MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 19, No. 1, Winter 2017, pp. 99–113
ISSN 1523-4614 (print), ISSN 1526-5498 (online)
http://pubsonline.informs.org/journal/msom/

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

Tinglong Dai,[a] Mustafa Akan,[b] Sridhar Tayur[b]

[a] Carey Business School, Johns Hopkins University, Baltimore, Maryland 21202; [b] Tepper School of Business, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

Contact: dai@jhu.edu (TD); akan@andrew.cmu.edu (MA); stayur@andrew.cmu.edu (ST)

Received: January 20, 2016
Revised: April 13, 2016
Accepted: May 31, 2016
Published Online in Articles in Advance: November 14, 2016

https://doi.org/10.1287/msom.2016.0594

Abstract. Motivated by a collaborative study with one of the most comprehensive ocular imaging programs in the United States, we investigate the underlying three-way trade-off among operational, clinical, and financial considerations in physicians' decisions about ordering imaging tests. Laboratory tests may be processed in parallel and thus have a limited effect on patients' waiting times; imaging tests, by contrast, require patient presence and thus directly influence patients' waiting times. We use a strategic queueing framework to model a physician's decision of ordering imaging tests and show that insurance coverage is the key driver of overtesting. Our further analysis reveals the following: (i) Whereas existing studies hold that lower out-of-pocket expenses lead to higher consumption levels, we refine this statement by showing the copayment and the coinsurance rate drive the consumption in different directions. Thus, simply expanding patient cost sharing is not the solution to overtesting. (ii) Setting a low reimbursement ceiling alone cannot eliminate overtesting. (iii) The joint effect of misdiagnosis concerns and insurance coverage can lead to both overtesting and undertesting even when no reimbursement ceiling exists. These and other results continue to hold under more general conditions and are therefore robust. We enrich our model along two extensions: one with patient heterogeneity in diagnostic precision, and the other with disparities in health insurance coverage. Our findings have implications for other healthcare settings with similar trade-offs.

Funding: The authors also thank the National Science Foundation [CMMI 1351821 and CMMI 1334194] for financial support. Tinglong Dai acknowledges partial financial support from the inaugural Johns Hopkins Discovery Award.

Supplemental Material: The online appendix is available at https://doi.org/10.1287/msom.2016.0594.

Keywords: queueing games • service operations • healthcare management • consumer behavior • incentives in healthcare operations

1. Introduction

Excessive usage of imaging tests (hereafter referred to as "overtesting") is generally attributed to either health providers' misaligned monetary incentives (Gawande 2015) or physicians' misdiagnosis concerns, as noted by Pinker (2012, p. 506): "A physician may order tests that have a low probability of influencing decisions or treatment, to play it safe and protect himself from malpractice lawsuits, knowing that the patient will not bear the financial costs of the tests." Our collaborative study with the University of Pittsburgh Medical Center (UPMC) Eye Center, which houses one of the most comprehensive ocular imaging programs in the United States, revealed a different picture. In the existing payment model at the UPMC, insurance plans approve payment for only one test per day per patient. Moreover, depending on the type of test and disease, insurance firms limit the number of reimbursable tests. For instance, when a physician orders three tests for a patient, the physician understands the insurance firm will reimburse only one test, and the other two will not generate additional service revenue. However, overtesting has been consistently observed at the UPMC Eye Center despite physicians' general lack of direct monetary incentives and misdiagnosis concerns (Dai et al. 2009).

Based on our interviews with physicians and patients at the 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, longer waiting times lead to lower patient volume, all else being equal (Martin and Smith 2003). Third, patients value service quality, which (at least perception-wise) often increases in the quantity of diagnostic tests, though the marginal return from ordering additional tests may be diminishing (Mold et al. 2010). In practice, patient perception of service quality may be relevant to other factors such as responsiveness and empathy; we assume away from these factors and focus on the service intensity because it is more controllable from the viewpoint of physician decision making. These

three aspects echo the iron triangle of U.S. healthcare, namely, cost, access, and quality (Kissick 1994).

In modeling physician decision making in the ocular imaging setting, we capture key financial, operational, and clinical incentives that govern the interactions between the physician and patients. Whereas the physician strikes a balance between system throughput and diagnostic certainty, patients optimally trade off between waiting time, out-of-pocket expense, and service quality. We characterize the physician's optimal service parameters and patients' queue-joining decisions, 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 is the loss of social welfare with respect to the socially efficient administration of imaging tests. Our model reveals several interesting results that help us understand physicians' decision making in the ocular imaging setting.

First, we show that even in the absence of the fee-for-service payment system and other commonly cited reasons, overtesting can still occur due to the insurance coverage that distorts the price signal. This result is aligned with the empirical literature, with the difference that overtesting occurs even in the absence of asymmetric information.

Second, although existing studies show that lower out-of-pocket expenses lead to higher consumption levels, we refine this statement by showing the copayment and the coinsurance rate can drive the consumption in opposite directions.

Third, when insurance firms impose reimbursement ceilings on physician practices, they essentially restrict physicians' pricing power. Under a reimbursement ceiling, we show overtesting can nonetheless occur even when the ceiling is low, and increasing the share of patients' cost sharing (e.g., coinsurance rate) can induce the ordering of more tests.

Fourth, in some situations, physicians are concerned about potential misdiagnosis and may perceive a "burden of proof" that decreases in the intensity of testing. Contrary to conventional wisdom, we show that both overtesting and undertesting are possible outcomes of the introduction of misdiagnosis concerns. The underlying intuition is that physicians' misdiagnosis concerns raise the socially efficient consumption level.

In addition, we consider two extensions on patient heterogeneity in diagnostic precision and disparities in health insurance coverage, respectively.

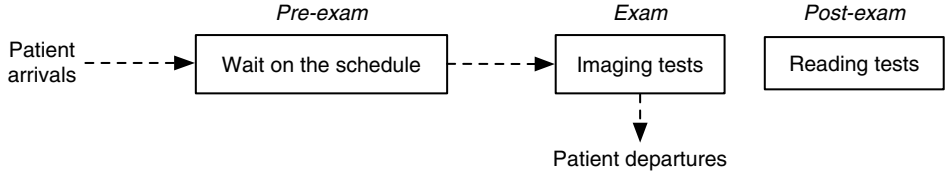
1.1. Salient Features of the UPMC Eye Center Scenario

We describe the salient features in ordering and conducting ocular imaging tests at the UPMC Eye Center, which is representative of many academic, elective outpatient settings. Figure 1 shows the flow schematic of conducting imaging tests.

At the UPMC Eye Center, when physicians order imaging tests, they typically order multiple tests all at once; this process is in contrast to a sequential testing process (i.e., start from one test and then decide whether to order further tests, depending on the information collected from the first test, and so on) in other clinical settings. A typical order is a combination of tests such as OCT (short for "optical coherence tomography," which provides cross-sectional analysis of birefringent tissues in the eye), GDx (short for "glaucoma diagnosis," which measures the thickness of the retinal nerve to determine the occurrence of structural damage), and HRT3 (short for "Heidelberg Retina Tomograph," which provides a 3D topography image of the optic nerve). Few patients require a second batch of tests to be ordered for the same complaint.

As mentioned previously, the payment system is non-fee-for-service, but the phenomenon of overtesting has been observed at the UPMC Eye Center, as demonstrated by significant variation in test-ordering patterns even for patients with comparable conditions (Dai et al. 2012). Because of extensive testing for patients, a high utilization rate exists and leads to patients' long waiting times for imaging tests. More generally, Hopp and Lovejoy (2012, p. 368) contend that in a typical imaging-test unit, the time waiting on schedule is the single largest source of delay in obtaining images for most patients. Hopp and Lovejoy (2012, p. 367) also highlight that the key distinction between laboratory tests (e.g., blood tests) and imaging tests is patient contact: "For laboratory tests, all that is needed from a patient is a specimen, which can be collected remotely in a hospital room, clinic, physician's office, or even by the patient himself at home. With a few exceptions . . . patients must generally come to a central

Figure 1. Flow Schematic for Imaging Services



Source. Adapted from Hopp and Lovejoy 2012.

location to be scanned to produce images." Therefore, the service rate associated with imaging tests limits the patient throughput.

1.2. Broader Implications

Although our research was initially motivated by the operations of an ophthalmology practice at an academic hospital (UPMC), our model applies to more (but not all) types of in-house imaging practices. In particular, we consider an elective care setting in which (1) patients call in for nonurgent health issues, (2) discretion in testing allows different levels of thoroughness, (3) waiting times are driven by the fact that imaging is the constraint of patient flow (we found this feature is not suitable for small practices), and (4) more tests do not necessarily generate higher revenue. Beyond the academic hospital with which we collaborated (UPMC), we have found that the above features are typical in mid- to large-size ocular imaging practices (20 or more physicians). Furthermore, this model applies to examples other than the eye clinic. For example, the imaging units in the orthopaedics setting share multiple operational features.

Admittedly, the results in this paper are based on a set of assumptions drawn from our observational study in the academic ophthalmology setting. But the results have broader implications to other healthcare settings in which (1) the quality of a patient's received care is known to correlate positively with the amount of time and resources dedicated to the patient; (2) physicians value clinical, financial, and operational aspects in making procedural decisions; and (3) patients value quality of care as well as their waiting time and out-of-pocket expenses. By not explicitly modeling the fee-for-service payment structure, our model isolates its well-known and obvious incentive effect and helps focus our attention on investigating nonobvious aspects of physicians' test-ordering behavior.

Our work joins in the first attempts of the operations management community to understand the immensely complex web of incentives in the U.S. healthcare system, widely known today for its inefficiency. One major aspect of the inefficiency is overtesting (Gawande 2009, 2015). Unfortunately, conventional cost-containment strategies are ineffective, because as Young and Saltman (1985) argue in the book *The Hospital Power Equilibrium: Physician Behavior and Cost Control*, these strategies offer little to help bridge the divide between administrators and physicians, with conflicting objectives and incentives. Young and Saltman (1985) further argue that the cost of healthcare is largely driven by physician behavior. To curb the phenomenon of overtesting, the first step entails understanding its underlying drivers (Rao and Levin 2012).

1.3. Literature Review

Our study continues the themes of expert services literature, for which Dulleck and Kerschbamer (2006) provide an extensive review. Shumsky and Pinker (2003) study the incentive compensation scheme for "gatekeepers" who are imperfectly capable of solving customers' problems and may need to refer them to specialists. Debo et al. (2008) model a monopolist expert who offers a service with unverifiable duration and hence has the incentive to delay the service. While embedding asymmetric information, their model does not address the differences in service quality. Debo and Veeraraghavan (2014), similar to us, assume service time and service quality are positively correlated, but they focus on analyzing strategic consumers' queue-joining behavior. Paç and Veeraraghavan (2015) model the interaction between customers with problems that can be either major or minor by nature and an expert who may choose not to reveal the true nature of various problems to sell more extensive services.

In recent years, literature on service design under congestion has flowered. Wang et al. (2010) consider the problem of a diagnostic service manager who needs to strike a balance between service accuracy, waiting time, and staffing costs. Alizamir et al. (2013) examine how to dynamically manage the trade-off between diagnostic accuracy and system congestion. Kostami and Rajagopalan (2013) analyze the intertemporal trade-off between speed and quality in a general service setting. Tong and Rajagopalan (2014) study the pricing strategy for discretionary services when the service outcome is contractible and is directly driven by the service provider's service choice. Our paper addresses a trade-off similar to as those in Wang et al. (2010) and Alizamir et al. (2013) but focuses on the economic side of physicians' test-ordering behavior.

Most relevant to our paper is the study by Anand et al. (2011) on a service provider's quality–speed trade-off when customers are strategic. They show the customer intensity of the industry is a major determinant of the service provider's decision. Furthermore, they analyze the competition among multiple service providers and show that higher prices and service quality can result from a more intense competition. Our paper differs from Anand et al. (2011) in that we consider insurance coverage in the healthcare market and emphasize the profound impact of insurance structure on the service usage under various service environments. Furthermore, we compare service consumption levels in the market equilibrium and in the social optimum. In addition, we consider a scenario in which the service provider chooses more than one service rate to achieve the same quality target across different patient types.

Our paper is relevant to several papers in the operations literature on reimbursement design for healthcare providers, ranging from early work by Dada and

White (1999) and So and Tang (2000) to recent work by Gupta and Mehrotra (2015). The most salient difference separating our paper from this literature is that whereas the literature focuses on healthcare providers' financial risks, we study the effect of patients' out-of-pocket expenses on physicians' medical decision making.

The supplier-induced demand (SID) literature contends that doctors, as service providers, can directly influence patients' service usage. Patients seek advice from doctors largely because they cannot reach informed medical decisions on their own. Whereas early SID models often view patients as perfectly informed but passive consumers, later studies treat patients as Bayesian decision makers whose information-acquisition mechanism affects physicians' behavior. Our paper differs from the SID literature in three ways. First, SID models generally assume physicians can observe patients' private information at no cost. Second, although waiting time limits the patients' access to healthcare, SID models treat it as a mechanism to control utilization and hence reduce the cost of ex post moral hazard (Gravelle and Siciliani 2008): the waiting time is the healthcare provider's unilateral decision rather than an outcome of physician–patient interaction. Third, the SID literature typically assumes a fee-for-service payment model. For example, Sorensen and Grytten (1999) work on the premise that only contract physicians in Norway, whose incomes derive exclusively from patient visits or laboratory tests, have the incentive to induce demand. Our paper, by addressing the trade-off among cost, access, and quality, justifies incentives to overtest even when more services do not imply additional revenue.

This paper proceeds as follows. In Section 2, we present our modeling framework. Section 3 analyzes the effect of insurance structure, reimbursement ceiling, and misdiagnosis concerns. Then we enrich our model along two directions: Section 4 considers patient heterogeneity in diagnostic precision; Section 5 considers disparities in health insurance coverage. Section 6 summarizes the key implications from our study for policy makers. In Section 7, we conclude. All the technical proofs are relegated to Online Appendix A.

2. Model

In this section, we model the interaction between a physician and a group of patients with exogenous demand. We start by formalizing the relationship between imaging-test ordering and service quality. We then model the trade-offs the patients and the physician face. Finally, we characterize the market equilibrium and the social optimum, which gives the condition of overtesting.

2.1. Imaging-Test Ordering and Service Quality

We capture the physician's test-ordering decision by the service rate μ that measures the speed of the overall imaging service, based on our observation that a higher service rate results from demanding fewer imaging tests. In reality, the service rate is chosen from a discrete set. (In the case of the UPMC Eye Center, each ordered imaging test is assigned a slot of a given length, and the number of tests being ordered for a patient directly influences the service rate of conducting imaging tests.) For tractability, we assume the service rate is a continuous variable. The service quality given average service rate μ is defined as

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

where Q_c denotes the baseline service quality; μ_c refers to the baseline service rate, that is, $Q(\mu_c) = Q_c$; α describes the rate at which the service quality improves when the service rate decreases. It follows from (1) that $Q(\mu)$ is a decreasing affine function of μ, meaning more tests (i.e., a slower service rate) lead to higher service quality. This model is aligned with an elective outpatient setting such as an ophthalmology clinic, where additional imaging tests do not lead to the phenomenon of "overdiagnosis"; rather, more imaging tests lead to better image resolution and increased diagnosis accuracy (Dai et al. 2009). We use this function for simplicity of representation; our major insights extend to the case in which $Q(\cdot)$ is a general concave and nonmonotonic function.

2.2. Patient Utility

Patients' utility from the service depends on three key factors: service quality, waiting time, and out-of-pocket payment. Patients are covered by indemnity insurance and pay less than the nominal service charge. A patient's health insurance plan contains several key components: the deductible is the accumulative out-of-pocket expense to trigger insurance coverage; the copayment is the fixed charge the patient must pay out of pocket for each visit; the coinsurance rate is the percentage of the service fee, after accounting for the copayment, the patient has to pay out of pocket. We ignore the deductible to avoid the difficulty of defining the service fee below the deductible (Newhouse 1978); in practice, the deductible may play some role in influencing patient demand, and the importance of such a role may depend on the timing of the visit. All patients have the same insurance coverage with zero deductible, a copayment of π, and a coinsurance rate of β. Later, in Section 5, we allow patients to differ in their health insurance coverage. The premium is viewed as a sunk cost and ignored. Letting p denote the nominal service fee, the patient's out-of-pocket payment is hence $\pi + \beta(p - \pi)$, because we focus solely on the interesting case in which $p \geq \pi$.

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 normalized to be zero without loss of generality. The induced arrival rate can be denoted as a function of μ and p such that $\lambda(\mu, p) = \rho(\mu, p) \cdot \Lambda$. Λ is assumed to be large enough to avoid the situations with zero or full coverage (i.e., $\rho(\mu, p) = 0$ or 1) in equilibrium. This setting is consistent with the literature on the equilibrium behavior of customers and service providers in queueing systems (Hassin and Haviv 2003).

The potential demand for the service follows a Poisson process, a reasonable representation for arrival processes in healthcare applications (Green 2006); thus, the induced arrival process resulting from patients' joint randomized decisions also follows a Poisson process. For simplicity, we assume service time is exponentially distributed, and the service setting corresponds to an $M/M/1$ queue; our major results carry over to a general service-time distribution. Consistent with *money price* models (e.g., Coffey 1983), we define the patient's waiting time $W(\mu, \lambda)$ as the amount of time a patient spends in the system before a diagnosis is reached (i.e., the sojourn time); in our particular setting, the waiting time consists mostly of the patient's time on the schedule for an appointment. The expected time in the system is given by $W(\mu, \lambda(\mu, p)) = 1/[\mu - \lambda(\mu, p)]$. Let ω denote the patient's waiting cost per unit of time. In practice, ω can be estimated as the value of lost productivity while waiting in the service queue (Phelps and Newhouse 1974). The sum of out-of-pocket expense $\pi + \beta(p - \pi)$ and waiting cost $\omega W(\mu, \lambda(\mu, p))$ is referred to as the full price. Then, using the market-clearing condition $Q(\mu) = \pi + \beta(p - \pi) + \omega W(\mu, \lambda(\mu, p))$ that equates the service quality to its full price and substituting for W, we obtain the induced arrival rate

$$\lambda(\mu, p) = \mu - \omega[Q(\mu) - \pi - \beta(p - \pi)]^{-1}.$$

In reality, the physician classifies patients into a number of pretest types such that within each pretest type, the physician orders a virtually identical set of tests for each patient. Although such a patient mix faced by the physician is best described by multiple classes of arrivals in a queueing network, in our baseline model, we choose to focus on modeling patients who are classified as the same pretest type and ultimately face the set of imaging tests. This simplification allows us to analytically characterize patients' strategic queue-joining decisions in response to the service parameters. Later, in Section 4, we consider the physician's test-ordering problem in the presence of multiple patient types.

2.3. Physician Decision

We treat 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, p. 3); this assumption is supported by patients' free choice of physicians, meaning that "in any negotiation over price between a physician and an insurer physicians have substantial bargaining power" (Newhouse 2002, p. 10). Recognizing that prices are set administratively in many situations (Gaynor and Town 2012), we will discuss in Section 3.2 an alternative scenario in which the service fee is subject to a price ceiling. The physician chooses the target service quality (through choosing a service rate μ) and a service fee p to maximize the revenue rate $g(\mu, p) = p\lambda(\mu, p)$. This model is an extension of Anand et al. (2011) with the critical difference that each customer pays a linear function of price, which allows us to draw insights about the impact of insurance structure on the physician's test-ordering decisions. We have numerically shown that our main insights carry over if we relax the assumption that quality decreases linearly in service rate (e.g., the physician chooses an integer number of tests to reach the desired diagnostic precision, and the service time follows an Erlang distribution with an exogenous scale parameter), suggesting that this model is a good starting point of studying physicians' test-ordering decisions.

We assume $Q_c < \alpha\mu_c + (1-\beta)\pi$ to rule out the trivial case $\mu^* \geq \mu_c$. 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 improvement in service quality, and (2) $(1-\beta)\pi$, each patient's copayment net of β, which is covered by the insurance. We characterize the equilibrium below. Note that we require that $\beta > 0$; if $\beta = 0$, a reimbursement ceiling—an issue that we will address in Section 3.2—has to be in place.

Proposition 1. *A unique market equilibrium exists in which*

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

(ii) *the induced arrival rate is* $\lambda^* = \mu^* - \sqrt{\omega/\alpha}$;

(iii) *the average waiting time is* $W^* = \sqrt{\alpha/\omega}$.

In equilibrium, the waiting time does not depend on the insurance structure. Because the waiting time $W(\mu^*, \lambda^*) = (\mu^* - \lambda^*)^{-1}$ spent per patient in the system depends only on the "surplus" service level $\mu^* - \lambda^*$, the optimal solution balances the cost α of increasing the service rate (i.e., the reduced diagnostic quality) with the reduction ω in each patient's waiting costs.

2.4. Social Optimum and Overtesting Condition

The benchmark used to characterize overtesting is the social optimum that involves a social planner who determines the admission policy and the service rate to maximize the social welfare. Each physician–patient interaction generates a social surplus that is equal to the service quality, less patients' disutility from waiting. The expected social welfare rate is formulated as follows:

$$U(\mu, \lambda) = \lambda \cdot \{Q(\mu) - \omega W(\mu, \lambda)\}.$$

The following proposition gives the socially efficient service rate and arrival rate, denoted by μ^S and λ^S, respectively.

Proposition 2. *In the social optimum,*

(i) *the optimal service rate is* $\mu^S = (Q_c + \alpha\mu_c)/(2\alpha)$;

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

(iii) *the expected waiting time is* $W^S = \sqrt{\alpha/\omega}$.

The above social-optimum result coincides with Proposition 2 in Anand et al. (2011), which characterizes the market equilibrium when each customer pays the full amount of the service fee.

Next, we compare the market equilibrium with the social optimum.

Corollary 1. (i) *The physician orders at least as many tests as in the market equilibrium, that is,* $\mu^* \le \mu^S$.

(ii) *The arrival rate in the social optimum is at least as large as in the market equilibrium, that is,* $\lambda^* \le \lambda^S$.

(iii) *The average waiting time is the same under the social optimum and in the market equilibrium, that is,* $W^S = W^* = \sqrt{\alpha/\omega}$.

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

3. Analysis

This section analyzes the effects of insurance structure, reimbursement ceiling, and misdiagnosis concerns. We also briefly describe a robustness check of our key results.

3.1. Insurance Structure

We first examine the effect of the copayment and coninsurance components of the insurance plan on the physicians' test-ordering behavior. Proposition 1 implies the following result.

Corollary 2. (i) *The physician's optimal service rate* μ^* *decreases in the copayment* π *and increases in the coinsurance rate* β.

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

The literature often suggests that increasing the patients' out-of-pocket expenses leads to decreased consumption of medical resources. The above corollary, by contrast, reveals the copayment and the coinsurance rate can drive the consumption of imaging tests in *opposite* directions. In particular, the number of tests increases in the copayment π but decreases in the coinsurance rate β. To understand why, note the market-clearing condition

$$\pi + \beta(p^* - \pi) = Q(\mu^*) - W^*.$$

We have from Proposition 1 that $W^* = \sqrt{\alpha/\omega}$, which does not depend on π or β. Thus, for a service fee p^*, the change in π or β would need to be balanced by a change in μ^*. 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, necessitating more tests to match the patients' increased monetary burden. With a higher coinsurance rate, however, the physician would charge a lower service fee, which leads to a reduced out-of-pocket expense for each patient and justifies the physician ordering fewer tests. Note that in practice, it is possible that the coinsurance component significantly outweighs the copayment component; in this case, the effect of the coinsurance would dominate that of the copayment.

To the best of our knowledge, ours is the first analytical finding about the impact of per-visit copayment on physicians' test-ordering behavior. Supporting empirical evidence exists for this result. For example, under an outpatient setting, Jung (1998) shows that increasing the per-visit copayment significantly reduces the number of office visits but increases the intensity of medical resource consumption for each visit.

Our results have implications on the impact of the increased health insurance coverage made possible by the Patient Protection and Affordable Care Act, under which more individuals have gained insurance coverage. From our results derived under an outpatient imaging-unit setting, we find the structure of health insurance plays an important role in influencing physicians' service decisions. Thus, a singular focus on improving insurance coverage may not necessarily lead to a more efficient health system.

Next, we examine the effect of the insurance structure on the social welfare gap between the market equilibrium and the social optimum. The social welfare

gap, written as a function of β and π, is $\Delta U(\pi,\beta) = U(\mu^S,\lambda^S) - U(\mu^*,\lambda^*) = \pi^2(1-\beta)^2/(4\alpha)$, 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 we have the following corollary.

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 tests for each patient, but the induced arrival rate decreases. Combining the decreased arrival rate with the increased testing intensity per patient visit, we observe that fewer individuals consume more resources at any given period of time, widening the social welfare gap at a faster pace. This result explains why the social welfare gap is convex increasing in the copayment π. As the coinsurance rate increases, both the physician's test-ordering pattern and the equilibrium arrival rate converge to the social optimum.

3.2. Reimbursement Ceiling

Given that insurance coverage distorts the demand curve for imaging services, one natural proposal would be to introduce a reimbursement ceiling that sets the maximum reimbursable amount for each service session. This ceiling essentially restricts the maximum service fee the physician charges, which we denote by $p_{\max}$. Define $q_{\max} = \pi + \beta(p_{\max} - \pi)$ and $\tilde{p} = [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} - (1-\beta)\pi]/(2\beta)$. The proposition that follows characterizes the equilibrium.

Proposition 3. *If $p_{\max} \leq \tilde{p}$, then*

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

(ii) *the induced arrival rate is $\lambda^* = \mu^* - \sqrt{\omega/\alpha}$;*

(iii) *the average waiting time is $W^* = \sqrt{\alpha/\omega}$.*

When $p_{\max}$ is low, it becomes restricting such that the physician would choose a service fee that is exactly the same as $p_{\max}$. This scenario essentially corresponds to the setting in which the physician's service fee is capped by the insurance firm's reimbursement ceiling.

The following corollary illustrates how the presence of a reimbursement ceiling affects the physicians' test-ordering behavior.

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

(ii) *The physician's optimal service rate μ^* decreases in the coinsurance rate β if and only if the reimbursement ceiling exceeds $p_{\max} \leq \tilde{p}$.*

The intuitions behind Corollary 4 are threefold. 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, the physician responds to a decrease in the coinsurance rate β by ordering more tests because patients are less sensitive to the service fee. 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. Note from Proposition 3 that the expected waiting time W^* is independent of π and β; thus, the market-clearing condition $\pi + \beta(p_{\max} - \pi) = Q(\mu^*) - W^*$ indicates that an increase in π or β would need to be balanced by an increase in the service quality (i.e., a decrease in the service rate). Therefore, under a low reimbursement ceiling, increasing the ratio of a patient's out-of-pocket expense to the total service fee leads to a higher testing level, and vice versa. These findings are in line with the empirical findings by Danzon (1982) that Medicaid or Medicare patients—often with the lowest out-of-pocket expenses—experience fewer tests than patients with other insurance plans.

As in the baseline model, the social welfare gap is convex increasing in the copayment π, because both the service rate μ^* and the equilibrium arrival rate λ^* decrease in π. With a high reimbursement ceiling, as in the baseline model, the social welfare gap is convex decreasing in β. With a low reimbursement ceiling, however, Corollary 4 suggests both the arrival rate and the service rate decrease in β, meaning the social welfare gap is convex increasing in β.

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

Corollary 5. *If the reimbursement ceiling $p_{\max} \leq \tilde{p}$, the physician can order more or fewer tests than the socially efficient level; that is, both $\mu^* \leq \mu^S$ and $\mu^* > \mu^S$ are possible.*

The above corollary provides the condition under which overtesting occurs. Note from our analysis in Section 2 that with a high reimbursement ceiling, the physician always overtests. With a low reimbursement ceiling, however, Corollary 5 states the physician can either overtest or undertest, depending on whether each patient's out-of-pocket expense $q_{\max}$ is more than $(Q_c + \alpha\mu_c - 2\sqrt{\alpha\omega})/2$. To give a numerical example, consider $\mu_c = 8$, $Q_c = 50$, $\omega = 5$, $\alpha = 20$, $\beta = 0.2$, and $\pi = 50$; overtesting occurs as long as $q_{\max}$ exceeds \$58. This is because more tests compensate for a higher net payment, and vice versa. This result echoes the work of Yip (1998), who empirically identifies a high usage of medical procedures in the presence of low reimbursement ceilings.

Corollary 5 also helps uncover the puzzle that motivates our research. Recall from Section 1 that overtesting occurs even under the exogenous pricing scenario, that is, when the physician receives the same revenue 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 in which the reimbursement ceiling is set low enough and the physician always sets the service fee at the maximum possible amount (see Proposition 3(ii)). 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^S. In other words, overtesting is still possible even under exogenous pricing.

3.3. Misdiagnosis Concerns

In some scenarios, the physician bears the risk of misdiagnosis. For example, an inadequate number of tests can indicate a normal patient is abnormal, exposing patients to unnecessary treatments. Prior medical literature demonstrates the significance of misdiagnosis concerns in their scope and impact. Studdert (2006) find that 37% of malpractice claims do not involve any *real* medical errors but nevertheless account for 13%–16% of the system's total costs. In a study to reveal physicians' perceived risk of misdiagnosis, Carrier et al. (2010) confirm high malpractice concerns among physicians at all levels even when malpractice risks are sufficiently low by objective measures. They also find that common tort reforms do not ease such concerns. Baicker et al. (2007) show that increased malpractice risk drives higher consumption levels of healthcare services, especially when they are discretionary.

We now incorporate a misdiagnosis cost in our modeling of physician decision making. The misdiagnosis cost is a real cost incurred by the physician, and essentially captures the nonfinancial aspect of the physician's expected costs due to potential malpractice lawsuits. In practice, the physician can present tests in the court as evidence for providing adequate medical care in the case of a malpractice lawsuit. Thus, it is a "burden of proof" that decreases in the intensity of testing (i.e., increases in the chosen service rate). Kessler and McClellan (2002) highlight the notion of "malpractice pressure" and contend that such pressure can be both financial and nonfinancial. The financial part usually does not factor into individual physicians' decision making and would be canceled out in the social welfare equation, because the premium of malpractice insurance is community rated, and rarely depends on malpractice claims. However, Kessler and McClellan (2002, pp. 933–934) argue,

> No insurance is possible against the unpleasant experiences and considerable time commitment over months or years. For example, in discovery, a physician may be required to answer both written and oral questions about her competence and judgment and to respond to questions and other requests from lawyers for the patient, for the malpractice insurer, and for the hospital and its malpractice lawyer.

We model the physician's misdiagnosis concerns as a misdiagnosis cost function of the service rate, $\theta(\mu) := d \cdot \mu$, where d is a constant denoting the marginal increase in misdiagnosis concerns and can be interpreted as the parameter for "burden of proof." The misdiagnosis cost increases in μ, to align with the observation that fewer tests cause the physician to be more burdened by proof of due diligence in the event of a lawsuit. When μ is very small, indicating the physician orders a sufficiently large number of tests, the misdiagnosis cost approaches zero. Note that the real misdiagnosis risk has already been incorporated into the modeling of service quality $Q(\mu)$; $\theta(\mu)$ is a burden incurred by the physician only.

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

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

(i) *the physician chooses the service rate* $\mu_m^* = [Q_c + \alpha\mu_c - (1-\beta)\pi]/[2(\alpha + \beta d)]$ *and the service fee* $p_m^* = (\alpha + 2\beta d)[\mu_m^* - \sqrt{\omega/(\alpha + \beta d)}]/\beta$;

(ii) *the induced arrival rate is* $\lambda_m^* = \mu_m^* - \sqrt{\omega/(\alpha + \beta d)}$;

(iii) *the average waiting time is* $W_m^* = \sqrt{(\alpha + \beta d)/\omega}$.

Proposition 4 indicates that as the physician's "burden of proof" (d) increases, the optimal service rate decreases. In other words, a lower "burden of proof" leads to a higher system capacity, which is aligned with the empirical finding by Kessler and McClellan (2002, p. 953), empirical finding that "policies that reduce the time spent and the amount of conflict involved in defending against a claim [can] reduce defensive practices substantially."

The corollary below follows from Proposition 4.

Corollary 6. (i) *With misdiagnosis concerns, the physician's optimal service rate* μ_m^* *decreases in the copayment* π.

(ii) *If* $d < \alpha[(Q_c + \alpha\mu_c)/\pi - 1]^{-1}$, *the physician's optimal service rate* μ_m^* *increases in the coinsurance rate* β; *otherwise, the physician's optimal service rate* μ_m^* *decreases in the coinsurance rate* β.

An increase in the fixed per-visit charge increases the requirement for service quality and so justifies more tests. An increase in the coinsurance rate β, however, can lead to either an increase or a reduction in the optimal service rate μ_m^*, depending on the size of d. When d is low, similar to the case without misdiagnosis concerns, an increase in the coinsurance rate leads to a lower service fee and lower service quality (i.e., a higher service rate). When d is high, due to the high "burden of proof," the physician no longer finds it optimal to lower the intensity of testing. Rather, it is

optimal to compensate patients' higher expenses by increasing the intensity of testing.

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

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

(i) *the optimal service rate is* $\mu_m^S=(Q_c+\alpha\mu_c)/(2(\alpha+d))$;

(ii) *the optimal arrival rate is* $\lambda_m^S=\mu_m^S-\sqrt{\omega/(\alpha+d)}$;

(iii) *the expected waiting time is* $W_m^S=\sqrt{(\alpha+d)/\omega}$.

The following corollary is immediate from Propositions 4 and 5.

Corollary 7. *If the copayment* π *is higher than* $(Q_c+\alpha\mu_c)/(1+\alpha/d)$, *the physician orders more tests than the socially efficient level, that is,* $\mu_m^*<\mu_m^S$; *otherwise, the physician orders fewer tests than the socially efficient level.*

Corollary 7 is rather counterintuitive: when physicians suffer from "burden of proof" in the case of potential inaccurate medical judgment, they may either overtest or undertest (i.e., order either more or fewer tests than the socially optimal level). The corollary is especially surprising in view of Corollary 1, which states the physician always overtests in the absence of misdiagnosis concerns. To understand this result, we recall from Proposition 5 that misdiagnosis concerns increase the 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 exceeds $(Q_c+\alpha\mu_c)/(1+\alpha/d)$, the insurance coverage supplements the physician's efforts to induce demand. Furthermore, given Q_c and μ_c, the threshold decreases in the ratio of α and d. Consider the special case in which the physician's misdiagnosis concern is sufficiently low (i.e., d is small): the threshold is then close to zero, meaning the physician invariably overtests, which is consistent with Corollary 1.

Corollary 8. *The average waiting time in the social optimum is longer than in the market equilibrium, that is,* $W_m^S>W_m^*$.

Corollary 8 may initially seem surprising in that even when the physician orders more tests than in the social optimum, 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

$$p_m^S=Q(\mu_m^S)-\omega W_m^S=\frac{(\alpha+2d)(Q_c+\alpha\mu_c)}{2(\alpha+d)}-\sqrt{\omega(\alpha+d)}. \tag{2}$$

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

$$\pi+\beta(p_m^*-\pi)=\frac{(\alpha+2\beta d)(Q_c+\alpha\mu_c)+\alpha\pi(1-\beta)}{2(\alpha+\beta d)}-\sqrt{\omega(\alpha+\beta d)}. \tag{3}$$

Recall from Corollary 7 that when $\pi>(Q_c+\alpha\mu_c)/(1+\alpha/d)$, the physician overtests. In the meantime, comparing (2) and (3) gives that $\pi+\beta(p_m^*-\pi)>p_m^S$, meaning each patient is subject to a high out-of-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 in Hong Kong's healthcare system, where the average waiting times across public and private hospitals are significantly different: the average waiting time for a public-hospital physician is 74.7 days, whereas it is 24.3 days for a private physician (Harvard Team 1999).

4. Patient Heterogeneity in Diagnostic Precision

In this section, we consider the possibility that the physician orders a different number of tests for patients with different levels of diagnostic certainty. We start with a description of our model, which is a generalization of the baseline model in Section 2. We then characterize the market equilibrium, the social optimum, and the condition for overtesting, followed by a numerical study. We close this section with a discussion of follow-up visits using the model.

4.1. Modeling Patient Heterogeneity

Two types of patients, indexed by $i=H,L$, exist such that the diagnostic precision α can be either high (α_H) or low (α_L), where $\alpha_H>\alpha_L$; that is, the patients are heterogeneous with respect to the rate α at which the service quality improves when the service rate decreases. A given number of diagnostic tests result in higher diagnostic certainty for a type H patient than for a type L patient; that is, $Q_H(\mu)>Q_L(\mu)$ for all μ. Both types of patients have the same sensitivity to delay; that is, their waiting costs ω are identical. The probability that $\alpha=\alpha_i$ is q_i. Patients of class i arrive according to a Poisson process with rate λ_i (i.e., class i patients arrive according to a Poisson process with rate $\lambda_i=q_i\cdot\lambda$), where the total arrival rate of patients λ is a decision

variable. The total potential arrival rate Λ is, as in the baseline model, assumed to be sufficiently large such that full coverage is not possible.

Let μ_H, μ_L denote the service rates chosen for the two types of patients. The physician tries to achieve a common level of diagnostic certainty q across both types of patients. Hence, μ_H and μ_L are such that $q = Q_H(\mu_H) = Q_L(\mu_L)$; that is, $\mu_H = \alpha_L/\alpha_H \cdot \mu_L + (1 - \alpha_L/\alpha_H) \cdot \mu_c$. For a given quality level Q, the service rates are given by $\mu_i(Q) = \mu_c + Q_c/\alpha_i - Q/\alpha_i$ for $i = H, L$.

We assume the physician cannot discriminate among patients by adopting different admission policies or service fees based on the patients' conditions (dictated by the diagnostic precision). This assumption reflects the case in which the physician admits the same fraction of each patient type and thus controls the total arrival rate λ. Similar to the analysis of the base-case model in Section 2, we characterize and compare the optimal service allocations under various optimality criteria.

The physician chooses the service fee p, the total arrival rate λ, the service-rate vector $\boldsymbol{\mu} = (\mu_H, \mu_L)$, and the scheduling policy ϕ (which is now a decision variable). For simplicity, we assume that a first-come, first-served policy is obeyed among customers of the same class; the results extend to any nonanticipating and nonpreemptive regime because they all result in the same mean queueing times, and the patients' arrival decisions depend only on the mean time in the system. Let $W_i^{\phi}(\lambda, \boldsymbol{\mu})$ denote the operationally feasible steady-state expected waiting time in the system for class i patients under an admissible scheduling rule ϕ. In the market equilibrium, each patient chooses an individually optimal queue-joining probability. We assume the patients do not know the diagnostic certainty of their condition ex ante. Therefore, both types of patients choose the same probability of joining the service, and the following market-clearing condition is satisfied in equilibrium:

$$Q_i(\mu_i) = \pi + \beta(p - \pi) + \omega \sum_{i=H,L} q_i W_i^{\phi}(\lambda, \boldsymbol{\mu}) \quad \text{for } i = H, L. \tag{4}$$

4.2. Equilibrium Characterization

We now find the optimal scheduling policy $\phi \in \Phi$ from the physician's and social planner's perspectives. For a given total arrival rate λ and service-rate vector $\boldsymbol{\mu} = (\mu_H, \mu_L)$, the revenue rate and the expected net benefit are maximized whenever the system's expected delay-cost rate $\omega \sum_{i=H,L} q_i W_i^{\phi}(\lambda, \boldsymbol{\mu})$ is minimized. Therefore, we take minimizing the delay cost as our optimality criterion subject to operational feasibility. Then, the optimal scheduling policy to be followed is the shortest expected processing time policy, which gives strict (nonpreemptive) priority to the patient classes in decreasing order of their diagnostic precision α_i (see Online Appendix A for details). The resulting waiting time is given by

$$W_H(\lambda, \boldsymbol{\mu}) = \frac{\sum_{k=H,L} \lambda_k/\mu_k^2}{(1 - \lambda_L/\mu_L)(1 - \lambda_H/\mu_H)} + \frac{1}{\mu_H} \quad \text{and}$$

$$W_L(\lambda, \boldsymbol{\mu}) = \frac{\sum_{k=H,L} \lambda_k/\mu_k^2}{(1 - \lambda_H/\mu_H)(1 - \sum_{k=H,L} \lambda_k/\mu_k)} + \frac{1}{\mu_L}.$$

4.2.1. Market Equilibrium.

The physician's problem is to choose the total arrival rate λ, service-rate vector $\boldsymbol{\mu} = (\mu_H, \mu_L)$, and the price p to

$$\begin{aligned} \max_{\lambda, \mu_H, \mu_L, p} \quad & g(\lambda, \boldsymbol{\mu}) = p\lambda \\ \text{s.t.} \quad & Q_i(\mu_i) = \pi + \beta(p - \pi) \\ & \qquad + \omega[q_H W_H(\lambda, \boldsymbol{\mu}) + q_L W_L(\lambda, \boldsymbol{\mu})], \\ & \lambda_H/\mu_H + \lambda_L/\mu_L < 1, \\ & Q(\mu_H) = Q(\mu_L), \end{aligned}$$

where $\lambda_i = q_i\lambda$.

Note that we can rewrite the expected waiting time explicitly as follows:

$$\begin{aligned} W(\lambda, \boldsymbol{\mu}) = & \left(\frac{q_H/\mu_H^2 + q_L/\mu_L^2}{1/\lambda - q_H/\mu_H} \right) \\ & \cdot \left[q_H + \frac{q_L}{1 - \lambda(q_H/\mu_H + q_L/\mu_L)} \right] \\ & + \frac{q_H}{\mu_H} + \frac{q_L}{\mu_L}, \end{aligned}$$

which implies $W(\lambda, \boldsymbol{\mu})$ is strictly increasing in the total arrival rate λ. Therefore, for each price level p, the market-clearing condition uniquely defines an aggregate arrival rate $\lambda(\boldsymbol{\mu})$ as a function of the service rates. Substituting for p from the market-clearing condition, the physician's objective function can equivalently be stated as choosing $\lambda, \boldsymbol{\mu}$ to maximize

$$g(\lambda, \boldsymbol{\mu}) = \lambda \cdot \{Q_i(\mu_i) - \omega[q_H W_H(\lambda, \boldsymbol{\mu}) + q_L W_H(\lambda, \boldsymbol{\mu})] - \pi(1 - \beta)\}.$$

4.2.2. Social Optimum and Overtesting Condition.

The objective for the social planner is to maximize the expected net benefit received per unit of time by the collective of all patients:

$$\begin{aligned} \max_{\lambda, \mu_H, \mu_L} \quad & U(\lambda, \boldsymbol{\mu}) = \lambda_H Q_H(\mu_H) + \lambda_L Q_H(\mu_L) \\ & \qquad - \omega\lambda_H W_H(\lambda, \boldsymbol{\mu}) - \omega\lambda_L W_L(\lambda, \boldsymbol{\mu}) \\ \text{st.} \quad & \lambda_H/\mu_H + \lambda_L/\mu_L < 1, \\ & Q(\mu_H) = Q(\mu_L), \end{aligned}$$

where $\lambda_i = q_i\lambda$.

The following proposition summarizes the effect of diagnostic precision heterogeneity on the physician's overtesting behavior.

Proposition 6. *When service-rate differentiation is allowed under diagnostic precision heterogeneity, the physician orders more tests for both patient classes* ($\mu_i^* \le \mu_i^S$ *for* $i = H, L$) *and admits fewer patients in total* ($\lambda^* \le \lambda^S$) *in the market equilibrium than in the social optimum. Furthermore, when* ω *is sufficiently large, the average waiting time is lower in the market equilibrium than in the social optimum* ($W^* \le W^S$).

Proposition 6 states that with patient heterogeneity in diagnostic precision, largely consistent with the results from the baseline model, that is, Propositions 1 and 2 and Corollary 1(i)–(ii), the physician would always overtest. In addition, the equilibrium arrival rate would be lower than under the social equilibrium. Different from the baseline model, however, the physician would account for patient heterogeneity by choosing the service rate in such a way that the average waiting time may be lower than in the social optimum. (See Online Appendix A for the exact characterization of a sufficient condition for $W^* \le W^S$ to hold.) This result is similar to the one stated in Corollary 8.

4.3. Numerical Study

We conduct numerical experiments to provide insights into the effects of interpatient heterogeneity and the patient mix on the equilibrium behavior of agents.

We first study the effect of interpatient heterogeneity. Let $\Delta\alpha = \alpha_H - \alpha_L$ denote the range of the diagnostic precision of the high and low patient classes. We use the following parameters: $\mu_c = 0.8$, $Q_c = 5$, $\omega = 0.5$, $\pi = 10$, $\beta = 0.2$, and $q_H = q_L = 0.5$. We maintain the average diagnostic precision $(\alpha_H + \alpha_L)/2$ at 20 and vary $\Delta\alpha$, which allows us to examine the effect of interpatient heterogeneity; we include the case in which $\alpha_H = 20.01$ to demonstrate the discontinuity as a result of the transition from a homogeneous patient population (i.e., $\alpha_H = \alpha_L$) to a heterogeneous one. Table 1 shows that as $\Delta\alpha$ increases, in both the market equilibrium and social optimum, (1) the queueing system has a higher service-time variability in that the service rate for type H patients increases, whereas the service rate for the type L patients decreases; (2) the optimal service quality decreases; and (3) the total expected waiting time decreases. The reduction in quality comes from the fact that as interpatient heterogeneity increases, providing the same level of service quality requires a higher level of variability. Interestingly, the effect of diagnostic precision heterogeneity on the expected waiting time varies depending on patient type: in both the market equilibrium and the social optimum, as $\Delta\alpha$ increases, type H patients' expected waiting time decreases, whereas type L patients' expected waiting time first increases and then decreases.

Next, we study the effect of patient mix. We use the same set of parameters as above except that we vary (q_H, q_L) to examine the effect of the patient mix. In each generated scenario, we adjust the values of (α_H, α_L) in such a way that the average diagnostic precision $(\alpha_H + \alpha_L)/2$ is maintained at 20. Tables 2 and 3 provide the results for the cases in which $(q_H, q_L) = (0.3, 0.7)$ and $(q_H, q_L) = (0.7, 0.3)$, respectively. One observation from comparing these results is that as the proportion of type H patients increases (i.e., from $q = 0.3$ in Table 2 to $q_H = 0.5$ in Table 1 and then to $q_H = 0.7$ in Table 3), the service rates for both types increase, whereas the service quality decreases in both the market equilibrium and the social optimum. In addition, both types of patients experience a shorter expected waiting time.

4.4. Discussion on Follow-Up Tests

We now consider a variant of the above model with patient heterogeneity. There are two types of patients indexed by $i = H, L$, each accounting for q_i of the population and with different levels of diagnostic precision: high (α_H) and low (α_L). However, different from the above model, a proportion of the patients' types are not revealed until some preliminary tests have been ordered; we refer to these patients as type u patients. The physician provides a uniform service quality and uses two service rates: μ_H, corresponding to a "basic package" of tests, for all patients, and μ_L, corresponding to an "advanced package" of tests, for type L patients only. Each type L patient must complete the basic package before proceeding to the advanced package. Once a basic package is completed, all the type u patients' types are revealed: those who are type H do not need further tests, whereas those who are type L need to complete the advanced package.

Table 1. Effect of Patient Heterogeneity in Diagnostic Precision ($q_H = q_L = 0.5$)

		Market equilibrium							Social optimum						
α_H	α_L	μ_H^*	μ_L^*	Q^*	W_H^*	W_L^*	W^*	λ^*	μ_H^S	μ_L^S	Q^S	W_H^S	W_L^S	W^S	λ^S
20	20	0.3250	0.3250	14.50	6.324	6.324	6.324	0.1669	0.5250	0.5250	10.50	6.324	6.324	6.324	0.3669
20.01	19.99	0.3252	0.3247	14.50	5.200	7.449	6.3249	0.1669	0.5251	0.5248	10.50	3.950	8.699	6.325	0.3669
21	19	0.3504	0.3030	14.44	4.957	7.654	6.306	0.1680	0.5381	0.5106	10.50	3.891	8.757	6.324	0.3673
22	18	0.3809	0.2878	14.22	4.676	7.693	6.184	0.1685	0.5542	0.4995	10.41	3.814	8.753	6.283	0.3690
23	17	0.4137	0.2774	13.88	4.395	7.596	5.996	0.1675	0.5724	0.4921	10.24	3.725	8.688	6.207	0.3719
24	16	0.4465	0.2697	13.48	4.131	7.409	5.770	0.1645	0.5919	0.4879	9.993	3.630	8.569	6.100	0.3757
25	15	0.4780	0.2633	13.05	3.884	7.166	5.525	0.1589	0.6121	0.4868	9.698	3.533	8.406	5.969	0.3801

Table 2. Effect of Patient Heterogeneity in Diagnostic Precision ($q_H = 0.3, q_L = 0.7$)

		Market equilibrium							Social optimum						
α_H	α_L	μ_H^*	μ_L^*	Q^*	W_H^*	W_L^*	W^*	λ^*	μ_H^S	μ_L^S	Q^S	W_H^S	W_L^S	W^S	λ^S
20	20	0.3250	0.3250	14.50	6.324	6.324	6.324	0.1669	0.5250	0.5250	10.50	6.324	6.324	6.324	0.3669
20.01	19.996	0.3252	0.3249	14.50	4.942	6.917	6.325	0.1669	0.5251	0.5249	10.50	3.588	7.497	6.325	0.3669
21	19.57	0.3486	0.3156	14.48	4.727	7.002	6.320	0.1674	0.5380	0.5188	10.50	3.537	7.519	6.324	0.3670
22	19.14	0.3727	0.3089	14.40	4.523	7.030	6.278	0.1676	0.5514	0.5142	10.47	3.484	7.520	6.309	0.3676
23	18.71	0.3964	0.3040	14.28	4.341	7.017	6.214	0.1674	0.5648	0.5110	10.41	3.430	7.503	6.281	0.3686
24	18.71	0.4192	0.3002	14.14	4.179	6.977	6.137	0.1667	0.5782	0.5088	10.32	3.378	7.472	6.243	0.3698
25	17.86	0.4408	0.2970	13.98	4.036	6.918	6.053	0.1654	0.5912	0.5076	10.22	3.327	7.428	6.198	0.3713

Table 3. Effect of Patient Heterogeneity in Diagnostic Precision ($q_H = 0.7, q_L = 0.3$)

		Market equilibrium							Social optimum						
α_H	α_L	μ_H^*	μ_L^*	Q^*	W_H^*	W_L^*	W^*	λ^*	μ_H^S	μ_L^S	Q^S	W_H^S	W_L^S	W^S	λ^S
20	20	0.3250	0.3250	14.50	6.324	6.324	6.324	0.1669	0.5250	0.5250	10.50	6.324	6.324	6.324	0.3669
20.01	19.98	0.3252	0.3244	14.50	5.542	8.154	6.325	0.1669	0.5251	0.5246	10.50	4.510	10.560	6.325	0.3669
21	17.67	0.3572	0.2736	14.30	5.205	8.640	6.236	0.1693	0.5396	0.4904	10.47	4.430	10.72	6.318	0.3684
22	15.33	0.4118	0.2431	13.54	4.622	8.483	5.780	0.1696	0.5656	0.4637	10.16	4.269	10.655	6.185	0.3744
23	13	0.4753	0.2256	12.47	3.966	7.787	5.113	0.1587	0.6024	0.4505	9.544	4.040	10.28	5.912	0.3856
24	10.67	0.5372	0.2086	11.31	3.256	6.917	4.354	0.1252	0.6457	0.4528	8.703	3.774	9.624	5.530	0.4007
25	8.33	0.6027	0.2081	9.933	2.661	6.161	3.711	0.1000	0.6907	0.4720	7.733	3.497	8.748	5.072	0.4190

The requirement for uniform service quality (denoted by Q) gives

$$\begin{aligned} Q &= Q_c + \alpha_H(\mu_c - \mu_H) \\ &= Q_c + \alpha_L\left(\mu_c - \frac{1}{1/\mu_H + 1/\mu_L}\right), \end{aligned} \tag{5}$$

which gives

$$\mu_H(Q) = \frac{Q_c + \alpha_H\mu_c - Q}{\alpha_H} \quad \text{and}$$
$$\mu_L(Q) = \frac{(Q_c + \alpha_H\mu_c - Q)(Q_c + \alpha_L\mu_c - Q)}{(Q - Q_c)(\alpha_H - \alpha_L)}.$$

We can verify $\mu_H(Q)$ decreases in Q. For $\mu_L(Q)$ to decrease in Q, we would need $Q < Q_c + \sqrt{\alpha_H\alpha_L}\mu_c$, which is satisfied because we have from (5) that $Q < Q_c + \min\{\alpha_H, \alpha_L\}\mu_c$.

To ensure $\mu_L(Q) < \mu_H(Q)$, we would need $\alpha_H/\alpha_L > 2 + \alpha_L\mu_c/(Q - Q_c)$. Under the assumption, the queueing system with follow-up tests may be viewed as a queueing system with two classes of customers with service rates $\mu_H(Q)$ and $\mu_L(Q)$, respectively, where $\mu_L(Q) < \mu_H(Q)$ and $\mu_i'(Q) < 0$ for $i = H, L$. The proportions of these two classes of patients are $q_H' = 1/(1 + q_L)$ and $q_L' = q_L/(1 + q_L)$, respectively. Under the optimal scheduling rule, the second class of customers has a higher priority in the queueing discipline. We can then show that Lemmas A2–A6 (see the online appendices) remain valid and replicate the result in Proposition 6.

5. Disparities in Health Insurance Coverage

We have so far focused on the case where there is only one type of patient, with a copayment of π and a coinsurance rate of β. We now extend our baseline model by allowing a proportion γ of the population to have a new type of insurance plan characterized by a copayment of π' and a coinsurance rate of β' that satisfy the following: (1) $\beta' \geq \beta$, meaning the new type of patients' coinsurance rate is no lower than that of the original type; and (2) $(1 - \beta')\pi' \geq (1 - \beta)\pi$, meaning the new type of patients' residual copayment—the copayment less the proportion covered through coinsurance—is no lower than that of the original type. Note that one of the above two inequalities must be strict; otherwise, the new type would be exactly the same as the original type. Thus, for any service charge, the new type has a higher out-of-pocket amount. As in the case with patient heterogeneity in diagnostic precision (see Section 4), the physician attempts to achieve a common level of diagnostic certainty across patients and thus chooses a uniform service rate.

The following proposition provides the optimal service rate. For simplicity of exposition, we define $\hat{\mu} = [Q_c + \alpha\mu_c - (1 - \beta)\pi]/(2\alpha)$ as the optimal service rate when all the patients are of the original type, and $\hat{\mu}' = [Q_c + \alpha\mu_c - (1 - \beta')\pi']/(2\alpha)$ as the optimal service rate when all the patients are of the new type; it is clear that $\hat{\mu}' < \hat{\mu}$. We focus on the setting in which $\Lambda > \hat{\mu}' - \sqrt{\omega/\alpha}$;

that is, the total potential arrival rate is large enough, so that not all patients will be covered in equilibrium.

Proposition 7. *Under two types of insurance coverage, the optimal service rate is*

$$\mu^* = \begin{cases} \hat{\mu} & \text{if } 0 < \hat{\mu} - \sqrt{\omega/\alpha} \le (1-\gamma)\Lambda, \\ \hat{\mu}' & \text{otherwise}, \end{cases}$$

which is always lower than the socially optimal service rate $\mu^{SO} = (Q_c + \alpha\mu_c)/(2\alpha)$.

If γ, the proportion of patients with higher out-of-pocket expenses, is below 1/2, an increased γ indicates that the patients' health insurance coverage becomes more heterogeneous. From Proposition 7, we observe that as γ increases, the physician is more likely to choose a lower service rate, which further increases the efficiency gap between the market equilibrium and the social optimum. Broadly speaking, this result is aligned with the industry insight from the outpatient setting that the shift to a patient population with more diversified insurance coverage (reflected in more insurance types accepted by healthcare providers) requires efforts in expanding capacity (Levine et al. 2013).

We can generalize the above analysis to the scenario with $N \ge 2$ types of patients, and we relegate the details to Online Appendix A.

6. Implications for Policy Makers

We highlight implications from our results to policy makers.

First, the insurance structure (as opposed to the share of patient cost sharing) drives the intensity of testing. In our baseline model, we show that the copayment and the coinsurance components have differential effects on testing decisions: the intensity of testing increases in the copayment but decreases in the coinsurance rate (Corollary 2). The effect of the copayment and the coinsurance components can also be similar in certain cases; for example, in the presence of a price-setting reimbursement ceiling (Corollary 4) or significant misdiagnosis concerns (Corollary 7), increasing the copayment or the coinsurance rate leads to higher testing intensity.

Second, overtesting is a complex phenomenon, and a comprehensive understanding of the underlying financial, operational, and clinical drives is essential before embarking on any radical changes in public policy. Simple changes in the payment scheme, such as imposing a reimbursement ceiling or eliminating insurance coverage all at once, may not work as intended. In addition, we show that overtesting does not occur in the absence of a positive copayment (typically charged for physician visits), but eliminating copayments altogether may lead to undertesting, as suggested by Corollary 5 (the case with a reimbursement ceiling) and Corollary 6 (the case with misdiagnosis concerns).

Third, physicians' misdiagnosis concerns lead to overtesting only when bundled together with a certain incentive environment. Corollary 7 indicates that addressing the issue of overtesting (and undertesting) requires incorporating physicians' misdiagnosis concerns as one factor in designing the insurance structure.

Fourth, in healthcare settings, it is often the case that the physician solely decides on the service quality (i.e., the level of diagnostic accuracy), that the physician wishes to uniformly deliver to all customers. Subsequently, because of the inherent uncertainty in the process, the service rate varies across patients. We show that in serving patients of different levels of diagnostic precision, overtesting nevertheless occurs when compared to the social optimum (Proposition 6). Furthermore, as the patient mix becomes more diverse, to provide the same service quality entails longer waiting times because of increased system variability; thus, as our numerical study demonstrates, the optimal service quality decreases.

Last, when the patient population is characterized by an increased number of insurance types, we find the service rate tends to be lower and the gap between the market equilibrium and the social optimum widens, because a strictly positive surplus is required for certain patients (Proposition 7).

7. Concluding Remarks

This paper was initially motivated by an observational study in an ocular imaging setting. In the case of laboratory testing, as one would expect, many tests may be processed in parallel, and thus test ordering has a limited effect on patients' waiting times. In an imaging-testing environment, quite differently, the tests require patient presence and thus directly drive patients' waiting times. Thus, imaging testing provides a compelling venue for us to examine the effect of the physician's clinical decision making on a healthcare organization's operational, financial, as well as clinical performance. Moreover, our analysis and results may provide insights into physicians' decisions concerning other services with similar trade-offs.

To the best of our knowledge, this work is the first to analytically investigate financial (insurance and reimbursement), operational (system throughput and congestion), and clinical (service quality) incentives behind physicians' test-ordering behavior in an outpatient setting. Our approach has a similar spirit as expressed in a commentary by Eisenberg et al. (1987, p. 805):

> We herein assume that physicians respond to financial incentives provided by different payment schemes

> and to changes in those incentives. This is not to say that financial incentives are of primary importance to physicians. The primary goal of physicians is to ensure their patients' health . . . . Substantial variability among physicians in test ordering suggests, however, that clinical indications for diagnostic tests are rarely absolute and that other factors may influence physicians' ordering decisions. Some reasons for variation reflect the physician's role as the patient's agent and advocate. Patients may want tests done, and their desire for testing may be influenced by insurance coverage. Physicians not only respond to their patients' wishes, but also act on their patients' behalf by considering convenience and out-of-pocket costs.

Our model reveals that insurance coverage is a key driver of overtesting, and the copayment and the coinsurance rate affect the equilibrium service rate in opposite ways: with a higher copayment, the physician orders more tests; with a higher coinsurance rate, the physician orders fewer tests. We then show that setting a reimbursement ceiling alone cannot eliminate overtesting, and, surprisingly, overtesting can still occur even when such a ceiling is low. Furthermore, when physicians are concerned about inaccurate diagnosis, we show that both overtesting and undertesting are possible outcomes, and the waiting time in equilibrium is shorter than the socially efficient level. We also consider two extensions of our baseline model: one on patient heterogeneity in diagnostic precision, and the other on disparities in health insurance coverage.

Acknowledgments

This paper is based on Chapter 1 of Tinglong Dai's Ph.D. dissertation, titled "Incentives in U.S. Healthcare Operations," at Tepper School of Business, Carnegie Mellon University. A previous version of this paper received the first place for the 2012 Production and Operations Management Society Best Healthcare Paper Award. The authors thank Dr. Joshua Rheinbolt, Dr. Robert J. Noecker, and the University of Pittsburgh Medical Center Eye Center for collaborating in a patient-experience improvement project that motivated this study, for providing institutional details regarding physician decision making, and for offering experiential feedback for the models and results. The authors also benefited from the comments by Barış Ata, David Axelrod, Maqbool Dada, Laurens Debo, Kevin Frick, Refael Hassin, Kinshuk Jerath, Judith Lave, Phillip Phan, Lawrence Robinson, Alan Scheller-Wolf, Rachna Shah, Senthil Veeraraghavan, and Ruizhi Wang. The authors appreciate the helpful comments from seminar participants at the University of California, Berkeley, Carnegie Mellon University, Cornell University, the University of Hong Kong, Johns Hopkins University, the University of Minnesota, Stevens Institute of Technology, the University of Utah, and Yale University and session participants at the 2010–2012 INFORMS Annual Meetings, 2011 MSOM Conference Healthcare Special Interest Group, 2012 POMS Annual Meeting, and 2012 Trans-Atlantic Doctoral Conference at London Business School. Last, the authors appreciate the constructive comments from Stephen Graves, Christopher Tang, the associate editor, and three anonymous reviewers that helped improved this paper immensely.

References

Alizamir S, de Véricourt F, Sun P (2013) Diagnostic accuracy under congestion. *Management Sci.* 59(1):157–171.

Anand KS, Paç MF, Veeraraghavan SK (2011) Quality-speed conundrum: tradeoffs in customer-intensive services. *Management Sci.* 57(1):40–56.

Baicker K, Fisher ES, Chandra A (2007) Malpractice liability costs and the practice of medicine in the Medicare program. *Health Affairs* 26(3):841–852.

Carrier ER, Reschovsky JD, Mello MM, Mayrell RC, Katz D (2010) Physicians' fears of malpractice lawsuits are not assuaged by tort reforms. *Health Affairs* 29(9):1585–1592.

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

Dada M, White WD (1999) Evaluating financial risk in the Medicare prospective payment system. *Management Sci.* 45(3):316–329.

Dai T, Noecker RJ, Rheinbolt J, Tayur S (2009) Personal communication, meeting at Tepper School of Business, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Dai T, Tayur S, Rheinbolt J, Noecker RJ (2012) Patient experience improvement at UPMC Eye Center. Teaching Case, Tepper School of Business, Carnegie Mellon University, Pittsburgh.

Danzon PM (1982) *Economic Factors in the Use of Laboratory Tests by Office-Based Physicians* (RAND Corporation, Santa Monica, CA).

Debo L, Veeraraghavan SK (2014) Equilibrium in queues under unknown service times and service value. *Oper. Res.* 62(1): 38–57.

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

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

Eisenberg JM, Myers LP, Pauly MV (1987) How will changes in physician payment by Medicare influence laboratory testing? *J. Amer. Med. Assoc.* 258(6):803–808.

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

Gaynor M, Town RJ (2012) Competition in health care markets. Pauly MV, McGuire TG, Barros PP, eds. *Handbook of Health Economics*, Vol. 2 (North-Holland, Oxford, UK), 499–637.

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

Green LV (2006) Queueing analysis in healthcare. Hall RW, ed. *Patient Flow: Reducing Delay in Healthcare Delivery* (Springer, New York), 281–307.

Gupta D, Mehrotra M (2015) Bundled payments for healthcare services: Proposer selection and information sharing. *Oper. Res.* 63(4):772–788.

Gawande A (2009) The cost conundrum. *New Yorker* (June 1), 36–44.

Gawande A (2015) Overkill. *New Yorker* (May 11), 44–55.

Harvard Team (1999) Improving Hong Kong's health care system: Why and for whom? Report, Health and Welfare Bureau, Government of Hong Kong SAR. Accessed May 31, 2016, http://www.fhb.gov.hk/en/press_and_publications/consultation/HCS.HTM.

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

Hopp WJ, Lovejoy WS (2012) *Hospital Operations: Principles of High Efficiency Health Care* (FT Press, Upper Saddle River, NJ).

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

Kessler DP, McClellan MB (2002) How liability law affects medical productivity. *J. Health Econom.* 21(6):931–955.

Kissick W (1994) *Medicine's Dilemmas: Infinite Needs versus Finite Resources* (Yale University Press, New Haven, CT).

Kostami V, Rajagopalan S (2013) Speed-quality trade-offs in a dynamic model. *Manufacturing Service Oper. Management* 16(1): 104–118.

Levine E, Bauman N, Garrett B (2013) The impact of coverage shifts on hospital utilization. Report, McKinsey & Company. Accessed May 31, 2016, http://healthcare.mckinsey.com/sites/default/files/793546_Coverage_Shifts_on_Hospital_Utilization.pdf.

Martin S, Smith PC (2003) Using panel methods to model waiting times for national health service surgery. *J. Royal Statist. Soc.: Series A* 166(3):369–387.

Mold JW, Hamm RM, McCarthy LH (2010) The law of diminishing returns in clinical medicine: How much risk reduction is enough? *J. Amer. Board Family Med.* 23(3):371–375.

Newhouse JP (1978) *Insurance Benefits, Out-of-Pocket Payments, and the Demand for Medical Care: A Review of the Literature* (RAND Corporation, Santa Monica, CA).

Newhouse JP (2002) *Pricing the Priceless: A Health Care Conundrum* (MIT Press, Cambridge, MA).

Paç MF, Veeraraghavan S (2015) False diagnosis and overtreatment in services. Working paper, University of Pennsylvania, Philadelphia.

Pauly MV (1980) *Doctors and Their Workshops: Economic Models of Physician Behavior* (University of Chicago Press, Chicago).

Phelps CE, Newhouse JP (1974) Coinsurance, the price of time, and the demand for medical services. *Rev. Econom. Statist.* 56(3): 334–342.

Pinker EJ (2012) Can OR/MS be a change agent in healthcare? *Manufacturing Service Oper. Management* 14(4):495–511.

Rao VM, Levin DC (2012) The overuse of diagnostic imaging and the choosing wisely initiative. *Ann. Intern. Med.* 157(8): 574–576.

Shumsky RA, Pinker EJ (2003) Gatekeepers and referrals in services. *Management Sci.* 49(7):839–856.

So KC, Tang CS (2000) Modeling the impact of an outcome-oriented reimbursement policy on clinic, patients, and pharmaceutical firms. *Management Sci.* 46(7):875–892.

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

Studdert DM, Mello MM, Gawande AA, Gandhi TK, Kachalia A, Yoon C, Puopolo AL, Brennan TA (2006) Claims, errors, and compensation payments in medical malpractice litigation. *New England J. Med.* 354(19):2024–2033.

Tong C, Rajagopalan S (2014) Pricing and operational performance in discretionary services. *Production Oper. Management* 23(4): 689–703.

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

Yip WC (1998) Physician response to Medicare fee reductions: Changes in the volume of coronary artery bypass graft (CABG) surgeries in the Medicare and private sectors. *J. Health Econom.* 17(6):675–699.

Young DW, Saltman RB (1985) *The Hospital Power Equilibrium: Physician Behavior and Cost Control* (Johns Hopkins University Press, Baltimore).