of the waiting time only affects the service fee but does not affect the service rate. The intuition is as follows. Since the service queue is unobservable to patients and the physician adopts a static FCFS queueing policy, the primary concern in determining the optimal service rate is the generation of the service value. Patients' waiting costs are also an important factor in determining the service parameters but are secondary in that they are a result of the chosen service rate. Therefore, in weighing the queueing process, the physician adjusts the service fee but not the service rate.

Social Optimum. The social planner chooses μ and λ to maximize the social welfare that can be represented as $SW(\mu,\lambda)=\lambda\cdot\{Q(\mu)-\omega\mathbb{E}[W'(\mu,\lambda)]\}$. Following a similar method to the proof of Proposition 2, we obtain the socially efficient service rate and arrival rate as follows:

$$\mu^{SE}=\frac{Q_c+\alpha\mu_c}{2\alpha} \text{ and } \lambda^{SE}=\mu^{SE}-(Q_c+\alpha\mu_c)\sqrt{\frac{\omega}{\alpha[(Q_c+\alpha\mu_c)^2+4\alpha\omega]}}.$$

Considering Misdiagnosis Concerns. The procedure is similar to those shown above. The optimal service rate remains the same, that is, $\mu^*=[Q_c+\alpha\mu_c-(1-\beta)\pi]/[2(\alpha+\beta d)]$. The social efficient service level also remains unaffected, that is, $\mu^{SE}=[Q_c+\alpha\mu_c]/[2(\alpha+d)]$.

B.2 Distribution of Service Time

So far we have assumed that the service time follows an exponential distribution. We now extend to a general service time distribution with a squared coefficient of variation of C_S.[5] Given a service rate of μ and an arrival rate λ, each patient's average waiting time follows from the Pollaczek-Khinchine formula (Kleinrock 1975):

$$\mathbb{E}[W(\mu,\lambda)]=\frac{1}{\mu}\left[\frac{\lambda}{\mu-\lambda}\cdot\frac{C_S+1}{2}+1\right].$$

Solving the physician's problem with the new formulation of the waiting time using a similar procedure to the proof of Proposition 1 gives $\mu^*=[Q_c+\alpha\mu_c-(1-\beta)\pi]/(2\alpha)$, which is exactly the same as in the $M/M/1$ case. The squared coefficient of variation C_s is reflected in the service fee only.

Using the new formulation of the waiting time, we can also verify that the optimal service rates in the social optimum and with misdiagnosis concerns, respectively, remain the same as in the $M/M/1$ case.

[5] $C=1$ corresponds to the special case that the service time is exponentially distributed.

B.3 A General, Non-Monotonic Service Quality Function

We assume in the baseline model that the service quality $Q(\mu)$ is an affine function of μ for simplicity of analysis. Now we extend $Q(\mu)$ to a general function form and allow it to be non-monotonic, that is, excessive testing leads to decrease in the quality of service. To be more specific, there exists $\hat{\mu}$ such that $Q(\mu)$ increases in μ, that is, $Q'(\mu) > 0$, if $\mu < \hat{\mu}$, and $Q(\mu)$ decreases in μ, that is, $Q'(\mu) \leq 0$, otherwise. Here $\hat{\mu}$ is a threshold at which the potential harm due to excessive testing exactly offsets the marginal benefit of diagnostic tests. We maintain the concavity of $Q(\mu)$, that is, $Q''(\mu) < 0$, which makes sense for both of the following two cases: when $\mu \geq \hat{\mu}$, the concavity reflects the diminishing marginal return from diagnostic tests; when $\mu < \hat{\mu}$, the concavity reflects the increasing marginal damage caused by additional tests, as more tests expose patients to even more unnecessary tests and procedures.

Market Equilibrium. To characterize the market equilibrium, we first present the following lemma:

LEMMA 4. *Under the full-information setting, the physician never sets the service rate μ below $\hat{\mu}$.*

Proof. We arbitrarily choose μ_1 such that $\mu_c < \mu_1 < \hat{\mu}$, and a service fee p_1. It is sufficient to show that, if the physician chooses a service parameter combination (μ_1, p_1), then the physician is always better off switching her service rate from μ_1 to $\hat{\mu}$. The reason is as follows: since the physician charges the same service fee, we only need to compare the equilibrium arrival rates under two different service parameters, that is, to compare $\lambda_1 := \lambda(\mu_1, p_1)$ with $\hat{\lambda} := \lambda(\hat{\mu}, p_1)$. Clearly,

$$\lambda_1 = \mu_1 - \frac{\omega}{Q(\mu_1) - \pi - \beta(p_1 - \pi)} < \hat{\lambda} = \hat{\mu} - \frac{\omega}{Q(\hat{\mu}) - \pi - \beta(p_1 - \pi)}$$

since $\mu_1 < \hat{\mu}$ and $Q(\mu_1) < Q(\hat{\mu})$. □

The intuition of the above lemma is that, under full information about service value, the physician has no incentive to order more tests than the threshold level; otherwise, the perceived service value reduction will lead to reduction in the equilibrium arrival rate.

PROPOSITION 10. *The optimal service rate $\mu^* = \max\{\bar{\mu}^*, \hat{\mu}\}$, where $\bar{\mu}^*$ satisfies*

$$Q(\mu) + \mu \cdot Q'(\mu) - \pi(1 - \beta) = 0 \text{ at } \mu = \bar{\mu}^*. \tag{28}$$

Proof sketch. We use the following procedure similar to the proof of Proposition 1 in the paper to derive the optimal service rate assuming that $\mu < \hat{\mu}$. First, we show that the physician's revenue rate function is concave in the service fee p, and obtain the expression of the optimal service fee conditioning on μ. Then we substitute the condition expression into the objective function, which is verified to be unimodal in μ. Finally we incorporate Lemma 4 and complete the proof. □

Next, we show that, if $\mu^* < \hat{\mu}$, then as in the baseline model, μ^* decreases in π but increases in β. To see why this is true, we write the left-hand side of (28) as $\phi(\mu)$, which decreases in μ as its first-order derivative in terms of μ is nonpositive: $d\phi(\mu)/d\mu = 2Q'(\mu) + \mu Q''(\mu) \leq 0$. As π increases or β increases, $\phi(\mu)$ increases, justifying a lower μ to keep (28) balanced.

Social Optimum. The social planner's problem consists of solving

$$\max_{\mu,\lambda} \lambda \left[Q(\mu) - \frac{\omega}{\lambda - \mu} \right].$$

The following proposition characterizes the socially efficient service rate:

PROPOSITION 11. *The socially efficient service level* $\mu^{SE} = \max\{\bar{\mu}^{SE}, \hat{\mu}\}$*, where* $\bar{\mu}^{SE}$ *satisfies*

$$Q(\mu) + \mu \cdot Q'(\mu) = 0 \text{ at } \mu = \bar{\mu}^{SE}. \tag{29}$$

Proof. Similar to the proof of Proposition 10. □

Now we examine the interesting case that both μ^* and μ^{SE} are above $\hat{\mu}$. Since $\phi(\mu)$ decreases in μ, we conclude from (28) and (29) that $\mu^{SE} \geq \mu^*$, and the equality is true only when $\beta = 1$ or $\pi = 0$.

So far we have shown that our main insights extend to the case considering the false positive effect as long as we keep the full-information assumption in the baseline model. When patients possess biased information about the service value of diagnostic tests, however, our analysis in the section might no longer be valid, and the cases that $\mu^* < \hat{\mu}$ can indeed occur. We leave this as a direction for future research.

B.4 Misdiagnosis Cost Function

Finally, we consider a general misdiagnosis cost function $\theta(\mu)$, which is convex increasing in μ, that is, $\theta'(\mu) > 0$ and $\theta''(\mu) > 0$. This assumption is realistic in that, when $\mu = \mu_c$, the physician orders no diagnostic tests and is most concerned about reaching inaccurate diagnosis; when μ approaches zero, however, the physician chooses an extremely detailed diagnosis process, the misdiagnosis cost gets close to zero but there is virtually no room for reducing the misdiagnosis cost. Under this condition, the concavity and unimodality properties of the physician's objective function hold. We can show that the optimal service rate μ^* satisfies the following optimality equation:

$$Q_c + \alpha_c \mu_c - (1-\beta)\pi - \mu[2\alpha_c + \beta \cdot \theta'(\mu)] - \beta \cdot \theta(\mu) = 0 \text{ at } \mu = \mu^*. \tag{30}$$

We write the left-hand side of (30) as $\bar{\phi}(\mu)$, which decreases in μ since $\partial\bar{\phi}(\mu)/\partial\mu = -2\beta \cdot \theta'(\mu) - \mu\beta \cdot \theta''(\mu) - 2\alpha_c < 0$. Therefore, we can prove that μ^* decreases in π by noticing that $\bar{\phi}$ decreases in π, meaning that a higher π needs to be compensated by a lower μ^* to keep that $\bar{\phi}(\mu) = 0$ at $\mu = \mu^*$. We can also prove that μ^* can either increase or decrease in β depending on the range of the copayment π: if $\pi > \mu^* \cdot \theta'(\mu^*) + \theta(\mu^*)$, then μ^* increases in β; otherwise, μ^* decreases in β.

Both Misdiagnosis-cost and Service Quality Functions are in General Forms. If $Q(\mu)$ is in general form, however, the optimal service rate μ^* satisfies $Q(\mu) + \mu \cdot Q'(\mu) - (1-\beta)\pi - \beta\mu \cdot \theta'(\mu) - \beta \cdot \theta(\mu) = 0$ at $\mu = \mu^*$ if $Q(\mu) = Q_c + \alpha_c(\mu_c - \mu)$. Defining $\tilde{\phi}(\mu) := Q(\mu) + \mu \cdot Q'(\mu) - (1-\beta)\pi - \beta\mu \cdot \theta'(\mu) - \beta \cdot \theta(\mu)$, we prove that $\tilde{\phi}(\mu)$ decreases in μ by showing that its first-order derivative $\partial\tilde{\phi}(\mu)/\partial\mu = 2Q'(\mu) + \mu \cdot Q''(\mu) - 2\beta \cdot \theta'(\mu) - \mu\beta \cdot \theta''(\mu) < 0$. It is not difficult to observe that μ^* decreases in π by noticing that $\tilde{\phi}$ decreases in π, as a higher π needs to be compensated by a lower μ^* to keep that $\bar{\phi}(\mu) = 0$ at $\mu = \mu^*$. We can also show that whether μ^* increases or decreases in β depends on the range of the copayment π: if $\pi > \mu^* \cdot \theta'(\mu^*) + \theta(\mu^*)$, then μ^* increases in β; otherwise, μ^* decreases in β.

B.5 Follow-up Visits

In the baseline model, we assume that at any time every patient with medical needs adopts the same randomized queue-joining strategy; as a consequence, a patient can make multiple visits over time, but different visits are considered to be independent of each other. In practice, however, a proportion of patients making initial visits are advised by the physician to make follow-up visits, which involve consultations, repetitive or new tests to collect additional information. In this section, we address this important issue under our analytical framework, and show that our main insights from the baseline model remain directionally unchanged.

Assume that with probability γ (referred to as "follow-up rate"), a patient making initial visits is determined by the physician to need to make a follow-up visit. According to our experience at UPMC Eye Center, follow-up visits are typically more brief and less expensive compared to initial visits. In addition, returning patients are usually more waiting-time-sensitive than new patients. We assume that patients are strategic and do not always comply with the physician's advice; as a result, only a proportion of patients join the queue for suggested follow-up visits. This proportion is referred to as "compliance rate" and denoted by θ. Here we highlight the key difference between the follow-up rate γ and the compliance rate θ: while the follow-up rate γ is determined by this physician's medical judgment, the compliance rate θ is driven by patients' willingness to make return visits.

We model the service system with both new and returning patients as a classed Jackson network. To be more specific, there are two classes of patients, denoted by n (new) and r (returning), respectively. The physician orders different numbers of tests that are dictated by service rates $\mu_i, i=n,r$, and charges them different service fees $p_i, i=n,r$. The two classes of patients have different unit waiting costs, denoted by $\omega_i, i=n,r$, and our analysis applies to both of the following two cases: (1) $\omega_r > \omega_n$ and (2) $\omega_r \leq \omega_n$. The two classes of patients jointly choose their queue-joining probabilities to maximize their expected utility. The resultant equilibrium arrival rates are denoted by $\lambda_i(p_n,p_r,\mu_n,\mu_r), i=n,r$. We have the following constraints that have to be satisfied:

$$\begin{aligned} &Q(\mu_n)-\pi-\beta(p_n-\pi)-\omega_n\mathbb{E}[W(\mu_n,\mu_r,\lambda_n(p_n,p_r,\mu_n,\mu_r),\lambda_r(p_n,p_r,\mu_n,\mu_r))] \\ &\quad+\theta\gamma\cdot\{Q(\mu_r)-\pi-\beta(p_r-\pi)-\omega_r\mathbb{E}[W(\mu_n,\mu_r,\lambda_n(p_n,p_r,\mu_n,\mu_r),\lambda_r(p_n,p_r,\mu_n,\mu_r))]\}\geq 0, \qquad (31) \\ &Q(\mu_r)-\pi-\beta(p_r-\pi)-\omega_r\mathbb{E}[W(\mu_n,\mu_r,\lambda_n(p_n,p_r,\mu_n,\mu_r),\lambda_r(p_n,p_r,\mu_n,\mu_r))]\geq 0, \qquad (32) \\ &\lambda_r(p_n,p_r,\mu_n,\mu_r)-\theta\cdot\gamma\cdot\lambda_n(p_n,p_r,\mu_n,\mu_r)=0, \qquad (33) \end{aligned}$$

where $\mathbb{E}[W(\mu_n,\mu_r,\lambda_n(p_n,p_r,\mu_n,\mu_r),\lambda_r(p_n,p_r,\mu_n,\mu_r))]$ is the expected waiting time of the classed Jackson network with two classes of patients when the service rates are μ_n and μ_r, and the equilibrium arrival rates are $\lambda_n(p_n,p_r,\mu_n,\mu_r)$ and $\lambda_r(p_n,p_r,\mu_n,\mu_r)$), respectively. (31) is a participation constraint ensuring that each new patient's expected total net surplus from the initial visit and a potential return visit is nonnegative. (32) is another participation constraint ensuring that each returning patient's expected net surplus is nonnegative; this constraint can be eliminated in certain cases, as we will show later on. (33) relates the arrival rates of

new and returning patients; it means that only a proportion (θ) of γ of the new patients will make returning visits.

The physician's problem is now to solve the following revenue-rate-maximization problem:

$$\begin{aligned} \max_{p_n,p_r,\mu_n,\mu_r} \quad & p_n \cdot \lambda_n(p_n,p_r,\mu_n,\mu_r) + p_r \cdot \lambda_r(p_n,p_r,\mu_n,\mu_r) \\ \text{subject to} \quad & (31)\text{–}(33). \end{aligned}$$

Below we analyze the physician's problem by examining two scenarios where the compliance rate θ is exogenous and endogenous, respectively.

Exogenous Compliance Rate. If the compliance rate θ is exogenous, then the constraint (32) becomes a redundant one, and (31) must be tight. In addition, we can eliminate (33) and substitute it into the physician's objective function. The physician's problem can thus be rewritten as

$$\begin{aligned} \max_{p_n,p_r,\mu_n,\mu_r} \quad & (p_n+\theta\gamma p_r)\cdot\lambda_n(p_n,p_r,\mu_n,\mu_r) \\ \text{subject to} \quad & [Q(\mu_n)+\theta\gamma Q(\mu_r)] - (1-\beta)(1+\theta\gamma)\pi - \beta(p_n+\theta\gamma p_r) \\ & - (\omega_n+\theta\gamma\theta_r\omega_r)\mathbb{E}[W(\mu_n,\mu_r,\lambda_n(p_n,p_r,\mu_n,\mu_r),\theta\gamma\cdot\lambda_n(p_n,p_r,\mu_n,\mu_r))] = 0, \end{aligned}$$

which can be further rewritten as

$$\begin{aligned} \max_{\hat{p},\mu_n,\mu_r} \quad & \hat{p}\cdot\hat{\lambda}(\hat{p},\mu_n,\mu_r) \\ \text{subject to} \quad & [\hat{Q}(\mu_n,\mu_r)] - (1-\beta)\hat{\pi} - \beta\hat{p} - \hat{\omega}\mathbb{E}[\hat{W}(\mu_n,\mu_r,\hat{\lambda}(\hat{p},\mu_n,\mu_r))] = 0, \end{aligned}$$

by defining

$$\begin{aligned} \hat{p} &= p_n+\theta\gamma p_r, \\ \hat{\lambda}(\hat{p},\mu_n,\mu_r) &= \lambda_n(p_n,p_r,\mu_n,\mu_r), \\ \hat{Q}(\mu_n,\mu_r) &= Q(\mu_n)+\theta\gamma Q(\mu_r), \\ \hat{\pi} &= (1+\theta\gamma)\pi, \\ \hat{\omega} &= \omega_n+\theta\gamma\omega_r, \\ \hat{W}(\mu_n,\mu_r,\hat{\lambda}(\hat{p},\mu_n,\mu_r)) &= W(\mu_n,\mu_r,\lambda_n(p_n,p_r,\mu_n,\mu_r),\theta\gamma\cdot\lambda_n(p_n,p_r,\mu_n,\mu_r)). \end{aligned}$$

Hence we see that when θ is exogenous, the physician has three decision variables, namely, $\hat{p}, \mu_n$, and μ_r. The problem can be analytically solved (see the end of this section for details). For ease of comparison, consider a special case where $\mu_n = \mu_r$, then the physician's optimal decision can be characterized by the following proposition:

PROPOSITION 12. *When considering the effect of follow-up visits, with exogenous θ and the additional constraint that $\mu_n = \mu_r$, there exists a unique equilibrium as follows.*

i) The physician chooses the service rate $\mu_n^* = \mu_r^* = \frac{Q_c+\alpha\mu_c-(1-\beta)\pi}{2\alpha}$, and an arbitrary combination of service fees (p_n^*, p_r^*) that satisfies $p_n+\theta\gamma p_r = \frac{1+\gamma\theta}{2\beta}\left[Q_c+\alpha\mu_c-(1-\beta)\pi\right] - \frac{\sqrt{(1+\theta\gamma)\alpha(\omega_n+\gamma\omega_r)}}{\beta}$.

ii) The induced arrival rate for initial visits is $\lambda_n(\mu^*,p^*) = \frac{1}{1+\theta\gamma}\cdot\left[\frac{Q_c+\alpha\mu_c-(1-\beta)\pi}{2\alpha} - \sqrt{\frac{\omega_n+\gamma\omega_r}{(1+\theta\gamma)\alpha}}\right]$, and that for follow-up visits is $\lambda_r(\mu^*,p^*) = \frac{\theta\gamma}{1+\theta\gamma}\cdot\left[\frac{Q_c+\alpha\mu_c-(1-\beta)\pi}{2\alpha} - \sqrt{\frac{\omega_n+\gamma\omega_r}{(1+\theta\gamma)\alpha}}\right]$.

iii) The expected waiting time is $\mathbb{E}[W(\mu^*, \lambda(\mu^*, p^*))] = \sqrt{\frac{(1+\theta\gamma)\alpha}{\omega_n+\gamma\omega_r}}$.

Proposition 12, admittedly based on a set of restrictive assumptions, has several interesting implications. First, comparing Proposition 12 with Proposition 1, we see that the physician chooses exactly the same service rate as in the baseline model. If, however, we drop the constraint that $\mu_n = \mu_r$, then the results will deviate from Proposition 1, but one would expect that $\mu_n < [Q_c + \alpha\mu_c - (1-\beta)\pi]/(2\alpha) < \mu_r$; in other words, the physician will still choose her service rates for new and returning patients around the optimal service rate in the baseline model, and our analysis about the impact of insurance and potentially other service environments remains directionally valid. Second, the physician enjoys more freedom to choose the service fees for new and returning visits; in some cases, if the physician sets p_r^* much higher than p_n^*, $(\theta\gamma)/(1+\theta\gamma)$ of patients who make both initial and follow-up visits would pay much more than others. Third, the arrival rate of new patients is decreasing in θ, meaning that a higher expectation of followup visits reduces patients' desire to choose a specific clinic and hence increases their probability of resorting to outside options.

Endogenous Compliance Ratio. When the compliance ratio θ is endogenous, the optimization problem can be solved by utilizing the product-form property of the Jackson classed network (Harchol-Balter 2011). Furthermore, the solution procedure can be significantly streamlined by the analysis below.

We first consider the case that the physician chooses service parameters that lead to $\theta = 1$, meaning that every new patient will comply with the physician's advice regarding returning visits. In this case, it is evident that (32) is not binding, that is,

$$Q(\mu_r) - \pi - \beta(p_r - \pi) - \omega_r \mathbb{E}[W(\mu_n, \mu_r, \lambda_n(p_n, p_r, \mu_n, \mu_r), \lambda_r(p_n, p_r, \mu_n, \mu_r))] > 0. \tag{34}$$

(31), however, must be binding due to the large potential demand, which means

$$Q(\mu_n) - \pi - \beta(p_n - \pi) - \omega_n \mathbb{E}[W(\mu_n, \mu_r, \lambda_n(p_n, p_r, \mu_n, \mu_r), \lambda_r(p_n, p_r, \mu_n, \mu_r))] < 0. \tag{35}$$

The two conditions (34) and (35) jointly give the following observation that provides an interesting insight on the inducement of patients' compliance behavior:

Observation 3. *In order to induce all the patients to comply with the physician's advice on follow-up visits, the physician must set service parameters such that each patients experiences a negative expected net surplus during his initial visit, but then a positive expected net surplus during his returning visit (if any).*

The following lemma states a necessary and sufficient condition for both (34) and (35) to be true.

Lemma 5. *The induced compliance rate* $\theta = 1$ *if and only if*

$$\frac{Q(\mu_r) - \beta p_r - (1-\beta)\pi}{Q(\mu_n) - \beta p_n - (1-\beta)\pi} > \frac{\omega_r}{\omega_n}, \tag{36}$$

in which case the physician's problem can be solved as if θ *is exogenous.*

If (36) is violated, then the induced compliance rate $\theta < 1$, that is, not all of the patients comply with the physician's advice. In this case, the constraint (32) is tight, and the service constraint that the physician faces can be rewritten as

$$Q(\mu_n) - \pi - \beta(p_n - \pi) - \omega_n \mathbb{E}[W(\mu_n, \mu_r, \lambda_n(p_n, p_r, \mu_n, \mu_r), \lambda_r(p_n, p_r, \mu_n, \mu_r))] = 0, \tag{37}$$

$$Q(\mu_r) - \pi - \beta(p_r - \pi) - \omega_r \mathbb{E}[W(\mu_n, \mu_r, \lambda_n(p_n, p_r, \mu_n, \mu_r), \lambda_r(p_n, p_r, \mu_n, \mu_r))] = 0, \tag{38}$$

Jointly solving (37) and (38) gives the following equation:

$$\frac{Q(\mu_r) - \beta p_r - (1-\beta)\pi}{Q(\mu_n) - \beta p_n - (1-\beta)\pi} = \frac{\omega_r}{\omega_n},$$

which, together with Lemma 5, leads to the following proposition:

PROPOSITION 13. *The physician will always choose* (μ_n, μ_r, p_n, p_r) *such that*

$$\frac{Q(\mu_r) - \beta p_r - (1-\beta)\pi}{Q(\mu_n) - \beta p_n - (1-\beta)\pi} \geq \frac{\omega_r}{\omega_n}, . \tag{39}$$

There are two cases depending on whether (39) *is binding or not:*

a) when (39) *is non-binding, patients will invariably comply with the physician's advices regarding follow-up visits, that is,* $\theta = 1$*;*

b) when (39) *is binding, only* $\theta < 1$ *of new patients will comply with the physician's advices regarding follow-up visits.*

In both of the above cases, the procedure for deriving the optimal service parameters is also similar to the case where θ *is exogenous. As a special case, if the physician restricts that* $\mu_n = \mu_r$*, then it is optimal for the physician to induce a compliance rate of either 100% (i.e.,* $\theta = 1$*) or 0, and, in the former case, in equilibrium,*

i) the physician chooses the service rates $\mu_n^* = \mu_r^* = \frac{Q_c + \alpha\mu_c - (1-\beta)\pi}{2\alpha}$*, and a combination of service fees* (p_n^*, p_r^*) *that satisfies* $p_n^* + \gamma p_r^* = \frac{1+\gamma}{2\beta}\left[Q_c + \alpha\mu_c - (1-\beta)\pi\right] - \frac{1}{\beta}\sqrt{(1+\gamma)\alpha(\omega_n + \gamma\omega_r)}$*, and* $p_n > \frac{Q_c + \alpha\mu_c - (1-\beta)\pi}{2\beta} - \frac{\omega_n}{\beta}\sqrt{\frac{(1+\gamma)\alpha}{\omega_n + \gamma\omega_r}}$*;*

ii) the equilibrium arrival rates are $\lambda_n(p_n^*, p_r^*, \mu^*, \mu^r) = \frac{1}{1+\gamma}\left[\frac{Q_c + \alpha\mu_c - (1-\beta)\pi}{2\alpha} - \sqrt{\frac{\omega_n + \gamma\omega_r}{(1+\gamma)\alpha}}\right]$*, and* $\lambda_r(p_n^*, p_r^*, \mu^*, \mu^r) = \frac{\gamma}{1+\gamma}\left[\frac{Q_c + \alpha\mu_c - (1-\beta)\pi}{2\alpha} - \sqrt{\frac{\omega_n + \gamma\omega_r}{(1+\gamma)\alpha}}\right]$*;*

iii) the expected waiting time is $\sqrt{\frac{(1+\gamma)\alpha}{\omega_n + \gamma\omega_r}}$*.*

Proof sketch. Depending on whether the induced compliance rate is 100% or not, there are two cases to consider:

Case a) When (39) is non-binding, patients will invariably follow the physician's advices regarding returning visits, that is, $\theta = 1$. If the physician restricts that $\mu_n = \mu_r$, then by adopting a procedure similar to the proof of Proposition 1, the equilibrium is characterized as follows:

i) the physician chooses the service rates $\mu_n^* = \mu_r^* = \frac{Q_c + \alpha\mu_c - (1-\beta)\pi}{2\alpha}$, and a combination of service fees (p_n^*, p_r^*) that satisfies $p_n^* + \gamma p_r^* = \frac{1+\gamma}{2\beta}\left[Q_c + \alpha\mu_c - (1-\beta)\pi\right] - \frac{1}{\beta}\sqrt{(1+\gamma)\alpha(\omega_n + \gamma\omega_r)}$, and $p_n > \frac{Q_c + \alpha\mu_c - (1-\beta)\pi}{2\beta} - \frac{\omega_n}{\beta}\sqrt{\frac{(1+\gamma)\alpha}{\omega_n + \gamma\omega_r}}$;

ii) the equilibrium arrival rates are $\lambda_n(p_n^*, p_r^*, \mu^*, \mu^r) = \frac{1}{1+\gamma}\left[\frac{Q_c + \alpha\mu_c - (1-\beta)\pi}{2\alpha} - \sqrt{\frac{\omega_n + \gamma\omega_r}{(1+\gamma)\alpha}}\right]$, and $\lambda_r(p_n^*, p_r^*, \mu^*, \mu^r) = \frac{\gamma}{1+\gamma}\left[\frac{Q_c + \alpha\mu_c - (1-\beta)\pi}{2\alpha} - \sqrt{\frac{\omega_n + \gamma\omega_r}{(1+\gamma)\alpha}}\right]$;

iii) the expected waiting time is $\sqrt{\frac{(1+\gamma)\alpha}{\omega_n + \gamma\omega_r}}$.

Case b) When (39) is binding, the induced compliance rate $\theta < 1$. If the physician restricts that $\mu_n = \mu_r$, then by adopting a procedure similar to the proof of Proposition 1, the equilibrium is characterized as follows:

i) the physician chooses the service rates $\mu_n^* = \mu_r^* = \frac{Q_c+\alpha\mu_c-(1-\beta)\pi}{2\alpha}$, and the service fees $p_n = \frac{Q_c+\alpha\mu_c-(1-\beta)\pi}{2\beta} - \frac{\omega_n}{\beta}\sqrt{\frac{(1+\theta^*\gamma)\alpha}{\omega_n+\theta^*\gamma\omega_r}}$, and $p_r = \frac{Q_c+\alpha\mu_c-(1-\beta)\pi}{2\beta} - \frac{\omega_r}{\beta}\sqrt{\frac{(1+\theta^*\gamma)\alpha}{\omega_n+\theta^*\gamma\omega_r}}$;

ii) the equilibrium arrival rates are $\lambda_n(p_n^*, p_r^*, \mu^*, \mu^r) = \frac{1}{1+\gamma}\left[\frac{Q_c+\alpha\mu_c-(1-\beta)\pi}{2\alpha} - \sqrt{\frac{\omega_n+\theta^*\gamma\omega_r}{(1+\theta^*\gamma)\alpha}}\right]$, and $\lambda_r(p_n^*, p_r^*, \mu^*, \mu^r) = \frac{\gamma}{1+\gamma}\left[\frac{Q_c+\alpha\mu_c-(1-\beta)\pi}{2\alpha} - \sqrt{\frac{\omega_n+\theta^*\gamma\omega_r}{(1+\theta^*\gamma)\alpha}}\right]$;

iii) the expected waiting time is $\sqrt{\frac{(1+\theta^*\gamma)\alpha}{\omega_n+\theta^*\gamma\omega_r}}$.

In the above solution, the induced compliance rate θ^* is the maximizer of

$$\frac{1}{4\alpha\beta}\left[Q_c+\alpha\mu_c-(1-\beta)\pi-2\sqrt{\frac{\alpha(w_n+\theta\gamma w_r)}{1+\theta\gamma}}\right]^2. \tag{40}$$

Note that in Case b), (40) is monotonue in θ, meaning that $\theta^* = 1-\epsilon$ or $\theta^* = 0$, where ϵ is infinitesimally close to zero. In the former case, it is straightforward to see that the solution is always dominated by the solution in Case a), meaning that it is optimal for the physician to induce a compliance rate of 100%. Therefore, the optimal service parameters can be determined as in Case a). □

Proposition 13 provides a simplifying condition for determining the optimal service parameters and characterizing the equilibrium. It says that, upon a careful examination of the service constraints, the solution procedure for the case where θ is endogenous turns out to resemble the case where θ is exogenous. Furthermore, under the restriction that $\mu_n = \mu_r$, the optimal service rates for both new and returning patients will be exactly the same as the optimal service rate in the baseline model (cf. Proposition 1). If, however, the restriction that $\mu_n = \mu_r$ is eliminated, one can still expect the optimal service rates to be around the optimal service rate in the baseline model. We therefore conclude that our analytical framework in the baseline model remains valid, and many of our managerial insights remain directionally unchanged.

Social Optimum. Next, we consider the social planner's problem, that is, to choose the socially efficient arrival rates $\lambda_i^{SE}, i = n, r$, and service rates $\mu_i^{SE}, i = n, r$ to maximize the social welfare rate:

$$SW^f(\mu_n, \mu_r, \lambda_n, \lambda_r) = \lambda_n \cdot \{Q(\mu_n) - \omega_n \mathbb{E}[W(\mu_n, \mu_r, \lambda_n, \lambda_r)]\} + \lambda_r \cdot \{Q(\mu_r) - \omega_r \mathbb{E}[W(\mu_n, \mu_r, \lambda_n, \lambda_r)]\}.$$

We characterize the social optimum in the following proposition when adding to the constraint that $\mu_n = \mu_r$.

Proposition 14. *Under the additional constraint that $\mu_n = \mu_r$, in the social optimum,*

i) the service rates are $\mu_n^{SE} = \mu_r^{SE} = \frac{Q_c+\alpha\mu_c}{2\alpha}$;

ii) the arrival rates are $\lambda_n^{SE} = \frac{1}{1+\theta\gamma}\left[\frac{Q_c+\alpha\mu_c}{2\alpha} - \sqrt{\frac{\omega_n+\theta\gamma\omega_r}{(1+\theta\gamma)\alpha}}\right]$, and $\lambda_r^{SE} = \frac{\theta\gamma}{1+\theta\gamma}\left[\frac{Q_c+\alpha\mu_c}{2\alpha} - \sqrt{\frac{\omega_n+\theta\gamma\omega_r}{(1+\gamma)\alpha}}\right]$.

Proof sketch. Given the compliance rate θ, the solution procedure to the problem is similar to the proof of Proposition 2. We obtain that the service rates are $\mu_n^{SE} = \mu_r^{SE}(\theta) = \frac{Q_c+\alpha\mu_c}{2\alpha}$, and the arrival rates are $\lambda_n^{SE} = \frac{1}{1+\theta\gamma}\left[\frac{Q_c+\alpha\mu_c}{2\alpha} - \sqrt{\frac{\omega_n+\theta\gamma\omega_r}{(1+\theta\gamma)\alpha}}\right]$, and $\lambda_r^{SE} = \frac{\theta\gamma}{1+\theta\gamma}\left[\frac{Q_c+\alpha\mu_c}{2\alpha} - \sqrt{\frac{\omega_n+\theta\gamma\omega_r}{(1+\theta\gamma)\alpha}}\right]$. □

Comparing the above proposition with Proposition 2, we see that the socially efficient service rates are exactly the same, indicating that the explanations for overtesting remain valid. When the constraint that $\mu_r = \mu_n$ is removed, however, we expect that the solution will be different, but our key managerial insights are directionally unchanged.

Removing Equal-Service-Rate Constraint. So far, we have primarily used the additional equal-service-rate constraint ($\mu_r = \mu_n$) to generate analytical results for simplicity of results and ease of comparison with the baseline model. Removing this constraint, however, does not change our main insights. Consider a general case where the physician imposes that $\mu_r = m \cdot \mu_n$, where $m \geq 1$ is a multiplier that can be either exogenously given or endogenously determined. In this case, the optimal service rates are characterized in the proposition that follows:

PROPOSITION 15. *If the physician restricts that $\mu_r = m\mu_n, m > 1$, then the optimal service rates are:*

$$\mu_n^* = \frac{m+\theta\gamma}{m(1+\theta\gamma)} \cdot \frac{Q_c + \alpha\mu_c - (1-\beta)\pi}{2\alpha}, \text{ and } \mu_r^* = \frac{m+\theta\gamma}{1+\theta\gamma} \cdot \frac{Q_c + \alpha\mu_c - (1-\beta)\pi}{2\alpha}.$$

Proof. By applying the Pollaczek-Khinchine formula (Kleinrock 1975). The rest of the proof is similar to that of Proposition 1. □

It is straightforward to observe from Proposition 15 that

$$\mu_n^* \leq \frac{Q_c + \alpha\mu_c - (1-\beta)\pi}{2\alpha} \leq \mu_r^*.$$

In addition, the service system's mean service time is

$$\frac{1}{1+\theta\gamma} \cdot \frac{1}{\mu_n^*} + \frac{\theta\gamma}{1+\theta\gamma} \cdot \frac{1}{\mu_r^*} = \frac{2\alpha}{Q_c + \alpha\mu_c - (1-\beta)\pi}, \tag{41}$$

which is to say, the system's mean service rate is always exactly $\frac{Q_c + \alpha\mu_c - (1-\beta)\pi}{2\alpha}$ regardless of the multiplier m and the compliance rate θ! Our analysis about the physician's test-ordering behavior, therefore, remain valid if we examine the two classes of patients—both new and returning—as a whole.

Note that, even though we derive (41) under given θ and m, it remains valid even when the physician can endogenously choose θ and m. This confirms our intuition that the physician always chooses the service rates around the optimal service rate derived in the baseline model, and demonstrates the robustness of our analytical framework.

B.6 Effort-Averse Physician

We now consider the case where the physician is effort-averse such that the physician incurs an effort cost of c per unit of service time. The physician's problem is to choose μ and p to maximize her utility rate:

$$\left(p - \frac{c}{\mu}\right) \cdot \left[\mu - \frac{\omega}{Q(\mu) - \pi - \beta(p-\pi)}\right].$$

We characterize the optimal service rate in the corollary that follows:

COROLLARY 12. *When the physician is effort-averse, the optimal service rate is the same as in Proposition 1 of the paper.*

The corollary reveals that our baseline model is a robust one in predicting physicians' test-ordering behavior. Nevertheless, in response to different values of c, the physician will adjust the service fee and induce different equilibrium arrival rates.

B.7 Time-Dependent Compensation

Throughout the paper, we consider a fixed pricing scheme in which the physician sets a fixed service fee for each patient. In this section, we show that our results extend to the case where the service fee depends on the *actual* service time. We show that our main insights regarding physicians' test-ordering behavior remain qualitatively unchanged.

Consider a physician who sets the fee per unit of service time, denoted by φ, and the service rate μ; the physician's service fee, by abuse of notation, is denoted by a function $p(\tau)=\varphi\cdot\tau$, where τ is a random variable denoting the actual service time with $\mathbb{E}[\tau]=1/\mu$.

The equilibrium arrival rate $\lambda(\mu,\varphi)$ can be determined from the following market-clearing condition:

$$Q(\mu)-\pi-\beta\mathbb{E}[(\varphi\tau-\pi)^+]-\frac{\omega}{\mu-\lambda(\mu,\varphi)}=0,$$

which gives

$$\lambda(\mu,\varphi)=\mu-\frac{\omega}{Q(\mu)-\pi-\beta\mathbb{E}[(\varphi\tau-\pi)^+]}. \tag{42}$$

In the denominator of $\lambda(\mu,\varphi)$, by defining $f(\xi):=e^{-\pi\xi}$, the term $\mathbb{E}[(\varphi\tau-\pi)^+]$ can be approximated using Taylor series expansion as

$$\begin{aligned}\mathbb{E}[(\varphi\tau-\pi)^+]&=\int_{\frac{\pi}{\varphi}}^{\infty}(\varphi\tau-\pi)\mu e^{-\mu\tau}d\tau\\&=\frac{e^{-\frac{\pi}{\varphi}\cdot\mu}\varphi}{\mu}\\&=\frac{\varphi}{\mu}f\left(\frac{\mu}{\varphi}\right)\\&\approx\frac{\varphi}{\mu}\left[f(0)-\pi\cdot f'(0)\cdot\left(\frac{\mu}{\varphi}-0\right)\right]\\&=\frac{\varphi}{\mu}\left(1-\pi\frac{\mu}{\varphi}\right)=\frac{\varphi}{\mu}-\pi.\end{aligned} \tag{43}$$

The above approximation is accurate enough as μ/φ is usually close to zero in practical situations: consider, for example, $\mu=2$ patients/hour, $\varphi=\$200$/hour, and $\pi=\$10$, then $\mu/\varphi=0.01$, and the upper bound of the error, according to Taylor's theorem, is $R_2(\mu/\varphi)=f''(0)/2\cdot(\mu/\varphi)^2=0.005$.[6] Hence we can rewrite (42) as

$$\lambda(\mu,\varphi)=\mu-\frac{\omega}{Q(\mu)-\pi-\beta(\frac{\varphi}{\mu}-\pi)}.$$

Now the physician solves the revenue rate maximization problem

$$\max_{\mu,\varphi}\ \mathbb{E}[\varphi\tau\lambda(\mu,\varphi)]=\varphi\lambda(\mu,\varphi)\varphi\mathbb{E}[\tau]=\frac{\varphi}{\mu}\cdot\lambda(\mu,\varphi),$$

and obtains the optimal service parameters that are characterized in the proposition that follows:

Proposition 16. *When the physician's compensation is determined by the actual service time, there exists a unique equilibrium as follows.*

[6] Although the derivation of (43) was based on the assumption that the service time is exponentially distributed, the same result can be obtained without using Taylor series expansion when the service time τ, with a general distribution, satisfies the following condition: $\frac{\Pr(\tau<1/k)}{k/\mu-1}$ is small enough, where $k:=\frac{\varphi}{\pi}$. When $\mu=2$ patients/hour, $\varphi=\$200$/hour, and $\pi=\$10$, this means that we need $\Pr(\tau<0.05)/9$ to be small enough.

i) The physician chooses the service rate $\mu^* = \frac{Q_c+\alpha\mu_c-(1-\beta)\pi}{2\alpha}$*, and the fee per unit of service time* $\varphi^* = \frac{Q_c+\alpha\mu_c-(1-\beta)\pi}{4\alpha\beta}\left[Q_c+\alpha\mu_c-(1-\beta)\pi-2\sqrt{\alpha\omega}\right]$.

ii) The induced arrival rate is $\lambda(\mu^*,\varphi^*) = \frac{Q_c+\alpha\mu_c-(1-\beta)\pi}{2\alpha} - \sqrt{\frac{\omega}{\alpha}}$.

iii) The expected waiting time $\mathbb{E}[W(\mu^*,\lambda(\mu^*,\varphi^*))] = \sqrt{\frac{\alpha}{\omega}}$.

Proof. Similar to the proof of Proposition 1. □

Comparing Proposition 16 with Proposition 1 in the paper, we see that the physician chooses the same service rate as in the baseline model, and the physician achieves the same revenue rate. We hence conclude that all the results remain unchanged except that different patients pay different service fees depending on the realization of their actual service time.