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

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, we show that 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. 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

> Every time you walk into a doctor's office, it's implicit that someone else will be paying most or all of your bill; for most of us, that means we give less attention to prices for medical services than we do to prices for anything else. Most physicians, meanwhile, benefit financially from ordering diagnostic tests, doing procedures, and scheduling follow-up appointments. Combine these two features of the system with a third—the informational advantage that extensive training has given physicians over their patients, and the authority that advantage confers—and you have a system where physicians can, to some extent, generate demand at will.
>
> "How American Health Care Killed My Father" by David Goldhill, in *The Atlantic*, September 2009

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 $700 billion per year, or 5 percent of the nation's GDP, is spent on tests and treatments 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. 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, ... 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 toward *reverse* directions.

•**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.

•**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 §3.4 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). 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. These two observations lead to a simplified 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.

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. 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.

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. 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. 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 $1-\rho(\mu,p)$. 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). We further assume that all the service times are exponential, but our major results carry over to a general service time distribution. 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). 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

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

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). The physician's decision consists of choosing the service rate μ and the service fee p to maximize the revenue rate, that is,

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

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 service rate* μ^* *decreases in the copayment* π.

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

iii) The physician's 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, 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 $p_{\max} \leq \frac{Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} - (1-\beta)\pi}{2\beta}$. 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 service rate μ^* decreases in the copayment π.*

ii) The physician's 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 $p_{\max} \leq [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} - (1 - \beta)\pi]/(2\beta)$, then the physician can order more or fewer tests than the socially efficient level, that is, both $\mu^* \leq \mu^{SE}$ 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.

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. 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.

The threshold $(Q_c + \alpha\mu_c)/(1 + \alpha/d)$ in Part ii) of Corollary 6 has an intuitive interpretation. The numerator is the unattainable upper bound of the service quality, and the denominator 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.

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.

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

$\omega_i \leq \omega_i^*$ 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

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

and hence

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

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 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 1 provides a time line depicting the sequence of events. At $t = -1$, the physician 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

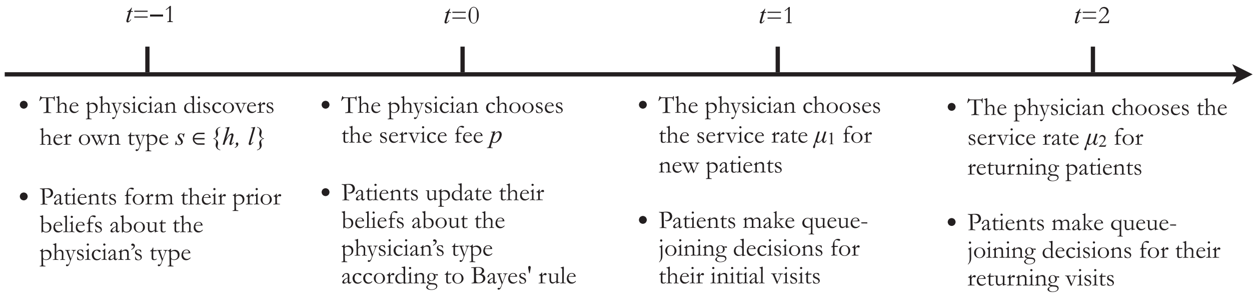
t=−1
t=0
t=1
t=2
• The physician discovers her own type s ∈ {h, l}
• Patients form their prior beliefs about the physician's type
• The physician chooses the service fee p
• Patients update their beliefs about the physician's type according to Bayes' rule
• The physician chooses the service rate μ1 for new patients
• Patients make queue-joining decisions for their initial visits
• The physician chooses the service rate μ2 for returning patients
• Patients make queue-joining decisions for their returning visits

Figure 1 Time line of the model with physician type uncertainty

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

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

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.

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 6. *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:*

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

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 (15).

Proposition 7. *There exists no pooling equilibrium that satisfies the intuitive criterion.*

The intuition behind Proposition 7 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 7 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 2, 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.

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

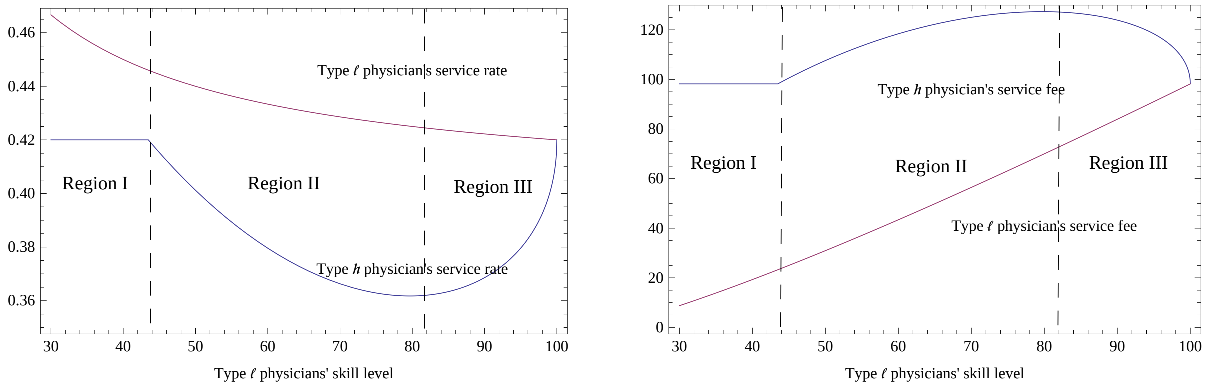

Figure 2 The impact of the different skill level differences among physicians on the physicians' service rates and fees. Parameters are: $\mu_c = 0.8, Q_c = 8, \omega = 5, \pi = 5, \beta = 0.2, \alpha_h = 100$.

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 different service environments: i) reimbursement ceiling, ii) misdiagnosis concerns, iii) patient heterogeneity, and iv) uncertainty about the physician type. We first 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. Last, 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 when bundled together with a distorted insurance structure. It is therefore imperative to create a legal and incentive 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.

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.

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.

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

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.

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

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.

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.

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

Proof of Proposition 1. We first show that the physician's objective function specified in (3) is concave in p since $\partial^2 g(\mu,p)/\partial p^2 = -2p\beta^2\omega/[Q(\mu) - \pi - \beta(p-\pi)]^3 - 2\beta\omega/[Q(\mu) - \pi - \beta(p-\pi)]^2 < 0$. 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 $\mu(Q_c + \pi(\beta-1) - \alpha\mu + \alpha\mu_c) = \mu[Q(\mu) + \pi(\beta-1)] > \mu[Q(\mu) - (\beta p + \pi(1-\beta))] \geq \frac{\mu\omega}{\mu-\lambda(\mu,p^*(\mu))} \quad = \omega \cdot \frac{\mu}{\mu-\lambda(\mu,p^*(\mu))} > \omega$, which gives $\sqrt{\mu(Q_c + \pi(\beta-1) - \alpha\mu + \alpha\mu_c)\omega} - \omega > 0$ since $\mu(Q_c + \pi(\beta-1) - \alpha\mu + \alpha\mu_c) > \omega$. Hence we see that the sign of

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

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 $\lambda(\mu^*,p^*)$: $\mathbb{E}[W(\mu^*,\lambda(\mu^*,p^*))] = 1/[\mu^* - \lambda(\mu^*,p^*)] = \sqrt{\alpha/\omega}$. *Q.E.D.*

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 $W(\mu^{SE}, \lambda^{SE}) = 1/(\mu^{SE} - \lambda^{SE}) = \sqrt{\alpha/\omega}$. *Q.E.D.*

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. *Q.E.D.*

Proof of Proposition 3. Similar to the proof of Proposition 1. *Q.E.D.*

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 $\mu^* = (Q_c + \alpha\mu_c - q_{\max} - \sqrt{\alpha\omega})/\alpha \geq (Q_c + \alpha\mu_c)/(2\alpha) = \mu^{SE}$. *Q.E.D.*

Proofs of Proposition 4 and Lemma 1. Similar to the proof of Proposition 1. *Q.E.D.*

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 $g_h(\mu_h^*, p_h^*) \geq g_h(\mu_l^*, p_l^*) > g_l(\mu_l^*, p_l^*)$. *Q.E.D.*

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 6. To prove Proposition 6, 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

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

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. *Q.E.D.*

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 6.

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. The full proof is available upon request.

Now that we have established Lemmas 2 and 3, we proceed to prove Proposition 6. We first consider Part i). When $\Delta\alpha \geq \Delta\alpha^*$, the single-crossing condition, specified by Lemma 3, gives

$$g_h(\mu_h^*, p_h^*) + \hat{g}_l(p_h^*) \leq 2g_l(\mu_l^*, p_l^*). \tag{16}$$

To understand the condition specified by (16), 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

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

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. *Q.E.D.*

Proof of Proposition 7. 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

$$\hat{g}_h(p) + \hat{g}_l(p) = 2g_l(\mu_l^*, p_l^*), \tag{17}$$

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

$$g_l(\mu_l^*, p_l^*) + \hat{g}_h(p_l^*) = 2\hat{g}_h(p). \tag{18}$$

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 (17) and (18), we have the following equation:

$$2\hat{g}_h(\bar{p}_2) = \hat{g}_h(\bar{p}_1) + \hat{g}_l(\bar{p}_1) + \hat{g}_h(p_l^*) - \hat{g}_l(p_l^*). \tag{19}$$

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 (19), gives

$$2\hat{g}_h(\bar{p}_2) = \hat{g}_h(\bar{p}_1) + \hat{g}_l(\bar{p}_1) + \hat{g}_h(p_l^*) - \hat{g}_l(p_l^*) > 2\hat{g}_h(\bar{p}_1). \tag{20}$$

Equation (20), 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 (17) and (18), after some algebra, jointly give

$$2\hat{g}_h(\underline{p}_2) = \hat{g}_h(\underline{p}_1) + \hat{g}_l(\underline{p}_1) + \hat{g}_h(p_l^*) - \hat{g}_l(p_l^*). \tag{21}$$

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 (21), gives

$$2\hat{g}_h(\underline{p}_2) = \hat{g}_h(\underline{p}_1) + \hat{g}_l(\underline{p}_1) + \hat{g}_h(p_l^*) - \hat{g}_l(p_l^*) < 2\hat{g}_h(\underline{p}_1). \tag{22}$$

Equation (22), 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 6 that a costly separating equilibrium always prevails whenever a costless equilibrium does not exist. *Q.E.D.*