PHYSICIAN INCENTIVES AND TREATMENT CHOICE*

KEVIN E. PFLUM†

ABSTRACT. I analyze how the countervailing incentives the insurer and patient place on physicians impact treatment selection. Patients have preferences over treatments based on their type and are responsive to physician treatment practices. This response induces inefficient choices by physicians that cannot always be overcome with supply-side payment rules alone. First-best may not be achievable when the demand response to physician treatment practices is low relative to the number of physicians in the market. Increasing the treatment practice elasticity of demand can improve efficiency. Lastly, I explore how noisy signals of illness type and diagnostic testing complicate the problem.

1. INTRODUCTION

The moral hazard and information asymmetries that characterize health markets have long posed a challenge in designing insurance contracts that limit over-utilization. The literature has explored accomplishing this through a variety of demand-side and supply-side means; however, the problem of over-utilization ultimately can occur only if physicians prescribe more than the efficient level of treatment.[1] Indeed, there has long been concern that physicians may be able to use their information superiority to prescribe not only more treatment than is necessary, but to prescribe more treatment than the patient would request if he had perfect information; a problem commonly referred to as physician-induced demand (PID). That insurers can exert some control over a physician's treatment choice by utilizing payments that work to align the physicians' incentives with its own is well known (See McGuire (2000) and Léger (2008) for detailed

Date: December 18, 2012.

*I am grateful to Paul Pecorino, Paan Jindapon, Paula Cordero Salas, Paul J. Healy, Suhui Li and participants of the Ohio State University theory workshop and the 2012 Annual Conference of the Southern Economic Association for helpful discussions and comments.

†Address: Rm 245 Alston Hall, Stadium Drive, Tuscaloosa, AL, 35487. Email: kpflum@cba.ua.edu.

[1]It is well known, of course, that demand-side cost sharing can reduce moral hazard (Pauly, 1968; Zeckhauser, 1970); though in practice, this approach is not without problem (e.g., Wong et al. (2001); Tamblyn et al. (2001); Trivedi et al. (2008)).

reviews of the literature); nevertheless, one overlooked limitation to controlling medical use through physicians is that patients may seek second opinions or switch providers if they are sufficiently unhappy with their physician's treatment recommendation (Givens, 1957; Macpherson et al., 2001; Berry, 2007). In consequence, if enough patients are responsive to physician treatment practices, then competition—often seen as a way to produce higher quality care at lower costs—may instead exacerbate moral hazard by more strongly aligning physicians' incentives with patients' preferences.

The objective of this study is to explore how competitive pressure among physicians impacts physician treatment selection and to identify how and when supply-side payment rules can overcome these incentives to induce physicians to follow an insurer's preferred treatment practice. Much of the literature examining optimal insurance and physician payment rules consider scenarios in which a physician chooses a quantity or "intensity" of medical care to provide a patient for a particular diagnosis or illness (examples include Rochaix (1989), Selden (1990), Ellis and McGuire (1986, 1990), Ma and McGuire (1997), Gal-Or (1999) and more recently Choné and Ma (2011)). Although appropriate for many settings, modeling treatment as a continuous quantity overlooks the fact that treatment decisions are often a discrete choice among treatments generating discontinuous jumps in both the expected benefits and costs. Furthermore, the cost of providing the same quantity of treatment may differ between patients having different illness severities. To represent this aspect of medical treatment choice physicians in the current model must choose one of two discrete treatments. Patients are heterogeneous in their severity of illness or "type" and the cost and benefit of each treatment is dependent on their type so that it is more efficient to use one treatment over the other for each type and the more efficient treatment varies with type. For example, surgery may be a more appropriate treatment choice for some patients diagnosed with a particular cancer while a less invasive treatment option is more appropriate for others.

Patients have preferences over treatments that depend on their type and out-of-pocket costs while insurance creates a wedge between the private and social benefit of each treatment by shielding patients from the full cost. Because of the discrete nature of the treatment and the

differences in their benefits, insurance does not drive all patients to demand the most costly treatment, even if there is no difference in the patient's out-of-pocket costs congruent with the argument made by Mendel et al. (2012) that patients are not likely to demand, or alternatively accept, the recommendation for a more costly procedure such as surgery simply because they are insured. Physicians have a preference over treatments that depends on their cost of care, their compensation, and patient demand. Because patients can choose a physician based on their treatment practice physicians gain market share when their treatment practice more closely matches patient preferences vis-á-vis other physicians. The insurer's challenge is to design its payment rules in such a way as to overcome this competitive pressure and induce physicians to follow its preferred treatment practice.

An insurer's ability to induce its preferred treatment practice through simple payment rules depends on how responsive demand is to treatment practices relative to the number of physicians competing in the market. For a monopolist physician the insurer is limited only by the difference in costs between treatment options: first-best may not be attained when the differences in treatment costs are too large, in which case the less costly treatment will be used more frequently in the second-best. Otherwise the insurer can induce the physician to internalize the social gain from the treatments. Competition increases the disparity between the size of the payments because of the need to counteract the physicians' incentive to follow a treatment practice that is preferred by patients in order to attract or retain them. If patients are insufficiently responsive relative to the number of physicians in the market, then an insurer will not have the ability to induce some preferred treatment practice as the insurer cannot simultaneously prevent physicians from dumping some high-cost patients by utilizing a lower-cost treatment and from gaining market share by utilizing more of the higher-cost treatment. The availability of diagnostic testing further complicates the insurer's problem as a physician's private gain from the test may be too large to overcome through the payment mechanism resulting in more testing than socially optimal.

The literature on optimal insurance under physician agency is substantial;[2] however, this paper shares many features with Dranove (1988), Chernew, Encinosa and Hirth (2000) and Liu and Ma (2011). Dranove (1988) explores how PID can arise in equilibrium even when patients know the physician's recommendation strategy. Dranove also explores how physician competition will impact the degree of PID similarly showing that if patients are responsive to physicians' recommendation strategies, then the physicians' incentives will be more aligned with the patients' reducing PID. However, the focus of the study is on how PID can arise in equilibrium and not on how an insurer can use payments to alter the physician's treatment decision. Similar to the current model, Chernew et al. (2000) and Liu and Ma (2011) also model the physician's treatment decision as a discrete choice between multiple treatment options. Chernew et al. (2000) are primarily interested in the relationship between patients and the insurer, however, so focus on the use of demand-side cost-sharing rules (copays) to induce patients to select the appropriate treatment. Liu and Ma (2011) are interested in supply-side cost sharing rules that may induce a physician to choose the correct treatment plan—where the outcome of each treatment is either a success or a failure—and the physician can move to the next treatment when one is found not to work. Neither paper considers the effect physician competition may have on a physician's treatment choice; nor do either consider the effect of diagnostic testing on those payment rules.

There are also a few papers which study the effect of physician competition on treatment choice and insurance. In a seminal paper on optimal insurance, Ma and McGuire (1997) consider the effect of physician competition in an extension to their main model. The authors derive the optimal health insurance and physician payment plans in a setting where a patient selects a quantity of treatment after observing a physician's choice of effort, a vertical quality characteristic of the physician's services. Ma and McGuire found that competition increases effort if

[2]Gaynor (1994) and McGuire (2000) provide detailed reviews of much of this literature. Gaynor (1994) emphasizes research that examines physician-insurer agency whereas McGuire (2000) places an emphasis on research examining PID. Léger (2008) provides a review of the more recent literature also focusing on PID and the issue of over-treatment.

treatment quantity and effort are substitutes. Otherwise, when quantity and effort are complements, competition does not add anything to the insurer's problem since it can already induce the optimal level of effort through the payment mechanism.

Allard, Léger and Rochaix (2009) utilize a model similar to Ma and McGuire (1997) to examine physician competition in a dynamic game. As with Ma and McGuire (1997) the patient chooses some quantity of treatment but, in contrast to Ma and McGuire (1997), they do so simultaneously with the physicians's choice of effort. Patients can infer the physician's level of effort ex post based on the quantity of health care chosen and their health outcome. The authors found that because patients can switch physicians, the physicians will always choose some minimal level of effort, and, under the right circumstances, competition may lead physicians to provide the insurer's desired level of effort.

In short, these studies find that competition generally serves to *better* align the physicians' incentives with the insurer. In contrast I find that under the right circumstances competition can compound the insurer's problem by more strongly aligning the physicians' incentives with the patients'. This result bears some resemblance to a result by Ellis (1998) in which providers overprovide care to low-severity patients in order to attract them and "cream-skim." In the current model physicians may want to increase demand from profitable patients by using the patients' preferred treatment on more types relative to other physicians or decrease demand from less profitable patients by choosing a less preferred treatment option. In both models physicians adjust their treatment in order to manipulate their demand; though, in the current model this manipulation is done vis-á-vis the treatment practice of other physicians in the market.

The rest of the paper is organized as follows. Section 2 introduces the model and derives the first-best when there is no diagnostic testing. Section 3 derives the optimal payment rules for both a monopoly physician and for the case of competing physicians. Section 4 analyzes how diagnostic testing affects the physicians' treatment practices and the insurer's payment rules. Finally, section 5 ends with some concluding remarks. All proofs are in the appendix.

2. A BASIC MODEL

Consider a market for a particular disease or ailment for which patients would like treatment. Because the probability of becoming ill just scales the premium, the analysis is simplified by assuming that every patient in the market becomes ill. A patient's illness type is the realization of the random variable $\theta \in [\underline{\theta}, \overline{\theta}] = \Theta$ from distribution F with density $f > 0$ satisfying the monotone hazard rate property, $d\{F(\theta)/f(\theta)\}/d\theta > 0 > d\{(1 - F(\theta))/f(\theta)\}\ \forall\theta \in \Theta$. I initially assume that a patient's type is observable by both the physician and the patient but not the insurer.[3] In Section 4 I relax the assumption that the patient's type is perfectly observed by the patient and physician and introduce a costly diagnostic test that reveals the type. The distribution of illness severities and its properties are common knowledge.

Physicians have available two treatments T_1 and T_2 that can be provided at a cost of $c_k(\theta)$, $k \in \{1, 2\}$, where c_k is strictly increasing and $\mathcal{C}^2$. The cost of treatment reflects all of the physician's opportunity costs of providing the treatment. Furthermore, the increase in cost by patient type captures the notion that a physician may have to perform more procedures with a particular treatment for sicker patients for the same illness or that sicker patients are more likely to have a negative outcome triggering a costly malpractice lawsuit. The physician's costs are known by the insurer but unverifiable reflecting the notion that they represent opportunity costs and not accounting costs.[4]

The physician's utility, or profit, from treating a patient with illness severity θ takes the simple representation $R(\theta) - c(\theta)$, where $R(\theta)$ is some payment that may or may not be dependent on a patient's type. The physician's utility is thus assumed to be linear in payment and the physician

[3] Instead of knowing their illness severity patients could receive a signal which is correlated with their true severity. The qualitative results of the paper hold as long as the signal reduces the set of possible severities. If, however, the patients' signals do not reduce the set of possible illness severities, then the competitive case will be identical to the monopoly case. This is discussed further at the end of Section 3.

[4] If the costs are verifiable such that they can be directly contracted upon, then the insurer's problem is trivial as the physician's cost identifies the patient's type. There are two ways in which this triviality can be avoided. First, for each discrete treatment the physician could still provide different intensities resulting in different possible costs for a given type and treatment choice. Or, second, the insurer could have some uncertainty regarding a physician's cost of providing care. Both cases have been explored to varying degrees in the literature on insurance as well as procurement. What is important in the problem of discrete treatment selection are the differences in costs for the same illness severities so to focus on this I make the simplifying assumption that costs cannot be directly contracted upon.

is not capacity constrained (i.e., there is no marginal utility from leisure). This eliminates the possibility of income effects; however, since the focus of the paper is on how physicians choose between treatments and not on how the physician chooses an overall quantity of treatment the utility function is not restrictive. Moreover, the exclusion of income effects is more appropriate if the physician represents a physicians group or hospital.

Given a patient's type θ and the physician's treatment choice, T_k, the patient's benefit from treatment is the realization of a random variable μ having conditional distribution $H(\cdot \mid \theta, T_k)$; i.e., a patient's benefit from treatment depends on his illness type and treatment. A patient's expected benefit from treatment can be expressed as

$$\psi(\theta_i, T_k) \equiv \mathbb{E}[\mu \mid \theta_i, T_k] = \int \mu \, dH(\mu \mid \theta, T_k), \ \forall i \text{ and } k \in \{1, 2\}. \tag{1}$$

Because higher types are more severely ill they will benefit more from treatment; i.e., $d\psi/d\theta \geq 0$ for all treatments T_k.

A patient's indirect utility is composed of an additively separable monetary and health component and patient i's utility from treatment can be expressed as:

$$V_i = U(Y - P) - L(\theta_i) + \psi(\theta_i, T_k), \ k \in \{1, 2\}, \tag{2}$$

where Y is expenditures on other goods and services, P is the health insurance premium, and $L(\theta_i)$ is patient i's loss in utility upon falling ill. The function U is strictly increasing, strictly concave and $\lim_{x \downarrow 0} U'(x) \to \infty$. L is continuous and strictly increasing in the severity of illness.

My focus is on how an insurer can induce a physician to make the socially optimal treatment choice when more expensive treatments do not necessarily leave the patient better off. To ensure that there exists the ability to choose sub-optimal treatment options I impose the following conditions for the benefit and cost of treatment:

A1: $\psi(\underline{\theta}, T_1) - c_1(\underline{\theta}) > \psi(\underline{\theta}, T_2) - c_2(\underline{\theta})$,

A2: $\psi(\overline{\theta}, T_1) - c_1(\overline{\theta}) < \psi(\overline{\theta}, T_2) - c_2(\overline{\theta})$,

A3: $d\{\psi(\theta, T_1) - \psi(\theta, T_2)\}/d\theta \leq 0$ for all $\theta \in \Theta$, and

A4: $d\{c_1(\theta) - c_2(\theta)\}/d\theta \geq 0$ for all $\theta \in \Theta$.

A1 establishes that the social value of treating the lowest type with T_1 is higher than the social value of treating the lowest type with T_2. Similarly A2 establishes that the social value of treating the highest type with T_2 is higher than the social value of treating the highest type with T_1. A3 and A4 are regularity conditions establishing that patients' preferences over treatments as well as the social value of the treatments are both monotone over types. That is, let θ^P solve $\psi(\theta^P, T_1) = \psi(\theta^P, T_2)$ then A3 indicates that all types below θ^P will prefer treatment T_1 and all types above θ^P will prefer treatment T_2. Similarly, A3 and A4 together indicate that if there is a type θ^E such that $\psi(\theta^E, T_1) - c_1(\theta^E) = \psi(\theta^E, T_2) - c_2(\theta^E)$ then it is socially optimal to treat all types below θ^E with T_1 and all types above θ^E with T_2. A1 and A2 guarantee that θ^E exists.[5] Additionally, A4 guarantees that a physician's optimization problem will be well behaved.[6]

Several additional remarks can be made regarding assumptions A1–A4. First, note that the results of the model would hold if instead of assumption A4 the cost of treatment was invariant to the patient's type and physicians' received some utility from the patients' health benefit similar to Ellis and McGuire (1986), Chalkley and Malcomson (1998), Jack (2005) and Choné and Ma (2011) as the utility from patient benefit would continue to generate a monotone difference in opportunity costs for the physician.[7] Second, the model could be generalized to allow for m treatments as long as the additional treatments include assumptions similar to A3 and A4 in which there is a monotonic difference in the social value of the treatments which allows for an ordering of the efficient treatments. Lastly, as patients are not exposed to the total cost of

[5]Note that these assumptions are not sufficient to insure that θ^P exists. However, because θ^E is guaranteed to exist when there does not exist any $\theta^P \in \Theta$ such that $\psi(\theta^P, T_1) = \psi(\theta^P, T_2)$, then it will be the case that $\theta^P = \underline{\theta}$ and all patients prefer the more expensive treatment T_2.

[6]To prove sufficiency in the insurer's second best optimization program (Proposition 3) it must be the case that $c_1''(\cdot)$ is not substantially larger than $c_2''(\cdot)$. This condition is trivially satisfied for linear cost functions. However, as this is the only case in which the restriction is needed, to maintain generality, I do not impose it. I discuss what is needed for sufficiency in the proof for Proposition 3.

[7]In the latter two papers the authors consider the optimal contracts for physicians exhibiting unknown levels of altruism.

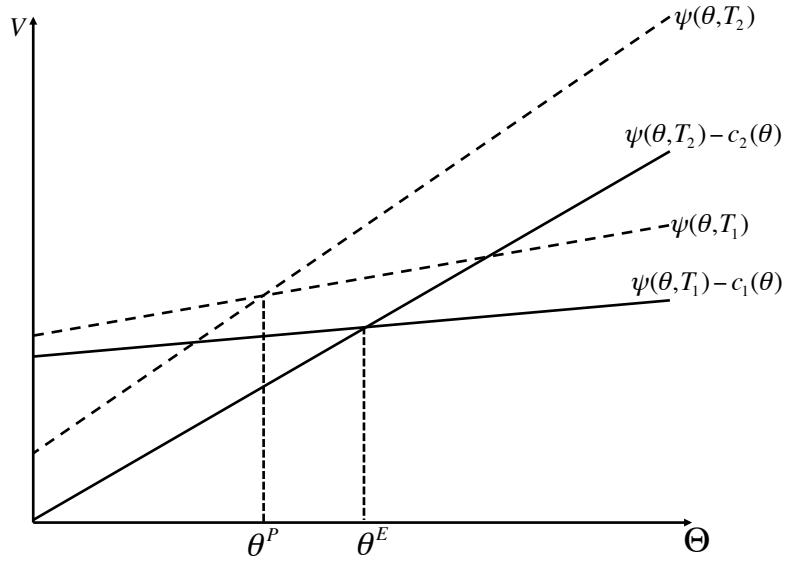


FIGURE 1. Illustration of the conflict between the preferences of the social planner and patients using linear cost and benefit functions. The dashed lines represent a patient's utility for each treatment and the solid lines represent the social value of each treatment. The private and socially optimal treatment is different only for types between θ^P and θ^E.

treatment, the socially efficient treatment cut-off will differ from the patients' preferred cut-off whenever $c_1(\theta^E) \neq c_2(\theta^E)$. As a matter of convention I assume $c_1(\theta^E) < c_2(\theta^E)$ so that $\theta^P < \theta^E$ and more types prefer to be treated with T_2 than is socially optimal.[8]

Illustrating the properties A1–A4 establish, Figure 1 identifies the socially efficient treatment cut-off θ^E and the patients' preferred treatment cut-off θ^P by plotting both the patients' expected benefit from treatment with both treatments as well as the net social value of the two treatments. Because treatment T_2 is more costly than T_1, patients prefer more of T_2 than is socially optimal. In this model, full insurance does not cause every patient to demand inefficient levels of treatment, but rather, it causes the marginal types between θ^P and θ^E to demand more treatment than is socially beneficial. Observe that PID would only occur if physicians chose to treat types below θ^P with T_2.

[8]This can be reversed of course, in which case patients will prefer T_1 over T_2 more than is socially optimal. Though the results will change accordingly, the intuition behind the results remains the same.

The patients, physician, and insurer play a simple game. The insurer designs its payment rule having observed the number of physicians in the market, the distribution of illness severities, and the cost and benefits of treatment. Next, given the insurer's payment rule and the number of physicians in the market, physicians simultaneously choose their treatment practice. Lastly, patients choose a physician and are treated. The choice of treatment is assumed to be verifiable to rule out the possibility that the physician can lie about which treatment was used;[9] however, the insurer does not know the patient's illness severity so cannot verify if the chosen treatment was appropriate.

The physician that a patient chooses may or may not depend on the physicians' treatment practice. For example, at one extreme, no patient may know or consider the physicians' treatment practices and instead select physicians based on characteristics, such as location, that are orthogonal to treatment practice. If physicians are uniformly distributed over these other characteristics, then they would capture equal shares of the market and there would be no competition in the treatment practice dimension.[10] Some proportion of patients—call them searchers—may consider physician treatment practices when choosing their physician, however. If there is a physician who follows a treatment practice that these patients prefer, then that physician can expect to garner a higher market share by attracting the searchers. It may also be the case that all patients choose their physician based on the physicians' treatment practices, but patients are not aware of all of the physicians in the market. In this case patients may learn about any given physician with some probability and so choose among the physicians they have learned about.[11]

To capture the notion that not all patients will respond to a physician's choice of treatment practice, let $\phi(n)$, where n is the number of physicians in the market and $\phi : \mathbb{N} \to [0, 1]$,

[9] Ma and McGuire (1997) show that when the patient and physician are responsible for a share of the costs of treatment, then in equilibrium they will not misreport the quantity of treatment used. That equilibrium would also hold here if I assume the patient is responsible for any arbitrarily small coinsurance. As my focus is on using payment rules to induce physicians to practice a preferred treatment style I simply assume the physician cannot misreport which treatment is used.

[10] Alternatively, consider a random utility model where patients receive some utility from physicians that takes the form of an independently and identically distributed random variable, then, in equilibrium, the physicians will have equal market shares.

[11] Note that this implies that patients are simply unaware of physicians they do not know about so they do not have a prior regarding the treatment practices of these physicians nor do they have a prior for the number of physicians that are in the market.

represents the probability that a patient observes (is aware of) a physician's treatment practice when there are n physicians in the market. As the number of physicians increases, $\phi(n)$ may be constant in the case that ϕ simply represents the portion of the market that is concerned with treatment practices or it may either increase or decrease as changing the number of physicians in the market alters the likelihood that patients learn about any specific physician (Satterthwaite, 1979; Pauly and Satterthwaite, 1981).[12] Patients who are not searchers or who are not aware of any physician's treatment practice are randomly matched with a physician. Note that by representing the responsiveness of the market $\phi(n)$ will impact the elasticity of demand with respect to treatment practice. For example, when no patients are searchers ($\phi(n) = 0$), then there is no elasticity of demand with respect to treatment practice. Demand will be at its most elastic when all patients are searchers ($\phi(n) = 1$) as even a small change in treatment practice can result in a large increase or decrease in a physician's demand. In this way, $\phi(n)$ defines the competitiveness of the market and is relevant only when there is more than one physician in the market. I will refer to $\phi(n)$ as the patient response to physician treatment practices.

With this model the socially optimal outcome is defined as the outcome that maximizes social surplus subject to a physician participation constraint and budget balance. As patients are risk averse and the physician is risk neutral there is a social cost to physician profit and the planner's problem can be expressed as maximizing patient utility subject to a break-even constraint for the physician. The following proposition identifies the socially optimal treatment practice.

Proposition 1. *In the first-best physicians will not treat patients when the patients' illness severity is below θ^*, use T_1 when the illness severity is between θ^* and θ^{**} and use T_2 when the illness severity is above θ^{**}, where θ^* solves $\psi(\theta^*, T_1) = U'(Y - P^*)c_1(\theta^*)$ and θ^{**} solves $\psi(\theta^{**}, T_1) - U'(Y - P^*)c_1(\theta^{**}) = \psi(\theta^{**}, T_2) - U'(Y - P^*)c_2(\theta^{**})$. The first-best premium is $P^* = \int_{\theta^*}^{\theta^{**}} c_1(\theta)dF(\theta) + \int_{\theta^{**}}^{\overline{\theta}} c_2(\theta)dF(\theta)$.*

[12] See Rochaix (1989) for a model of consumer search in the market for physicians. In the current model patients could be thought of as having a fixed budget which can be used to research the treatment practice of physicians. A constant research cost per physician would then imply that the probability of learning about any one physician decreases with the number of physicians in the market once the cost of researching all physicians exceeds the search budget.

The intuition behind the first-best outcome is straight-forward. The lower cut-off θ^*, which delineates when the patient should not be treated or should be treated with T_1, is defined by the illness type for which the expected benefit of treatment is equal to the social cost where the social cost is the private cost of treatment weighted by the patients' marginal utility of consumption. Increasing the number of patients who are treated necessitates a higher premium which increases the social cost of treatment. The higher cut-off θ^{**}, which delineates when the patient should be treated with T_1 or T_2, represents the cut-off with which the social value of treatment is the same for each treatment. Lastly the optimal premium represents the expected cost of treatment for a patient. Because patients are risk-averse while physicians are risk-neutral, first-best requires that physicians do not earn positive economic profit.

Moving forward, let $\psi(\underline{\theta}, T_1) - c_1(\underline{\theta}) > 0$ so that it is socially preferable to treat all patients.[13] With this constraint we can concentrate on the physician's decision to use one treatment over the other and define θ^E as the socially optimal treatment cut-off; i.e., $\theta^E = \theta^{**}$ where θ^{**} is defined in Proposition 1. Furthermore, as all patients are treated and there is only one cut-off, it will be notationally convenient to define the term $\mathbb{E}_\theta[c(\theta \mid \hat{\theta})]$—the expected cost of treatment given cut-off $\hat{\theta}$—as follows:

$$\mathbb{E}_\theta[c(\theta \mid \hat{\theta})] \equiv \int_{\underline{\theta}}^{\hat{\theta}} c_1(\theta) dF(\theta) + \int_{\hat{\theta}}^{\overline{\theta}} c_2(\theta) dF(\theta).$$

With these simplifications the first-best optimization program can be expressed as

$$\max_{P,\theta^*} \int_{\underline{\theta}}^{\theta^*} \big[U(Y-P) - L(\theta) + \psi(\theta, T_1)\big] dF(\theta) + \int_{\theta^*}^{\overline{\theta}} \big[U(Y-P) - L(\theta) + \psi(\theta, T_2)\big] dF(\theta),$$

subject to $P = \mathbb{E}_\theta[c(\theta \mid \theta^*)]$ and $\theta^* \in \Theta$.

3. Optimal Payment Rules

3.1. *Monopolist Physician*

The insurer, acting as a Stackelberg leader, chooses the payment rule that induces the physician to optimally choose the socially efficient cut-off and treat all patients below the cut-off

[13] Similarly T_1 can be thought of as a non-clinical treatment regime such as continued monitoring.

with T_1 and all patients above with T_2. The insurer's payment rule consists of a reimbursement for each treatment similar to the prospective payment system used by Medicare.[14,15] When the physician is a monopolist, she can choose the cut-off type that maximizes her profit without concern for how the choice impacts demand. When the physician is paid reimbursements r_1 and r_2 for treatments T_1 and T_2, respectively, her expected profit can be expressed as

$$\int_{\underline{\theta}}^{\overline{\theta}} \big[I_{\{\tau=T_1\}}(\theta)\big(r_1 - c_1(\theta)\big) + \big(1 - I_{\{\tau=T_1\}}(\theta)\big)\big(r_2 - c_2(\theta)\big)\big] dF(\theta), \tag{3}$$

where $I_{\{\tau=T_1\}}(\theta)$ is an indicator function that takes the value of 1 when the physician chooses treatment T_1 and 0 if she chooses T_2.

Similar to the socially optimal treatment choice, the physician's treatment choice is monotone in the illness types.[16] The cut-off type $\hat{\theta}$ thus serves as the physician's choice variable and her optimization program can be expressed as

$$\max_{\hat{\theta}} \int_{\underline{\theta}}^{\hat{\theta}} \{r_1 - c_1(\theta)\} dF(\theta) + \int_{\hat{\theta}}^{\overline{\theta}} \{r_2 - c_2(\theta)\} dF(\theta).$$

Given the insurer sets reimbursements r_1 and r_2, the following Proposition reports the unique payment rule that induces a monopolist physician to follow the socially optimal treatment practice.

[14] Although the prospective payment system used by Medicare is diagnosis (DRG) based in principle, in practice it allows for separate reimbursements based on the treatment chosen for a particular diagnosis. The most clear example of this is natural and cesarean delivery. However, even when a natural delivery is performed the physician can receive more compensation by indicating that there were complications such as would be the case if the physician needed to perform an episiotomy. Similarly chemotherapy is a separate DRG even though it is one of several procedures used to treat cancer.

[15] Unverifiable costs preclude physician cost sharing. If the explicit costs of treatment were constant as in Liu and Ma (2011), then they would not reveal a patient's type and cost sharing would be the same as focusing on a fixed reimbursement.

[16] To see this, suppose there exists a $\hat{\theta} \in \Theta$ such that $r_1 - c_1(\hat{\theta}) = r_2 - c(\hat{\theta})$. Then for any type $\theta > \hat{\theta}$ it must be the case that profit is maximized using treatment T_2 since $c_1'(\cdot) > c_2'(\cdot)$. Similarly, it must be the case that profit is maximized treating any type $\theta < \hat{\theta}$ with T_1. If such a $\hat{\theta} \in \Theta$ does not exist, then the physician will either treat all patients with T_1 or T_2 depending on which produces the most profit.

Proposition 2. *Let $\{r_1^*, r_2^*\}$ be the insurer's payment rule. The insurer can induce the socially optimal outcome if and only if the payments are non-negative and*

$$r_1^* = \mathbb{E}_\theta[c(\theta \mid \theta^E)] + \big[1 - F(\theta^E)\big]\big[c_1(\theta^E) - c_2(\theta^E)\big] \text{ and}$$
$$r_2^* = \mathbb{E}_\theta[c(\theta \mid \theta^E)] + F(\theta^E)\big[c_2(\theta^E) - c_1(\theta^E)\big].$$

The payments reported by Proposition 2 are possible because there is a monotone difference in the physician's cost between the two treatments.[17] This allows the insurer to induce the physician to choose any type $\theta \in \Theta$ as the treatment cut-off type by making the physician internalize the social gain. The payments that induce the socially optimal cut-off and leaves the physician with zero rents can be described as taking the expected cost of treating a patient and adding to it the probability that the patient should receive the *other* treatment times the difference in costs between the treatments at the optimal cut-off. In this way the physician is exactly indifferent at the cut-off type and strictly prefers to treat lower types with T_1 and higher types with T_2.

The discrete nature of the treatment decision means that changes in the relative payments will not cause the physician to stat using that treatment on all of her patients, but rather causes the physician to adjust her marginal type for one treatment over the other. Gruber et al. (1999) observe this behavior in the rates at which cesarean delivery versus normal delivery is performed when the Medicaid program lowered the reimbursement amount for a cesarean delivery. The authors find that reducing the fee differential between natural and cesarean delivery reduced the cesarean rate for Medicaid patients and argue that the difference in fees between the privately insured and Medicaid patients explains the majority of the difference in cesarean rates between the populations. In the current model, if T_2 represented a cesarean delivery, then a reduction in r_2 will shift the cut-off type upward resulting in more natural deliveries.

The payment rule reported by Proposition 2 does have one limitation. If the difference in costs are sufficiently large at the socially optimal cut-off, then r_1^* may need to be negative if

[17] Again, the results of Proposition 2 would also be obtained if instead of assumption A4 the costs were constant and the physician was not completely self-interested and had some preference for patient health benefit as well.

the physician is to internalize the social gains. When the insurer is restricted to a non-negative payment, it must increase both treatment payments to induce the desired cut-off, θ^*.[18] However, $\theta^* = \theta^E$ will no longer be the socially-optimal cut-off since the patient's utility function is concave in income and the physician receives rents. Indeed, it can be shown that if $c_1(\theta^E) \ll c_2(\theta^E)$ such that $r_1^* < 0$, then the second-best optimal cut-off is higher than the first-best cut-off and the physician may earn positive profit. That is, in the second best, treatment T_1 will be used on more patients than is socially optimal. Furthermore, though it has been assumed that $c_1(\theta^E) < c_2(\theta^E)$, this assumption could be reversed and it could be the case that $c_1(\theta^E) \gg c_2(\theta^E)$ such that $r_2^* < 0$. In this case the second-best cut-off is lower than the first-best (similarly moving it further from the patients' preferred cut-off) and the physician may be left with positive profit. The following proposition summarizes this limitation.

Proposition 3. *Let $\{r_1^*, r_2^*\}$ be the insurer's payment rule defined in Proposition 2. If $r_1^* < 0$, then $\theta^{SB} > \theta^E$ and the physician earns expected profit $\big[1 - F(\theta^{SB})\big]\big[c_2(\theta^{SB}) - c_1(\theta^{SB})\big] - \mathbb{E}_\theta[c(\theta \mid \theta^{SB})] \geq 0$. If instead $r_2^* < 0$, then $\theta^{SB} < \theta^E$ and the physician earns expected profit $F(\theta^{SB})\big[c_1(\theta^{SB}) - c_2(\theta^{SB})\big] - \mathbb{E}_\theta[c(\theta \mid \theta^{SB})] \geq 0$.*

When there is a large disparity in the treatment costs, Proposition 3 states that the physician must earn rents and the second-best treatment practice will be distorted in the direction of the less expensive treatment. An insurer can avoid leaving the physician with rents (and obtain the first-best outcome) if it can charge the physician an "access fee" for having the right to treat its patients where the fee extracts the physician's rents and is independent of the quantity of patients treated. Ma and McGuire (1997) encounter a similar problem when quantity and effort are substitutes and there is competition between physicians.

[18] Utilizing some form of provider cost-sharing does not help here either because whenever $r_1^* < 0$ the physician must be losing more than her total cost $c_1(\theta^E)$ if she is to internalize the social gain of using one treatment over the other.

3.2. *Competing Physicians*

Competition for market share will impact physicians' treatment practices when demand exhibits some elasticity with respect to those practices. For instance, in the market for obstetricians expecting mothers may look at the frequency at which obstetricians perform a cesarean section or episiotomy compared to other obstetricians. If an obstetrician is observed to resort to a cesarean delivery more or less frequently than patients prefer or think appropriate, then patients may substitute away from her.[19] Knowing this, obstetricians may try to balance the rate at which they perform either procedure based on the relative difference in payments and the treatment rates of other obstetricians. The insurer still acts as a Stackelberg leader in this environment; however, the insurer's problem is made more difficult because the payment policy must be designed to prevent competitive deviations that don't exist when demand is inelastic to treatment practice.

The equilibrium of interest is the symmetric equilibrium in which all physicians choose the same treatment cut-off as any other equilibrium necessarily results in an inefficient level of treatment.[20] In a symmetric equilibrium all of the physicians choose the same cut-off so we just need to consider the impact of a unilateral deviation by some physician from the equilibrium cut-off. First, suppose all of the physicians choose cut-off $\theta^* > \theta^P$ and one physician deviates upward by choosing some $\hat{\theta} > \theta^*$. In this case the physician has moved to a cut-off that is less preferred by patients than θ^*. Consequently, she will lose demand from those marginal types between θ^* and $\hat{\theta}$ who are responsive to physician treatment practices (i.e., the $\phi(n)$ proportion of the marginal types who are searchers). Although the physician loses demand, this deviation is profitable when the cost of using T_1 is substantially lower that the gain in profit from each patient the physician keeps exceeds the loss in profit from losing market share.

Instead of deviating to a less preferred treatment cut-off and losing some patients, a physician could deviate to a $\hat{\theta} < \theta^*$. When $\hat{\theta} \geq \theta^P$ the deviation represents a treatment practice that is

[19] Asking about an obstetrician's cesarean delivery rate and views on cesarean delivery is a common recommendation made by family planning and pregnancy websites such as babycenter.com, babyzone.com, and familyresource.com.

[20] Similarly, any mixed strategy requires that physicians choose some sub-optimal cut-off with positive probability.

more preferred by marginal types between $\hat{\theta}$ and θ^* and the physician attracts additional patients. It need not be profitable for the physician to treat one of the marginal types with T_2 instead of T_1 for this to represent a profitable deviation as the physician's increase in demand may be sufficiently large that she is able to increase her total profit. This strategy is not a "cream-skimming" strategy in the traditional sense as the physician is not attracting the least costly patients; however, she is attracting the patients that are the most profitable when treated using the high cost treatment and makes up for any loss in profit relative to using treatment T_1 through an increase in demand.

If an insurer is to induce its preferred treatment cut-off, it must of course design the payments to ensure that these types of deviations are not profitable. The following proposition identifies the necessary and sufficient conditions for a payment rule that will prevent these profitable deviations at some $\theta^* \in (\theta^P, \overline{\theta})$.

Proposition 4. *Given a payment rule* $\{r_1, r_2\}$*,* $n \geq 2$ *physicians, and patient response* $\phi(n) \in [0, 1)$*, a cut-off* θ^**, where* $\theta^P < \theta^* < \overline{\theta}$*, represents an equilibrium if and only if* $\mathbb{E}_\theta[\Pi(\theta \mid r_1, r_2)] \geq 0$ *and the payment rule satisfies:*

$$1 + \phi(n)(n-1) \leq \frac{r_1 - c_1(\theta^*)}{r_2 - c_2(\theta^*)} \leq \big(1 - \phi(n)\big)^{-1}. \tag{4}$$

Proposition 4 provides a bounds on the relative profit between treatments that a physician can earn. If the profit from using treatment T_2 is too low relative to using T_1 at the cut-off θ^*, then she can increase her profit by raising her cut-off. The physician will lose some demand, but will earn a higher return on those patients she retains. This deviation generates the upper bound of (4). On the other hand, if the profit from treatment T_2 is sufficiently high, then she can increase her demand by lowering her cut-off type and attracting an additional $\phi(n)(n-1)/n$ of the marginal types. This deviation generates the lower bounds in (4), which represents the proportional gain in demand when the physician deviates to a lower (more preferred) cut-off.

Ensuring that a physician treat all types of patients is as straight-forward as telling patients to report when a physician refuses to treat them and penalizing that physician. Consequently the payment rule can be designed so that one treatment subsidizes the other; i.e.,

$\int_{\underline{\theta}}^{\theta^*}\big(r_1 - c_1(\theta)\big)dF(\theta) = -\int_{\theta^*}^{\overline{\theta}}\big(r_2 - c_2(\theta)\big)dF(\theta)$. However, when physicians compete for patients, cross-subsidization of treatments becomes problematic if all patients are responsive to all physicians' treatment practices; i.e., if $\phi(n) = 1$. This is because a physician who incurs a loss from treatment T_2 can raise her cut-off to $\overline{\theta}$ and dump *all* of the high-cost patients preventing the insurer from inducing an equilibrium at any $\theta^* \in (\theta^P, \overline{\theta})$. In consequence, the high-cost treatment cannot be completely subsidized by the lower cost treatment and the payment rule identified in Proposition 4 is insufficient when $\phi(n) = 1$. The following corollary to Proposition 4 reports this finding.

Corollary 1. *If $\phi(n) = 1$ and $n \geq 2$, then cross-subsidization of treatments is not possible and a physician can be induced to choose cut-off $\theta^* \in (\theta^P, \overline{\theta})$ only if $\int_{\theta^*}^{\overline{\theta}}[r_2 - c_2(\theta)]d\theta \geq 0$.*

The bounds on the physicians' relative profit at the treatment cut-off are defined by the total demand that comes from a unilateral deviation in the treatment cut-off. A payment rule can only satisfy the bounds if they are consistent. That is, a payment rule can only satisfy the bounds when $1 + \phi(n)(n-1) < \big(1 - \phi(n)\big)^{-1}$. For sufficiently low observation probabilities these two payment constraints will be in conflict, however. When $1 + \phi(n)(n-1) > \big(1 - \phi(n)\big)^{-1}$ the insurer will only be able to either prevent demand-stealing downward deviations or patient dumping upward deviations, but not both. This limitation exists because of how the demand response asymmetrically impacts the physician's deviation profit. A physician that deviates to a higher cut-off loses demand from only the share of patients who both respond to her treatment practice and would have visited her if she chose the same cut-off as the other physicians but don't because of the deviation while a physician who deviates downward takes some market share from *all* of the physicians in the market.[21] When there are a large number of physicians, even a small demand response can result in a significant increase in demand for the deviating physician. Consequently the insurer must set the payments so that the profit differential between treating the marginal types with T_1 over T_2 is sufficiently great that such a deviation is

[21]The extent of the deviation will crucially depend on the spare capacity of the physician. It is not clear how much capacity physicians generally have, but Gruber and Owings (1996) and Dunn and Shapiro (2012) observed that physicians are able to increase their treatment quantity following payment changes suggesting they may generally have spare capacity.

discouraged; but such a large difference in profit will then encourage the physician to deviate upward when it will not lose that many patients as a result. For this reason, successfully inducing its preferred treatment practice requires some minimal demand response from patients that is dependent on the number of physicians in the market.[22] When the insurer loses the ability to choose a specific interior cut-off it must decide whether to induce the patients' preferred cut-off θ^P, to induce physicians to treat all patients with treatment T_1 by inducing the cut-off $\overline{\theta}$, or to allow physicians to randomly choose an interior cut-off. The following Proposition summarizes these findings.

Proposition 5. *If $0 < \phi(n) < \frac{n-2}{n-1}$, then an insurer cannot induce $\theta^* \in (\theta^P, \overline{\theta})$ using any payment policy $\{r_1, r_2\}$, where $r_1, r_2 \in \mathbb{R}_+$, and the only equilibria are either mixed strategy equilibria in which the choice of cut-off is stochastic or one of the boundaries θ^P and $\overline{\theta}$.*

The bad news for the insurer is that the treatment elasticity of demand must increase with the number of physicians if it is to induce the desired treatment practice. In this case policies that increase the transparency of physician practices by providing more information to patients may have a large impact on welfare since they can generate market conditions that permit the insurer to induce the socially efficient treatment cut-off. Although it is not clear how much patients consider physician treatment practices when choosing their physician there is strong evidence that many patients do seek out physicians based on their treatment recommendations (e.g.,McCarthy (1985); Macpherson et al. (2001)) indicating that there may be some opportunity to increase treatment efficiency by increasing transparency.

Before concluding this section, I return to the assumption that patients know their own type. This assumption is important because it determines which patients are impacted by and respond to physician treatment practices. For example, if all physicians choose the cut-off θ^* and one deviates to a lower cut-off $\hat{\theta}$, then only the marginal types between $\hat{\theta}$ and θ^* will benefit by selecting that physician. Instead of knowing their type, however, patients may only receive a signal (e.g., symptoms) that provide noisy information about their true type. For example each

[22]Note that when there is no demand response, then all physicians effectively act as monopolists with respect to their treatment practice.

patient i may receive a noisy signal $\tilde{\theta}_i = \theta_i + \epsilon_i$ where θ_i is i's true type and ϵ_i is a zero mean random variable. With this information structure the bounds on ϵ will define the marginal types that select a physician based on her treatment practice relative to the other physicians' practices. For example, suppose the ϵ_i have support $[-\epsilon, \epsilon]$ where $\epsilon > 0$, then any patient receiving signal $\tilde{\theta}_i \in [\hat{\theta} - \epsilon, \theta^* + \epsilon]$ may benefit by selecting the deviating physician while patients receiving signals outside these bounds will not benefit. If the bound ϵ is sufficiently large, then any signal may indicate that there is a positive probability that the patient's true type is between $\hat{\theta}$ and θ^* and all patients have the potential to benefit by selecting the deviating physician. This case is similar to that of Bertrand price competition because a physician will attract all patients who observe her treatment practice if she shifts it by any arbitrarily small amount away from θ^* towards the patients' preferred cut-off θ^P. Following the proof for Proposition 4 it is easy to show that the payment rule in this case is equivalent to the monopoly payment rule reported in Proposition 3 as the payments that maximize the physician's profit at the desired cut-off will also be maximal when all other physicians choose the same cut-off. This case may better represent patient choice for a family physician where the choice is typically made prior to falling ill and learning one's type. On the other hand, physician treatment practices are likely to impact specific patients when they have some information about their illness type or their prognosis with various treatments such as may be the case with the chronically ill or those requiring the services of a specialist.

3.3. *Comparative Statics*

In order to identify the comparative statics for the equilibrium cut-off, an equilibrium cut-off must be pinned down. Proposition 4 provides a bounds on the physicians' profit for each treatment at some cut-off. The interval of types implied by these bounds is degenerate, however, only when $\phi(n) = 0$. Otherwise every θ within that interval represents a Nash equilibrium. The equilibrium can be refined, though, by assuming physicians coordinate on the cut-off type which is not Pareto dominated by another type within the interval. The Pareto dominant cut-off is easy to find by observing that when $n \geq 2$ and $\phi(n) > 0$, both the upper and lower bounds of the ratio of physician profit are greater than one. Furthermore, in a symmetric equilibrium, the

physician's profit is maximized when $r_1 - c_1(\theta^*) = r_2 - c_2(\theta^*)$; i.e., when the ratio of profit at the cut-off is unity. Given the concavity of the physicians' problem, this implies that their profit is highest at the lowest bound; i.e., at $r_1 - c_1(\theta^*) = \big(1 + \phi(n)(n-1)\big)\big(r_2 - c_2(\theta^*)\big)$. This refinement is summarized in the following lemma.

Lemma 1. *When the payment rule $\{r_1, r_2\}$ satisfies condition (4) of Proposition 4 the Pareto dominant equilibrium cut-off is the $\hat{\theta}$ at which the ratio equals the lower bound; i.e., the $\hat{\theta}$ at which $r_1 - c_1(\hat{\theta}) = \big(1 + \phi(n)(n-1)\big)\big(r_2 - c_2(\hat{\theta})\big)$.*

If physicians coordinate on the Pareto dominant cut-off when given a particular payment rule, then the comparative statics can be derived for the cut-off type and payments with respect to changes in the number of physicians and the transparency of their treatment practice. These comparative statics are reported in the following Proposition.

Proposition 6. *(Comparative Statics) For a fixed payment rule $\{r_1, r_2\}$, the Pareto dominant equilibrium cut-off θ^* satisfies the following statics:*

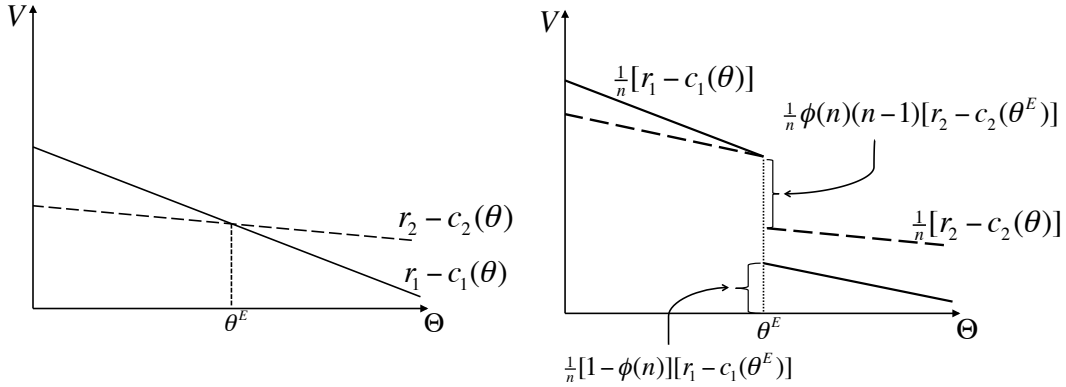
$$\frac{d\theta^*}{d\phi(n)} < 0 \text{ and } \frac{d\theta^*}{dn} \lesseqgtr 0 \text{ when } (n-1)\phi'(n) \lesseqgtr -\phi(n).$$

Furthermore the optimal payment rule satisfies the following statics:

$$\frac{dr_2^*}{d\phi(n)} < 0 < \frac{dr_1^*}{d\phi(n)} \text{ and } \frac{dr_2^*}{dn} \lesseqgtr 0 \lesseqgtr \frac{dr_1^*}{dn} \text{ when } (n-1)\phi'(n) \lesseqgtr -\phi(n).$$

The comparative statics are divided into two sets. The comparative statics for $d\theta^*/d\phi(n)$ and $d\theta^*/dn$ identify how the equilibrium cut-off will change when there is a change in the market responsiveness to treatment practices or the number of physicians and the payments remained fixed.[23] The intuition for $d\theta^*/d\phi(n) < 0$ follows from the fact that as more of the market becomes responsive to the physicians' treatment practice, there is a larger incentive to deviate towards patients' preferred treatment cut-off. Similarly, increasing the number of physicians in the market decreases the physicians' market shares, increasing the incentive to

[23]This latter circumstance has been used in the literature to identify physician induced demand. See, for example, Fuchs (1978); Auster and Oaxaca (1981); Pauly and Satterthwaite (1981).



(a) $n = 1$ (b) $n > 1$

FIGURE 2. Physician profit by treatment choice and type with linear cost functions. The solid line represents a physician's profit from using treatment T_1 and the dashed line represents a physician's profit from using treatment T_2. When there are multiple physicians, the profits are conditional on all other physicians providing the socially optimal treatment.

increase market share by deviating towards patients' preferred treatment cut-off. However, increasing the number of physicians may actually decrease the responsiveness of the market to any one physician's treatment practice. The sign depends on which effect dominates. The implication of these statics are that we should expect to see variation in the treatment rates for a particular illness across markets of varying competitiveness when the payments across those markets are the same. This may provide an additional explanation for some of the variation in treatment practices observed across various regions (Phelps, 2000).

The second set of comparative statics indicate the direction that the payments must be adjusted in order to maintain the *same* equilibrium cut-off when either the demand response changes or there is a change in the number of physicians in the market. Similar to $d\theta^*/dn$ the signs for both dr_1^*/dn and dr_2^*/dn are dependent on $sign[\phi'(n)]$ since increasing the number of physicians may decrease the market's responsiveness to any one physician. When $\phi(n)$ increases or an increase in n increases the incentive to deviate downward, the payments must be adjusted so that it becomes more profitable to treat higher types with T_1 and less profitable to treat lower types with T_2 to counteract the incentive to deviate towards the patients' preferred cut-off.

Using linear cost functions, Figure 2 illustrates the difference in a physician's profit between when she is a monopolist from when she competes with other physicians. In both cases the payments are set so that the physician is indifferent in what treatment to choose at θ^E. When the physician is a monopolist this simply means the physician earns the same economic profit at θ^E for each treatment. However, when the physician competes, then she will earn substantially different profits at θ^E depending on the treatment chosen. Nevertheless, the physician is indifferent at θ^E because the potential gain from deviating generates a discontinuity in the profit, which drives the comparative statics. Figure 2(b) illustrates the Pareto dominant equilibrium in which the lower bound of the interval was chosen; i.e., where $r_1 - c_1(\theta^E) = \big(1+\phi(n)(n-1)\big)\big(r_2 - c_2(\theta^E)\big)$. If $\phi(n)(n-1)$ increases, then the discontinuity from shifting cut-off types must also widen. The interval of types θ that can be supported as an equilibrium with the same payment rules can be found by extending the portion of $r_1 - c_1(\theta^E)$ and $r_2 - c_2(\theta^E)$ that is to the right of θ^E to the north-west until they intersect. Any type between the intersection and θ^E represents a Nash equilibrium with the payments r_1 and r_2. Lastly, note that although Figure 2 represents an equilibrium it does not depict the optimal payment rule because in both figures the physician is left with positive economic profit.

4. Diagnostic Testing

Physicians may of course not know precisely which treatment is optimal for a particular patient. Moreover, there is concern that diagnostic testing is driving some of the higher utilization rates and cost of medical care (e.g., Smith-Bindman, Miglioretti and Larson, 2008; Lehnert and Bree, 2010) so it is important to understand the incentives behind testing. There are many illnesses in which a diagnostic test (or tests) is always necessary as a component to treatment or is needed in order to identify the patient's illness; for example, an x-ray will be ordered when a bone is obviously broken in order to identify the nature of the break and how best to reset it and a patient presenting with a set of symptoms that could be attributed to a multitude of illnesses may have a blood test to narrow the diagnosis. Testing in these situations is simply part of the cost of treatment and captured by the no-diagnostic testing regime. The kind of testing that

this section is concerned with is testing that is not always medically necessary but may be used to help the physician determine the optimal treatment. To identify how this type of diagnostic testing alters the payment rule the following modifications are made to the model. First, instead of observing a patient's illness type θ_i directly, the physician and patient observe the signal $\xi_i \in [\underline{\xi}, \overline{\xi}] = \Xi$ for patient i's illness, which could be thought of as the patient's symptoms and are drawn from cumulative distribution function Γ. The patient's illness type is related to the signal through the conditional CDF $G(\theta \mid \xi)$ where $G(\cdot \mid \xi_1)$ first order stochastically dominates $G(\cdot \mid \xi_2)$ for all $\xi_1, \xi_2 \in \Xi$ and $\xi_1 > \xi_2$; i.e., higher signals indicate that the illness type is likely to be higher as well. Patient responsiveness to physician testing practices will be represented by $\phi(n)$.

Physicians have available a diagnostic test which completely reveals the patient's type and incur a cost $D > 0$ when they administer the test.[24] Because the physician incurs a cost, however, it is not always optimal to perform the test even though it will eliminate the risk of choosing a suboptimal treatment.[25] This follows, of course, because the value of eliminating the risk may be less than the cost. Given that this trade-off exists, the following Proposition establishes the first-best testing and treatment practice.

Proposition 7. *The first-best diagnostic practice is as follows. Let* $\Delta\psi(\theta) = \psi(\theta, T_2) - \psi(\theta, T_1)$ *and* $\Delta c(\theta) = c_2(\theta) - c_1(\theta)$ *and define* ξ^* *and* ξ^{**} *as*

$$\xi^* = \min \left\{ \xi \in \Xi \,\middle|\, \int_{\theta^E}^{\overline{\theta}} \left[\Delta\psi(\theta) - U'(Y - P^*)\Delta c(\theta)\right] dG(\theta \mid \xi) \geq U'(Y - P^*)D \right\}$$

$$\xi^{**} = \max \left\{ \xi \in \Xi \,\middle|\, \int_{\underline{\theta}}^{\theta^E} -\left[\Delta\psi(\theta) - U'(Y - P^*)\Delta c(\theta)\right] dG(\theta \mid \xi) \geq U'(Y - P^*)D \right\},$$

[24] The cost to the physician or patient is again not necessarily monetary, but rather can reflect the opportunity cost of their time, or in the case of the patient, the opportunity cost incurred from any discomfort or pain experienced in the course of the test.

[25] Note that D is the total cost for administering the test and who incurs the cost is irrelevant for first-best. That the physician is the one who incurs the cost only matters when deriving the optimal payment rules.

where

$$P^* = \int_{\underline{\xi}}^{\xi^*} \mathbb{E}_\theta[c(\theta) \mid \overline{\theta}) \mid \xi] d\Gamma + \int_{\xi^*}^{\xi^{**}} \mathbb{E}_\theta[c(\theta) \mid \theta^*) \mid \xi] d\Gamma + \int_{\xi^{**}}^{\overline{\xi}} \mathbb{E}_\theta[c(\theta) \mid \underline{\theta}) \mid \xi] d\Gamma.$$

When both ξ^* and ξ^{**} exist the physician does not perform the diagnostic test and treats patients with T_1 when $\xi < \xi^*$ and T_2 when $\xi > \xi^{**}$; and the physician performs the diagnostic test when $\xi \in [\xi^*, \xi^{**}]$ and uses the treatment rule presented in Proposition 1 based on the realized value of θ. When either ξ^* or ξ^{**} does not exist the physician does not perform a diagnostic test and treats all patients below ξ^E with T_1 and all patients above ξ^E with T_2 where

$$\int_\Theta \big[\psi(\theta, T_1) - U'(Y - P^*) c_1(\theta)\big] dG(\theta \mid \xi) d\Gamma(\xi^E) = \int_\Theta \big[\psi(\theta, T_2) - U'(Y - P^*) c_2(\theta)\big] dG(\theta \mid \xi) d\Gamma(\xi^E), \tag{5}$$

and $P^* = \int_{\underline{\xi}}^{\xi^E} \mathbb{E}_\theta[c(\theta) \mid \overline{\theta}) \mid \xi] d\Gamma + \int_{\xi^E}^{\overline{\xi}} \mathbb{E}_\theta[c(\theta) \mid \underline{\theta}) \mid \xi] d\Gamma.$

The proposition identifies two diagnostic cut-offs. The lower cut-off, ξ^* represents the signal below which no diagnostic test is performed and the patient is treated with T_1 and above which the diagnostic test is performed. The upper cut-off, ξ^{**} represents the signal above which no diagnostic test is performed and the patient is treated with T_2 and below which the diagnostic test is performed. The conditions identifying ξ^* and ξ^{**} compare the social benefits of testing to the social cost where the left-hand-sides of both conditions represents the benefit of conducting the test and the right-hand-sides are the social costs. In the case of the lower diagnostic cut-off, ξ^*, the higher the value for ξ^*, the larger the expected gain from performing the diagnostic test and selecting the optimal treatment over not testing and treating with T_1 since higher signals indicate a higher probability of a high-type. Similarly, lower values for ξ^{**} result in larger expected gains from performing the diagnostic test and choosing the optimal treatment over not testing and treating with T_2 since a lower signal implies a higher probability that the type is low. When the gains from performing the test do not exceed the cost of the test then there is no

diagnostic cut-off and first-best requires the physician to use the treatment that maximizes the expected social value of treatment without performing test.[26]

An insurer wanting to induce some pair of diagnostic cut-offs, $\{\xi^*, \xi^{**}\}$ as well as a treatment cut-off θ^* may be limited when the private benefit for the physician exceeds the social benefit at the socially optimal cut-offs. This causes a problem because, although an insurer can differentially reimburse the physician for the treatment chosen based on whether or not she also conducted a diagnostic test, I impose the practical assumption that a physician can choose not to reveal she has administered the test if withholding that information is beneficial. In this way, a physician could be paid more if she administers the test, but it is impossible to pay the physician more *not* to conduct the test.[27] In consequence, if the physician benefits more by conducting the test at the the socially-optimal diagnostic cut-off, then there is no payment rule that can induce the optimal cut-off. The following lemmas provide the conditions under which this occurs.

Lemma 2. *If there is no cost to patients from the diagnostic test, then the insurer cannot induce both treatment cut-off $\theta^* > \theta^P$ and upper diagnostic cut-off $\xi^{**} \in (\underline{\xi}, \overline{\xi})$ with any payment rule $\{r_1, r_2, R\}$ where $r_1 \geq 0, r_2 \geq 0$ and $R \geq 0$ when*

$$\big[\big(1+\phi(n)(n-1)\big)\big(1-F(\theta^*)\big)-1\big]\left[c_2(\theta^*)-\frac{c_1(\theta^*)}{1+\phi(n)(n-1)}\right] > \big(1+\phi(n)(n-1)\big)\big[\mathbb{E}_\theta[c(\theta \mid \theta^*) \mid \xi^{**}]+D\big]-\mathbb{E}_\theta[c(\theta \mid \underline{\theta}) \mid \xi^{**}]. \tag{6}$$

Condition (6) reported by lemma 2 has a simple interpretation. The left-hand-side of (6) represents the minimum expected change in revenue from a payment rule $\{r_1, r_2\}$ that induces θ^* for a physician when she deviates to a higher diagnostic cut-off (tests more patients) and there is no additional payment for the test ($R = 0$). The right-hand-side represents the change in the physician's cost for this deviation. Observe that even when there is no demand response to the physicians' choice of testing cut-off ($\phi(n) = 0$) the left-hand-side will not exceed the right-hand-side for small values of D. This is problematic because many forms of diagnostic testing

[26] It may be the case that only one treatment is optimal for all possible diagnostic signals.

[27] This is in contrast to Ma and Riordan (2002) who allow an insurer to pay a physician not to treat patients. The difference is that in the current model the physician always treats the patient and administering the diagnostic test increases her profit. In contrast, in Ma and Riordan (2002) the physician's profit is lowered by treating the patient.

such as magnetic resonance imaging (MRI) are done by a third party and the Stark laws prevent physicians from self-referral. This can compound the insurer's ability to induce its preferred testing cut-off as the physician still benefits from the information provided by the exam while incurring zero additional cost (there may be a time cost to waiting for the test results). This can be overcome simply by requiring that the physician pay for the test; i.e, by having the physician internalize the cost.

The following lemma reports a similar limitation for the insurer trying to induce some lower diagnostic and treatment cut-offs ξ^* and θ^*, respectively.

Lemma 3. *If there is no cost to patients from the diagnostic test, then the insurer cannot induce both treatment cut-off $\theta^* > \theta^P$ and lower diagnostic cut-off $\xi^* \in (\underline{\xi}, \overline{\xi})$ with any payment rule $\{r_1, r_2, R\}$ where $r_1 \geq 0, r_2 \geq 0$ and $R \geq 0$ when*

$$\begin{aligned}\big(1 - F(\theta^*)\big)\big[\big(1+\phi(n)(n-1)\big)c_2(\theta^*) - c_1(\theta^*)\big] > \\ \big(1 + \phi(n)(n-1)\big)\big[\mathbb{E}_\theta[c(\theta \mid \theta^*) \mid \xi^*] + D\big] - \mathbb{E}_\theta[c(\theta \mid \overline{\theta}) \mid \xi^*].\end{aligned} \tag{7}$$

Observe that both lemmas report conditions in which a physician will expand the set of signals from which she will administer the diagnostic test. Specifically, Lemma 2 describes an upward deviation in the upper diagnostic cut-off and Lemma 3 describes a downward deviation in the lower diagnostic cut-off. These represent deviations that are more preferred by patients because they always benefit from the physician having better information.[28]

The preceding lemmas show that the physicians' incentives may be too strong to prevent it from performing more testing than is socially efficient. However, when the social benefit of the test exceeds the physician's private benefit at the socially optimal cut-offs, the insurer must provide extra compensation. The difference in profit for the physician at the two diagnostic cut-offs differ indicating that a single additional payment for the test will not induce both cut-offs. Instead, the insurer can provide different reimbursements for the two treatments conditional on whether or not a diagnostic test is administered. That is, the insurer can also offer payments

[28]This will not be the case when the diagnostic test imparts an opportunity cost on the patient and I discuss this further at the end of the section.

$\hat{r}_1$ and $\hat{r}_2$ for treatments T_1 and T_2, respectively, conditional on no diagnostic test. Proposition 4 continues to provide the necessary and sufficient conditions for a payment rule that induces a physician to choose cut-off θ^* given the diagnostic test has been performed. As with the no diagnostic testing case, the conditional reimbursements must be designed so that physicians have no incentive to take market-share or dump costly patients by deviating to a different cut-off. As reported by Proposition 8, this bounds the payments in a similar way to the no diagnostic case explored in Section 3.

Proposition 8. *Given a payment rule* $\{r_1, r_2, \hat{r}_1, \hat{r}_2\}$ *and patient response* $\phi(n) \in [0, 1)$*, diagnostic cut-offs* $\{\xi^*, \xi^{**}\}$ *and treatment cut-off* θ^**, where* $\xi^*, \xi^{**} \in \Xi$ *and* $\theta^P \leq \theta^* \leq \overline{\theta}$*, represent an equilibrium if and only if* $E_\theta[\Pi(\theta \mid r_1, r_2, \hat{r}_1, \hat{r}_2)] \geq 0$*,* $r_1 > \hat{r}_1$*,* $r_2 > \hat{r}_2$*, and the payments satisfy:*

$$1 + \phi(n)(n-1) \leq \frac{\hat{r}_1 - \mathbb{E}_\theta[c(\theta \mid \overline{\theta}) \mid \xi^*]}{\mathbb{E}_\theta[\Pi(\theta \mid \theta^*) \mid \xi^*] - D} \leq \big(1 - \phi(n)\big)^{-1},$$

$$1 + \phi(n)(n-1) \leq \frac{\hat{r}_2 - \mathbb{E}_\theta[c(\theta \mid \underline{\theta}) \mid \xi^{**}]}{\mathbb{E}_\theta[\Pi(\theta \mid \theta^*) \mid \xi^{**}] - D} \leq \big(1 - \phi(n)\big)^{-1},$$

$$1 + \phi(n)(n-1) \leq \frac{r_1 - c_1(\theta^*)}{r_2 - c_2(\theta^*)} \leq \big(1 - \phi(n)\big)^{-1}.$$

The last condition in Proposition 8 is simply a restatement of the condition identified by Proposition 4; however, the first two conditions generate bounds on the reimbursements conditional on no diagnostic testing. In the first expression the numerator of the middle term represents the physician's profit from not administering the test and always choosing T_1 when the diagnostic signal is ξ^*. The denominator represents the expected profit from administering the test when the diagnostic signal is ξ^*. The second expression has a similar interpretation, except that the numerator represents the physician's profit from not administering the test and choosing T_2 and both the numerator and denominator are evaluated at diagnostic signal ξ^{**}. A monopolist physician must be indifferent at both cut-offs whereas competition creates a wedge between the diagnostic and no-diagnostic profits at the cut-offs similar to the wedge that competition generates at the treatment cut-off since physicians can gain market share or dump costly patients vis-á-vis the diagnostic practices of other physicians.

Lastly, if patients also incur a sufficiently large cost from the diagnostic test, which may be the case for time consuming or uncomfortable tests such as a colonoscopy, then the physicians' incentives may be reversed at one or both of the diagnostic cut-offs. For example, if patients incur a cost $B > 0$ from the diagnostic test, then at lower diagnostic cut-off ξ^* patients will prefer to be tested less whenever $\mathbb{E}_\theta[\psi(\theta \mid \overline{\theta}) \mid \xi^*] > \mathbb{E}_\theta[\psi(\theta \mid \theta^*) \mid \xi^*] - B$. In this case, a physician gains market share by increasing her choice of lower diagnostic cut-off and testing *fewer* patients and an insurer must compensate a physician for the test in order to counteract this competitive incentive. Similarly, if B is sufficiently large that $\mathbb{E}_\theta[\psi(\theta \mid \underline{\theta}) \mid \xi^{**}] > \mathbb{E}_\theta[\psi(\theta \mid \theta^*) \mid \xi^{**}] - B$, then the insurer must compensate the physicians for performing the diagnostic test in order to overcome the competitive incentive to reduce testing.

5. Conclusions

I have shown that physician competition can work against an insurer by creating a countervailing incentive to provide the insurer's preferred treatment. Under certain circumstances competition creates a situation in which the insurer simply cannot induce its preferred treatment practice through payments alone because the incentive to gain market share or dump more costly patients are too strong. In this circumstance the insurer will have to resort mechanisms such as utilization review to induce its preferred treatment practice. The results rely on a couple well-known characteristics of the market for healthcare. First, because patients are insured there is moral hazard and patients will prefer the treatment maximizing their well-being. This will be true even when there is demand-side cost sharing as long as the difference in the patient's out-of-pocket costs between treatments is not equal to the true cost differences, which eliminates all moral hazard. Second, patients are (somewhat) responsive to the treatment practices of physicians. Their responsiveness to physicians' treatment practices combine to generate an incentive for physicians to alter their practice relative to other physicians to either take market share or treat less of the highest cost patients.

Adding diagnostic testing that may not be medically necessary to the model further complicates the insurer's problem. Because testing is not beneficial when the physician receives

extreme signals, indicating the patient is most likely to be a very high or low type, the insurer prefers the physician administer the test only for some interior set of signals resulting in two diagnostic cut-offs. If patients are responsive to the testing practices of physicians, then competition creates the same wedge in payments that is created with the choice of treatment and no diagnostic testing. When the benefit to the physician of performing the test is too strong the insurer will not be able to exert control over the physician through its choice of payments. Lastly, the insurer will have to compensate the physician more to induce the test when the test is costly to the patient.

These results indicate that competitive pressures can undermine the incentives generated by an insurer's payments. At the extreme competition completely removes any capacity for the insurer to control the physicians' treatment practices using simple payment rules and additional instruments such as utilization review, which removes the agency problem altogether, will be needed. Because utilization review is both common and costly, it would be particularly worthwhile to analyze how the practice can be optimally utilized. For example, can an insurer achieve sufficient control over physicians by instituting a random review process? Insurers also use networks and the threat of exclusion to increase their bargaining power vis-á-vis providers—particularly hospitals—but can the threat of exclusion based on higher costs also be used to better control physicians without exercising utilization review? In a study on prescription drug choice, Hellerstein (1998) found that physicians were more likely to prescribe a generic version of drug when patients are insured by HMOs suggesting that either through utilization review or selective contracting HMOs may be able to exert more control over physicians than indemnity insurers, though there are likely important demand-side differences between the two insurance plans as well.

REFERENCES

Allard, Marie, Pierre Thomas Léger, and Lise Rochaix, "Provider Competition in a Dynamic Setting," *Journal of Economics and Management Strategy*, 2009, *18* (2), 457–486.

Auster, Richard and Ronald Oaxaca, "Identification of Supplier Induced Demand in the Health Care Sector," *Journal of Human Resources*, 1981, *16* (3), 327–342.

Berry, Emily, "What do patients really want from you?," Technical Report, American Medical Association April 2007. http://www.ama-assn.org/amednews/2009/04/27/bil20427.htm

(accessed 06/01/2012).

Chalkley, Martin and James M. Malcomson, "Contracting for Health Services when Patient Demand Does Not Reflect Quality," *Journal of Health Economics*, 1998, *17*, 1–19.

Chernew, Michael E., William E. Encinosa, and Richard A. Hirth, "Optimal health insurance: the case of observable, severe illness," *Journal of Health Economics*, 2000, *19*, 585–609.

Choné, Philippe and Ching-To Albert Ma, "Optimal Health Care Contract under Physician Agency," 2011. forthcoming Annales and d' Économie et de Statistique.

Dranove, David, "Demand Inducement and the Physician/patient Relationship," *Economic Inquiry*, 1988, *26*, 251–298.

Dunn, Abe and Adam Hale Shapiro, "Physician Market Power and Medical-Care Expenditures," 2012. BEA Working Paper (WP2012-6).

Ellis, Randall P., "Creaming, Skimping and Dumping: Provider Competition on the Intensive and Extensive Margins," *Journal of Health Economics*, 1998, *17*, 537–555.

_____ **and Thomas G. McGuire**, "Provider Behavior under Prospective Reimbursement," *Journal of Health Economics*, 1986, *5*, 129–151.

_____ **and** _____, "Optimal Payment Systems for Health Services," *Journal of Health Economics*, 1990, *9*, 375–396.

Fuchs, Victor, "The Supply of Surgeons and the Demand for Operations," *Journal of Human Resources*, 1978, *13*, 35–56.

Gal-Or, Esther, "Optimal Reimbursement and Malpractice Sharing Rules in Health Care Markets," *Journal of Regulatory Economics*, 1999, *16*, 237–265.

Gaynor, Martin, "Issues in the Industrial Organization of the Market for Physician Services," April 1994. NBER Working Paper No. 4695.

Givens, John T., "Thirteen Reasons Why Patients Change Doctors," *Journal of the National Medical Association*, 1957, *49* (3), 174–175.

Gruber, Jon, John Kim, and Dina Mayzlin, "Physician fees and procedure intensity: the case of cesarean delivery," *Journal of Health Economics*, 1999, *18*, 473–490.

Gruber, Jonathan and M. Owings, "Physician Financial Incentives and Cesarean Section Delivery," *Rand Journal of Economics*, 1996, *27*, 99–123.

Hellerstein, Judith K., "The Importance of the Physician in the Generic versus Trade-Name Prescription Decision," *RAND Journal of Economics*, 1998, *29* (1), 108–136.

Jack, William, "Purchasing health care services from providers with unknown altruism," *Journal of Health Economics*, 2005, *24*, 73–93.

Léger, Pierre Thomas, *Physician Payment Mechanisms*, Wiley-VCH Verlag GmbH & Co. KGaA,

Lehnert, Bruce E. and Robert L. Bree, "Analysis of Appropriateness of Outpatient CT and MRI Referred From Primary Care Clinics at an Academic Medical Center: How Critical Is the Need for Improved Decision Support?," *Journal of the American College of Radiology*, 2010, *7* (3), 192–197.

Liu, Ting and Albert Ching-To Ma, "Health Insurance, Treatment Plan, and Delegation to Altruistic Physician," June 2011. Unpublished Manuscript.

Ma, Albert Ching-To and Thomas G. McGuire, "Optimal Health Insurance and Provider Payment," *American Economic Review*, 1997, *87* (4), 685–704.

Ma, Ching-To Albert and Michael H. Riordan, "Health Insurance, Moral Hazard, and Managed Care," *Journal of Economics and Management Strategy*, 2002, *11* (1), 81–107.

Macpherson, Alison K, Michael S Kramer, Francine M Ducharme, Hong Yang, and François P BÍanger, "Doctor shopping before and after a visit to a paediatric emergency department," *Pediatrics and Child Health*, 2001, *6* (6), 341–346.

McCarthy, Thomas R., "The Competitive Nature of the Primary-Care Physician Market," *Journal of Health Economics*, 1985, *4*, 93–117.

McGuire, Thomas G., *Handbook of Health Economics*, Vol. 1, Elsevier Science B.V.,

Mendel, Rosmarie, Eva Traut-Mattausch, Dieter Frey, Markus Bühner, Achim Berthele, Werner Kissling, and Johannes Hamann, "Do physicians recommendations pull patients away from their preferred treatment options?," *Health Expectations*, 2012, *15* (1), 23–31.

Pauly, Mark V., "The economics of moral hazard," *American Economic Review*, 1968, *58* (3), 531–537.

_____ **and Mark A. Satterthwaite**, "The pricing of Primary Care Physician Services: A Test of the Ruole of Consumer Information," *Bell Journal of Economics*, 1981, *12* (2), 488–506.

Phelps, Charles E., "Information diffusion and best practice adoption," in A. J. Culyer and J. P. Newhouse, eds., *Handbook of Health Economics*, Vol. 1 of *Handbook of Health Economics*, Elsevier, 2000, chapter 5, pp. 223–264.

Rochaix, Lisa, "Information Asymmetry and Search in the Market for Physicians' Services," *Journal of Health Economics*, 1989, *8*, 53–84.

Satterthwaite, Mark A., "Consumer Information, Equilibrium Industry Price and the Number of Sellers," *Bell Journal of Economics*, 1979, *10*, 483–502.

Selden, Thomas M., "A Model of Capitation," *Journal of Health Economics*, 1990, *9*, 397–409.

Smith-Bindman, Rebecca, Diana L. Miglioretti, and Eric B. Larson, "Rising Use of Diagnostic Medical Imaging In A Large Integrated Health System," *Health Affairs*, 2008, *27* (6), 1491–1502.

Tamblyn, Robyn, Rejean Laprise, James A. Hanley, Michael Abrahamowicz, Susan Scott, Nancy Mayo, Jerry Hurley, Roland Grad, Eric Latimer, Robert Perreault, Peter McLeod, Allen Huang, Pierre Larochelle, and Louise Mallet, "Adverse Events Associated With Prescription Drug Cost-Sharing Among Poor and Elderly Persons," *Journal of the American Medical Association*, 2001, *285* (4), 421–429.

Trivedi, Amal N., William Rakowski, and John Z. Ayanian, "Effect of Cost Sharing on Screening Mammography in Medicare Health Plans," *New England Journal of Medicine*, 2008, *358* (4), 375–383.

Wong, Mitchell D., Ronald Andersen, Cathy D. Sherbourne, Ron D. Hays, , and Martin F. Shapiro, "Effects of Cost Sharing on Care Seeking and Health Status: Results From the Medical Outcomes Study," *Public Health*, 2001, *91* (11), 1889–1894.

Zeckhauser, Richard J., "Medical insurance: a case study of the tradeoff between risk spreading and appropriate incentives," *Journal of Economic Theory*, 1970, *2* (1), 10–26.

APPENDIX A. MATHEMATICAL PROOFS

PROPOSITION 1

Proof. The social planner's objective is to maximize social surplus subject to a physician participation constraint and budget balance. Because patients are risk averse and the physician is risk

neutral, the planner's problem can be expressed as maximizing patient utility subject to a break-even constraint for the physician. The planner's choice variables consist of the reimbursement rates r_1 and r_2 as well as the treatment plans, τ_k where $k \in \{0, 1, 2\}$. Therefore the planner's problem can be expressed as

$$\max_{\{P,\tau_0,\tau_1,\tau_2\}} \int \big[U(Y-P) - L(\theta) + \sum_k \tau_k(\theta)\psi(\theta, T_k)\big] dF(\theta), \tag{A-1}$$

subject to a budget balance constraint

$$P = \int \sum_k \tau_k(\theta) c_k(\theta) dF(\theta),$$

and boundary conditions $0 \leq \tau_k(\theta) \leq 1$ for all $\theta \in \Theta$ and $k \in \{1, 2, 3\}$.

Ignoring the boundary conditions for the moment, the first-order conditions yield

$$-U'(Y-P) - \lambda = 0 \text{ and} \tag{A-2}$$

$$\psi(\theta, T_k) + \lambda c_k(\theta) = 0,\ k \in \{0, 1, 2\}, \tag{A-3}$$

where λ is the Lagrangian multiplier. Combining the FOCs yields the following condition:

$$\psi(\theta, T_k) = U'(Y-P)c_k(\theta),\ k \in \{0, 1, 2\}. \tag{A-4}$$

Notably, (A-4) does not depend on the value for τ_k thus we can choose the treatment T_k that result in the highest utility for each θ; i.e., the treatment resulting in the largest value for (A-3). If (A-3) is not positive for any T_k then no treatment is chosen.

By A1 – A3, if $\theta^* < \overline{\theta}$, then there exists some θ^{**} such that treatment T_2 provides the highest value for all types above θ^{**} and $\psi(\theta^{**}, T_1) - U'(Y-P)c_1(\theta^{**}) = \psi(\theta^{**}, T_2) - U'(Y-P)c_2(\theta^{**})$.

Lastly it is clear that the planner's problem is strictly quasi-concave thus P^* is unique, and because of the budget balance constraint must equal the total cost of treatment; i.e., $P^* = \int_{\theta^*}^{\theta^{**}} c_1(\theta) dF(\theta) + \int_{\theta^{**}}^{\overline{\theta}} c_2(\theta) dF(\theta)$. □

PROPOSITION 2

Proof. The insurer's problem is

$$\max_{P,\theta^*} \int_{\underline{\theta}}^{\theta^*} \big[U(Y-P) - L(\theta) + \psi(\theta, T_1)\big] dF(\theta) + \int_{\theta^*}^{\overline{\theta}} \big[U(Y-P) - L(\theta) + \psi(\theta, T_2)\big] dF(\theta),$$

subject to budget balance

$$P = \int_{\underline{\theta}}^{\theta^*} r_1 dF(\theta) + \int_{\theta^*}^{\overline{\theta}} r_2 dF(\theta),$$

a physician participation constraint

$$\mathbb{E}_\theta[\Pi(\theta \mid \theta^*)] = \int_{\underline{\theta}}^{\theta^*} [r_1 - c_1(\theta)] dF(\theta) + \int_{\theta^*}^{\overline{\theta}} [r_2 - c_2(\theta)] dF(\theta) = 0,$$

and the boundary conditions $\theta^* \in \Theta$ and $r_1, r_2 \geq 0$.

Given a payment rule $\{r_1, r_2\}$ the physician will choose the cut-off that maximizes her profit; thus, the insurer acts as a Stackelberg leader when establishing the payment rule. The first order condition of the physician's optimization problem shows that given reimbursement rates r_1 and r_2 her treatment practice will be defined by the cut-off $\theta \in \Theta$ with which she is indifferent between treatments; i.e., the FOC of the physician's optimization problem along with the monotonicity of the cost functions indicates that the physician will treat all patients of type less than $\hat{\theta}$ with T_1 and all types higher than $\hat{\theta}$ with T_2 where $\hat{\theta}$ solves

$$r_1 - c_1(\hat{\theta}) = r_2 - c_2(\hat{\theta}). \tag{A-5}$$

Using (A-5), the insurer's optimization program can be expressed as

$$\max_{P,\theta^*} \int_{\underline{\theta}}^{\theta^*} \big[U(Y-P) - L(\theta) + \psi(\theta, T_1)\big] dF(\theta) + \int_{\theta^*}^{\overline{\theta}} \big[U(Y-P) - L(\theta) + \psi(\theta, T_2)\big] dF(\theta),$$

subject to

$$P = \int_{\underline{\theta}}^{\theta^*} [r_2 + c_1(\theta^*) - c_2(\theta^*)] dF(\theta) + \int_{\theta^*}^{\overline{\theta}} r_2 dF(\theta), \tag{A-6}$$

(A-7)

$$\mathbb{E}_\theta[\Pi(\theta \mid \theta^*)] = \int_{\underline{\theta}}^{\theta^*} [r_2 + c_1(\theta^*) - c_2(\theta^*) - c_1(\theta)] dF(\theta) + \int_{\theta^*}^{\overline{\theta}} [r_2 - c_2(\theta)] dF(\theta) = 0,$$

and the boundary conditions $\theta^* \in \Theta$ and $r_1, r_2 \geq 0$.

The physician's FOC, Eq. (A-5), pins down r_1 relative to r_2 and the physician's participation constraint, Eq. (A-7), can be used to pin down r_2 relative to θ^E resulting in the payment rules identified in the Proposition.

Lastly, plugging these payments into P and taking the first-order condition of the insurer's maximization problem reveals that $\theta^* = \theta^E$ thus the physician is held to zero profit and the cut-off is equal to the socially-efficient cut-off. □

PROPOSITION 3

Proof. If negative reimbursements are not admissible, then the insurer's boundary condition for r_1 can be expressed as

$$r_1 = \mathbb{E}_\theta[c(\theta \mid \theta^*)] + \big[1 - F(\theta^*)\big]\big[c_1(\theta^*) - c_2(\theta^*)\big] \geq 0.$$

When the constraint is not binding the optimal payment rules are as defined in Proposition 2. When the constraint is binding, though, the insurer's optimization program can be more simply expressed as

$$\max_{P,\theta^*} \int_{\underline{\theta}}^{\theta^*} \big[U(Y-P) - L(\theta) + \psi(\theta, T_1)\big] dF(\theta) + \int_{\theta^*}^{\overline{\theta}} \big[U(Y-P) - L(\theta) + \psi(\theta, T_2)\big] dF(\theta),$$

subject to $P = \int_{\theta^*}^{\overline{\theta}} \big[c_2(\theta^*) - c_1(\theta^*)\big] dF(\theta)$ and $\mathbb{E}_\theta[\Pi(\theta \mid \theta^*)] \geq 0$.

When the physician's participation constraint binds θ^* solves

$$[1 - F(\theta^*)][c_2(\theta^*) - c_1(\theta^*)] = \mathbb{E}_\theta[c(\theta \mid \theta^*)].$$

However, this can only occur when r_1^*, defined in Proposition 2, is zero at θ^{FB}. Otherwise the physician must be paid more and her participation constraint will not bind.

Removing the physician's participation constraint from the insurer's optimization program and taking the first-order condition reveals that the second-best cut-off is the θ^{SB} satisfying

$$\begin{aligned}\psi(\theta^{SB}, T_1) - U'(Y-P)\left[c_1(\theta^{SB}) - \frac{1-F(\theta^{SB})}{f(\theta^{SB})}c_1'(\theta^{SB})\right] = \\ \psi(\theta^{SB}, T_2) - U'(Y-P)\left[c_2(\theta^{SB}) - \frac{1-F(\theta^{SB})}{f(\theta^{SB})}c_2'(\theta^{SB})\right].\end{aligned} \tag{A-8}$$

Because $c_1'(\cdot) > c_2'(\cdot)$ and $F(\cdot)$ satisfies the monotone hazard rate property the insurer's problem is concave in θ^{SB} as long as $c_1''(\cdot)$ is not significantly greater than $c_2''(\cdot)$. This is because quasi concavity of the insurer's optimization problem requires

$$\begin{aligned}&- U''(Y-P)\big[[c_1'(\theta) - c_2'(\theta)][1-F(\theta)] - [c_1(\theta) - c_2(\theta)]f(\theta)\big]^2 \\ &- \big[\psi_\theta(\theta, T_1) - \psi_\theta(\theta, T_2)\big] + U'(Y-P) \\ &\times \left[c_1'(\theta) - c_2'(\theta) - \frac{d}{d\theta}\left\{\frac{1-F(\theta)}{f(\theta)}\right\}\big[c_1'(\theta) - c_2'(\theta)\big] - \left(\frac{1-F(\theta)}{f(\theta)}\right)\big[c_1''(\theta) - c_2''(\theta)\big]\right] \\ &> 0.\end{aligned}$$

Every term except for $-U'(Y-P)\left(\frac{1-F(\theta)}{f(\theta)}\right)\big[c_1''(\theta) - c_2''(\theta)\big]$ is positive so the insurer's optimization program is quasi concave as long as $c_1''(\theta) - c_2''(\theta)$ is not too large. For linear cost functions, or for convex cost functions where $c_2'' > c_1''$, this property is trivially satisfied.

If the insurer's problem is a well-behaved quasi concave program then it must be the case that $\theta^{SB} > \theta^{FB} \equiv \theta^E$.

I have assumed that $c_1(\theta^E) < c_2(\theta^E)$, however, if instead $c_1(\theta^E) > c_2(\theta^E)$ so that more patients prefer treatment T_1 than is socially optimal, then it may be the case that $r_2^* < 0$. Following the same arguments as above the second-best cut-off is the θ^{SB} satisfying

$$\begin{aligned}\psi(\theta^{SB}, T_1) - U'(Y-P)\left[c_1(\theta^{SB}) + \frac{F(\theta^{SB})}{f(\theta^{SB})}c_1'(\theta^{SB})\right] = \\ \psi(\theta^{SB}, T_2) - U'(Y-P)\left[c_2(\theta^{SB}) + \frac{F(\theta^{SB})}{f(\theta^{SB})}c_2'(\theta^{SB})\right].\end{aligned} \tag{A-9}$$

Because $F(\cdot)$ satisfies the monotone hazard rate property the insurer's problem is concave in θ^{SB} without the added restriction that $c_1''(\theta) - c_2''(\theta)$ not be too large. Consequently it must be the case that $\theta^{SB} < \theta^{FB} \equiv \theta^E$. □

PROPOSITION 4

Proof. Because physicians compete against one another to attract patients the equilibrium is not defined by the first order condition of their objective function. Rather, a cut-off θ^* and payment rule $\{r_1, r_2\}$ is an equilibrium if and only if a physician cannot increase her profit by unilaterally deviating from θ^*. That is, let $\Pi(\theta \mid \theta_{-j})$ represent physician j's profit when she chooses cut-off θ and all of the other physicians choose cut-off θ_{-j}. Then θ^* is an equilibrium cut-off if and only if $\Pi(\theta^* \mid \theta_{-j} = \theta^*) \geq \Pi(\theta \mid \theta_{-j} = \theta^*)$ for all $\theta \neq \theta^*$ and for all physicians j.

For $\theta^* > \theta^P$ the physicians have three possible deviations. A physician may deviate upwards, or a physician may deviate downwards to a cut-off that is still greater than θ^P or to a cut-off that is below θ^P. The expected profit of a physician who deviates to a higher cut-off given the other physicians choose cut-off θ^* is

$$\begin{aligned}\mathbb{E}_\theta[\Pi_i(\hat{\theta} \mid \theta_{-i} = \theta^*, \theta^P \leq \theta^* \leq \hat{\theta})] = \\ \frac{1}{n}\int_{\underline{\theta}}^{\theta^*} \big(r_1 - c_1(\theta)\big) dF(\theta) + \big(1 - \phi(n)\big)\frac{1}{n}\int_{\theta^*}^{\hat{\theta}} \big(r_1 - c_1(\theta)\big) dF(\theta) \\ + \frac{1}{n}\int_{\hat{\theta}}^{\overline{\theta}} \big(r_2 - c_2(\theta)\big) dF(\theta).\end{aligned} \tag{A-10}$$

Eq. (A-10) shows that a physician who deviates upward will lose $\frac{\phi(n)}{n}\big[F(\hat{\theta}) - F(\theta^*)\big]$ patients, but will now treat $\big[F(\hat{\theta}) - F(\theta^*)\big](1 - \phi(n))/n$ with T_1 instead of T_2. Because $c_1'(\cdot) \geq c_2'(\cdot)$ the physician's change in profit from increasing her cut-off is monotonically decreasing in θ. Consequently, to check that an upward deviation is not profitable it is sufficient to check that it is not profitable to deviate upward at θ^*; i.e., the following is a necessary condition for an equilibrium at θ^*:

$$\frac{d}{d\hat{\theta}}\big\{\mathbb{E}_\theta[\Pi_i(\hat{\theta} \mid \theta_{-i} = \theta^*, \theta^P \leq \theta^* \leq \hat{\theta})]\big\}\Big|_{\hat{\theta}=\theta^*} \leq 0. \tag{A-11}$$

Evaluating (A-11) yields the condition

$$\big(1 - \phi(n)\big)\big(r_1 - c_1(\theta^*)\big) \leq r_2 - c_2(\theta^*). \tag{A-12}$$

Observe that when $\phi(n) = 0$, this condition is equivalent to the cut-off condition for the monopolist physician.

Next, the expected profit for a physician who deviates downwards, but chooses a $\hat{\theta} \geq \theta^P$, is

$$\begin{aligned}\mathbb{E}_\theta[\Pi_i(\hat{\theta} \mid \theta_{-i} = \theta^*, \theta^P \leq \hat{\theta} \leq \theta^*)] = \\ \frac{1}{n}\int_{\underline{\theta}}^{\hat{\theta}} \big(r_1 - c_1(\theta)\big) dF(\theta) + \left[\phi(n)\left(\frac{n-1}{n}\right) + \frac{1}{n}\right]\int_{\hat{\theta}}^{\theta^*} \big(r_2 - c_2(\theta)\big) dF(\theta) \\ + \frac{1}{n}\int_{\theta^*}^{\overline{\theta}} \big(r_2 - c_2(\theta)\big) dF(\theta).\end{aligned}$$

A physician who deviates down can thus expect to attract an additional $\phi(n)(1 - n^{-1})\big[F(\theta^*) - F(\hat{\theta})\big]$ patients since she chooses a more preferred treatment for types $\hat{\theta}$ to θ^*. Again, because $c_1'(\cdot) \geq c_2'(\cdot)$, it is sufficient to check that a downward deviation at θ^* is not profitable; i.e.,

$$\frac{d}{d\hat{\theta}}\big\{\mathbb{E}_\theta[\Pi_i(\hat{\theta} \mid \theta_{-i} = \theta^*, \theta^P \leq \hat{\theta} \leq \theta^*)]\big\}\Big|_{\hat{\theta}=\theta^*} \geq 0. \tag{A-13}$$

Evaluating (A-13) yields the condition

$$r_1 - c_1(\theta^*) \geq \big(1 + \phi(n)(n-1)\big)\big(r_2 - c_2(\theta^*)\big). \tag{A-14}$$

Again, when $\phi(n) = 0$ this condition is equivalent to the cut-off condition for the monopolist physician.

Lastly, a physician could choose to deviate to a cut-off below the patients' preferred cut-off θ^P, giving the expected profit:

$$\begin{aligned}\mathbb{E}_\theta[\Pi_i(\hat{\theta} \mid \theta_{-i} =&\theta^*, \hat{\theta} \leq \theta^P \leq \theta^*)] = \\ &\frac{1}{n}\int_{\underline{\theta}}^{\hat{\theta}} \big(r_1 - c_1(\theta)\big)dF(\theta) + \frac{1-\phi(n)}{n}\int_{\hat{\theta}}^{\theta^P} \big(r_2 - c_2(\theta)\big)dF(\theta) \\ &+ \left[\phi(n)\left(\frac{n-1}{n}\right) + \frac{1}{n}\right]\int_{\theta^P}^{\theta^*} \big(r_2 - c_2(\theta)\big)dF(\theta) \\ &+ \frac{1}{n}\int_{\theta^*}^{\overline{\theta}} \big(r_2 - c_2(\theta)\big)dF(\theta).\end{aligned} \tag{A-15}$$

By deviating to a cut-off that is below the patients' preferred cut-off θ^P, the physician gains $\phi(n)(1-n^{-1})\big[F(\theta^*) - F(\theta^P)\big]$ patients because she will utilize their preferred treatment when the other physicians do not but also loses $\big[F(\theta^P) - F(\hat{\theta})\big]\big(1-\phi(n)\big)/n$ patients because the other physicians treatment practice is preferred by these types. I have established that a physician will not deviate to θ^P if condition (A-13) is satisfied, thus I need only to check whether or not a physician's profit increases by deviating even lower to some $\hat{\theta} < \theta^P$. Once again, because $c_1'(\cdot) \geq c_2'(\cdot)$, it is sufficient to check that

$$\frac{d}{d\hat{\theta}}\big\{\mathbb{E}_\theta[\Pi_i(\hat{\theta} \mid \theta_{-i} = \theta^*, \hat{\theta} < \theta^P \leq \theta^*)]\big\}\Big|_{\hat{\theta}=\theta^P} \geq 0. \tag{A-16}$$

Evaluating (A-16) yields the condition $r_1 - c_1(\theta^P) \geq \big(1-\phi(n)\big)\big(r_2 - c_2(\theta^P)\big)$, which is true whenever condition (A-14) is true. Combining conditions (A-12) and (A-14) yield condition (4) of the Proposition. Lastly, I have shown that conditions (A-12) and (A-14) are necessary for θ^* to be an equilibrium cut-off; however, combined with the condition that the physicians' expected profit from choosing θ^P is positive thus the physician accepts the contract, it is clear that if the physician has no incentive to deviate at θ^* then θ^* must be an equilibrium establishing sufficiency. □

PROPOSITION 5

Proof. A payment rule can satisfy condition (4) in Proposition 4 only if $1 + \phi(n)(n-1) \leq \big(1-\phi(n)\big)^{-1}$. Rearranging this inequality shows that that the condition cannot be satisfied whenever $0 < \phi(n) < \frac{n-2}{n-1}$. Consequently the payments can induce the physicians to not want to deviate to a higher cut-off or towards lower cut-off, but not both simultaneously and $\theta^* \in (\theta^P, \overline{\theta})$ cannot represent a pure-strategy equilibrium. The insurer can induce $\overline{\theta}$ by setting $r_2 = 0$ and r_1 so that it satisfies $r_1 - c_1(\overline{\theta}) \geq -\big(1 + \phi(n)(n-1)\big)c_2(\overline{\theta})$. Lastly, the insurer can induce θ^P by setting r_1 and r_2 so that $1 - \phi(n) \leq \frac{r_1 - c_1(\theta^P)}{r_2 - c_2(\theta^P)} \leq \big(1-\phi(n)\big)^{-1}$. □

PROPOSITION 6

Proof. Given a payment rule $\{r_1, r_2\}$, the Pareto dominant equilibrium is the θ where

$$1 + \phi(n)(n-1) = \frac{r_1 - c_1(\theta)}{r_2 - c_2(\theta)}. \tag{A-17}$$

Because $c_1'(\cdot) > c_2'(\theta)$ anything that increases the left-hand-side of (A-17) lowers θ and anything that lowers the left-hand-side increases the equilibrium θ. The left-hand-side is increasing with $\phi(n)$ resulting in the first comparative static. Similarly when $\phi'(n)(n-1)+\phi(n) > 0$ the left-hand-side of (A-17) increases with n resulting in a lower equilibrium θ; and when $\phi'(n)(n-1)+\phi(n) < 0$ the left-hand-side of (A-17) decreases with n resulting in a higher equilibrium θ.

From the insurer's maximization program, if the physician is held to zero profit, then r_2^* must satisfy

$$r_2^*\big[1+\phi(n)(n-1)F(\theta^*)\big] = \mathbb{E}_\theta[c(\theta \mid \theta^*)] + F(\theta^*)\big[c_2(\theta^*) - c_1(\theta^*)\big] + \phi(n)(n-1)F(\theta^*)c_2(\theta^*). \tag{A-18}$$

Totally differentiating (A-18) yields

$$\frac{dr_2^*}{d\phi(n)} = \frac{-F(\theta^*)(n-1)\big(r_2^* - c_2(\theta^*)\big)}{1+\phi(n)(n-1)F(\theta^*)} < 0. \tag{A-19}$$

Similarly totally differentiating (A-18) with respect to r_2^* and n yields

$$\frac{dr_2^*}{dn} = \frac{-F(\theta^*)\big[\phi'(n)(n-1)+\phi(n)\big]\big[r_2 - c_2(\theta^*)\big]}{1+\phi(n)(n-1)F(\theta^*)}. \tag{A-20}$$

The RHS of (A-20) is positive whenever $\phi'(n)(n-1)+\phi(n) < 0$ and negative whenever $\phi'(n)(n-1)+\phi(n) > 0$.

Lastly, because θ^* is the same for every n, the total expected cost per patient to the physician remains the same for every n. Thus, because the physician is held to zero profit, an increase in r_2^* implies a decrease in r_1^* and vice versa. □

PROPOSITION 7

Proof. The insurer's problem can be expressed as

$$\max_{\{P,\sigma_1,\sigma_2,\sigma_3,\tau_1,\tau_2\}} \int_\Xi \int_\Theta \big\{U(Y-P) - L(\theta) + \sigma_1(\xi)\psi(\theta,T_1) + \sigma_2(\xi)\psi(\theta,T_2) + \sigma_3(\xi)\sum_\kappa \big[\tau_k(\theta)\psi(\theta,T_k)\big]\big\} dG(\theta \mid \xi) d\Gamma(\xi),$$

subject to

$$P = \int_\Xi \int_\Theta \big\{\sigma_1(\xi)c_1(\theta) + \sigma_2(\xi)c_2(\theta) + \sigma_3(\xi)\sum_\kappa \big[\tau_k(\theta)c_k(\theta) + D\big]\big\} dG(\theta \mid \xi) d\Gamma(\xi),$$

and $\sigma_j \in [0,1]$ for $j = 1,2,3$ and $\tau_k \in [0,1]$ for $k = 1,2$, where the σ_j and τ_k represent the probability of following the particular treatment practice; i.e., $\sigma_1(\xi)$ is the probability of treating the patient with T_1 without performing a diagnostic test, $\sigma_2(\xi)$ is the probability of treating the patient with T_2 without performing a diagnostic test, and $\sigma_3(\xi)$ is the probability of performing the diagnostic test, all given a particular ξ.

Ignoring the boundary conditions and using pointwise optimization, the FOCs for σ_1 and σ_2 yield:

$$\int_\Theta \big[\psi(\theta,T_j) - \lambda c_j(\theta)\big] dG(\theta \mid \xi) = 0 \text{ for } j \in \{1,2\}. \tag{A-21}$$

The FOC for σ_3 is:

$$\text{(A-22)} \qquad \int_{\Theta} \left[\sum_k \tau_k(\theta)\psi(\theta, T_k) - \lambda \sum_k \tau_k(\theta)[c_k(\theta) + D] \right] dG(\theta \mid \xi) = 0,$$

and the FOC for P is

$$\text{(A-23)} \qquad -U'(Y - P) - \lambda = 0,$$

where λ is the Lagrangian multiplier. The FOCs are all independent of the σ_j thus we can choose the treatment diagnostic and treatment plan that results in the highest utility for each ξ; i.e., the diagnostic and treatment plan resulting in the largest value for (A-21) and (A-22). Eq. (A-22) assumes its maximal value when $\tau_1(\theta) = 1$ for all $\theta < \theta^E$, $\tau_1(\theta) = 0$ otherwise, and $\tau_2(\theta) = 1 - \tau_1(\theta)$ for all $\theta \in \Theta$.

The lower diagnostic cut-off, ξ^*, represents the signal with which the FOCs for σ_1 and σ_3 are equal:

$$\text{(A-24)} \qquad \begin{aligned} &\int_{\underline{\theta}}^{\overline{\theta}} [\psi(\theta, T_1) - \lambda c_1(\theta)] dG(\theta \mid \xi^*) = \\ &\int_{\underline{\theta}}^{\theta^E} [\psi(\theta, T_1) - \lambda c_1(\theta)] dG(\theta \mid \xi^*) + \int_{\theta^E}^{\overline{\theta}} [\psi(\theta, T_2) - \lambda c_2(\theta)] dG(\theta \mid \xi^*) - \lambda D, \end{aligned}$$

and the upper diagnostic cut-off, ξ^{**}, represents the signal with which the FOCs for σ_2 and σ_3 are equal:

$$\text{(A-25)} \qquad \begin{aligned} &\int_{\underline{\theta}}^{\overline{\theta}} [\psi(\theta, T_2) - \lambda c_2(\theta)] dG(\theta \mid \xi^{**}) = \\ &\int_{\underline{\theta}}^{\theta^E} [\psi(\theta, T_1) - \lambda c_1(\theta)] dG(\theta \mid \xi^{**}) + \int_{\theta^E}^{\overline{\theta}} [\psi(\theta, T_2) - \lambda c_2(\theta)] dG(\theta \mid \xi^{**}) - \lambda D. \end{aligned}$$

If either one of the signals $\xi^*, \xi^{**} \in \Xi$ satisfying (A-24) and (A-25) do not exist then it is socially wasteful to conduct the diagnostic test and the socially optimal treatment is the one generating the highest expected social value given the signal.

□

LEMMA 2

Proof. If all other physicians choose some upper cut-off $\xi^u_{-i} = \xi^{**}$ then the profit for a physician who deviates to a higher upper cut-off is expressed as

$$\begin{aligned}\mathbb{E}_{\theta,\xi}[\Pi_i(\hat{\xi}^u \mid \xi^u_{-i} = \xi^{**}, \hat{\xi}^u \geq \xi^{**})] = \\ &\frac{1}{n}\int_{\underline{\xi}}^{\xi^*} \mathbb{E}_\theta[\Pi(\theta \mid \overline{\theta}) \mid \xi] d\Gamma(\xi) + \frac{1}{n}\int_{\xi^*}^{\xi^{**}} \big\{\mathbb{E}_\theta[\Pi(\theta \mid \theta^*) \mid \xi)] - D\big\} d\Gamma(\xi) \\ &+ \left[\phi(n)\left(\frac{n-1}{n}\right) + \frac{1}{n}\right] \int_{\xi^{**}}^{\hat{\xi}^u} \big\{\mathbb{E}_\theta[\Pi(\theta \mid \theta^*) \mid \xi)] - D\big\} d\Gamma(\xi) \\ &+ \frac{1}{n}\int_{\xi^{**}}^{\overline{\xi}} \mathbb{E}_\theta[\Pi(\theta \mid \underline{\theta}) \mid \xi] d\Gamma(\xi).\end{aligned}$$

As with a deviation in the treatment cut-off (Proposition 4) the deviation in the upper diagnostic cut-off results in an increase in demand because patients always prefer more diagnostic testing. Equilibrium requires that a physician is not better off from any deviation; i.e., the following is a necessary condition for equilibrium:

$$\frac{d}{d\hat{\xi}^u}\big\{\mathbb{E}_{\theta,\xi}[\Pi_i(\hat{\xi}^u \mid \xi^u_{-i} = \xi^{**}, \hat{\xi}^u \geq \xi^{**})]\big\}\Big|_{\hat{\xi}^u = \xi^{**}} \leq 0 \tag{A-26}$$

Evaluating (A-26) yields the condition

$$\big(\phi(n)(n-1)+1\big)\big[\mathbb{E}_\theta[\Pi(\theta \mid \theta^*) \mid \xi^{**}] - D\big] \leq \mathbb{E}_\theta[\Pi(\theta \mid \underline{\theta}) \mid \xi^{**}]. \tag{A-27}$$

The relationship between the treatment reimbursements r_1 and r_2 are limited by the physician's first-order condition if the insurer wants to induce some θ^*. Specifically, Proposition 4 along with the assumption that the physicians choose the un-dominated equilibrium implies that $\big(1 + \phi(n)(n-1)\big)\big(r_2 - c_2(\theta^*)\big) = r_1 - c_1(\theta^*)$. The lower possible revenue occurs when $r_1 = 0$, thus plugging in $r_1 = 0$ and $r_2 = c_2(\theta^*) - c_1(\theta^*)/\big(1 + \phi(n)(n-1)\big)$ into (A-27) yields

$$\begin{aligned}c_2(\theta^*) - \frac{c_1(\theta^*)}{1+\phi(n)(n-1)} - \mathbb{E}_\theta[c(\theta \mid \underline{\theta}) \mid \xi^{**}] \geq \\ &\big(1+\phi(n)(n-1)\big) \\ &\times \left[\big(1 - F(\theta^*)\big)\left[c_2(\theta^*) - \frac{c_1(\theta^*)}{1+\phi(n)(n-1)}\right] - \mathbb{E}_\theta[c(\theta \mid \theta^*) \mid \xi^{**}] - D\right]\end{aligned}$$

Rearranging results in (6). □

LEMMA 3

Proof. If all other physicians choose some lower cut-off $\xi^l_{-i} = \xi^*$ then the profit for a physician who deviates to a lower lower cut-off is expressed as

$$\begin{aligned}\mathbb{E}_{\theta,\xi}[\Pi_i(\hat{\xi}^l \mid \xi^l_{-i} = \xi^*, \hat{\xi}^l \leq \xi^*)] = \\ &\frac{1}{n}\int_{\underline{\xi}}^{\hat{\xi}^l} \mathbb{E}_\theta[\Pi(\theta \mid \overline{\theta}) \mid \xi] d\Gamma(\xi) \\ &+ \left[\phi(n)\left(\frac{n-1}{n}\right) + \frac{1}{n}\right]\int_{\hat{\xi}_l}^{\xi^*} \big\{\mathbb{E}_\theta[\Pi(\theta \mid \theta^*) \mid \xi)] - D\big\} d\Gamma(\xi) \\ &+ \frac{1}{n}\int_{\xi^*}^{\xi^{**}} \big\{\mathbb{E}_\theta[\Pi(\theta \mid \theta^*) \mid \xi)] - D\big\} d\Gamma(\xi) + \frac{1}{n}\int_{\xi^{**}}^{\overline{\xi}} \mathbb{E}_\theta[\Pi(\theta \mid \underline{\theta}) \mid \xi] d\Gamma(\xi).\end{aligned}$$

As with a deviation in the treatment cut-off (Proposition 4) the deviation in the lower diagnostic cut-off results in an increase in demand because patients always prefer more diagnostic testing. Equilibrium requires that a physician is not better off from any deviation; i.e., the following is a necessary condition for equilibrium:

$$\frac{d}{d\hat{\xi}^l}\big\{\mathbb{E}_{\theta,\xi}[\Pi_i(\hat{\xi}^l \mid \xi^l_{-i} = \xi^*, \hat{\xi}^l \leq \xi^*)]\big\}\Big|_{\hat{\xi}^l = \xi^*} \geq 0 \tag{A-28}$$

Evaluating (A-28) yields the condition

$$\big(\phi(n)(n-1)+1\big)\big[\mathbb{E}_\theta[\Pi(\theta \mid \theta^*) \mid \xi^*] - D\big] \leq \mathbb{E}_\theta[\Pi(\theta \mid \overline{\theta}) \mid \xi^*]. \tag{A-29}$$

The relationship between the treatment reimbursements r_1 and r_2 are limited by the physician's first-order condition if the insurer wants to induce some θ^*. Specifically, Proposition 4 along with the assumption that the physicians choose the un-dominated equilibrium implies that $\big(1 + \phi(n)(n-1)\big)\big(r_2 - c_2(\theta^*)\big) = r_1 - c_1(\theta^*)$. The lowest possible revenue occurs when $r_1 = 0$, thus plugging in $r_1 = 0$ and $r_2 = c_2(\theta^*) - c_1(\theta^*)/\big(1 + \phi(n)(n-1)\big)$ into (A-29) yields

$$\begin{aligned}\big(\phi(n)(n-1)+1\big)\big[\big(1 - F(\theta^*)\big)\left[c_2(\theta^*) - \frac{c_1(\theta^*)}{1+\phi(n)(n-1)}\right] \\ - \mathbb{E}_\theta[c(\theta \mid \theta^*) \mid \xi^*] - D\big] \leq -\mathbb{E}_\theta[c(\theta \mid \overline{\theta}) \mid \xi^*].\end{aligned}$$

Rearranging results in (7). □

PROPOSITION 8

Proof. First, $E_\theta[\Pi(\theta \mid r_1, r_2, \hat{r}_1, \hat{r}_2)] \geq 0$ is a standard individual rationality constraint. Of course it can always be satisfied with any payment rules if we also admit a lump-sum payment R. The conditions $r_1 > \hat{r}_1$ and $r_2 > \hat{r}_2$ follow from the fact that the physician cannot be compensated less for administering the diagnostic test. This is basically an incentive compatibility issue as the physician can always administer the test without reporting doing so to receive the higher payment.

If all other physicians have chosen some upper diagnostic cut-off $\xi^2_{-i} = \xi^{**}$ then the profit for a physician who deviates to a lower upper cut-off is expressed as

$$\begin{aligned}\mathbb{E}_{\xi,\theta}[\Pi_i(\hat{\xi}^2 \mid \xi^u_{-i} = \xi^{**}, \hat{\xi} \leq \xi^{**})] = \\ &\frac{1}{n}\int_{\underline{\xi}}^{\xi^*} \mathbb{E}_\theta[\Pi(\theta \mid \overline{\theta}) \mid \xi] d\Gamma(\xi) + \frac{1}{n}\int_{\xi^*}^{\hat{\xi}^2} \big\{\mathbb{E}_\theta[\Pi(\theta \mid \theta^*) \mid \xi] - D\big\} d\Gamma(\xi) \\ &+ \left(\frac{1-\phi(n)}{n}\right)\int_{\hat{\xi}^u}^{\xi^{**}} \mathbb{E}_\theta[\Pi(\theta \mid \underline{\theta}) \mid \xi] d\Gamma(\xi) \\ &+ \frac{1}{n}\int_{\xi^{**}}^{\overline{\xi}} \mathbb{E}_\theta[\Pi(\theta \mid \underline{\theta}) \mid \xi] d\Gamma(\xi).\end{aligned}$$

As with a deviation in the treatment cut-off (Proposition 4) the deviation in the upper diagnostic cut-off results in a loss in demand when patients prefer the diagnostic cut-off chosen by the other physicians. Equilibrium requires that a physician will not be better off from any deviation; i.e., the following is a necessary condition for equilibrium:

$$\frac{d}{d\hat{\xi}^u}\big\{\mathbb{E}_{\xi,\theta}[\Pi_i(\hat{\xi}^u \mid \xi^u_{-i} = \xi^{**}, \hat{\xi}^u \leq \xi^{**})]\big\}\Big|_{\hat{\xi}^u = \xi^{**}} \geq 0 \tag{A-30}$$

The derivative is positive because the physician is deviating downward. Evaluating (A-30) yields the condition

$$\mathbb{E}_\theta[\Pi(\theta \mid \theta^*) \mid \xi^{**}] - D \geq \big(1-\phi(n)\big)\mathbb{E}_\theta[\Pi(\theta \mid \underline{\theta}) \mid \xi^{**}]. \tag{A-31}$$

Moreover, the proof for lemma 2 shows that

$$\big(1+\phi(n)(n-1)\big)\big[\mathbb{E}_\theta[\Pi(\theta \mid \theta^*) \mid \xi^{**}] - D\big] \leq \mathbb{E}_\theta[\Pi(\theta \mid \underline{\theta}) \mid \xi^{**}]. \tag{A-32}$$

Combining (A-31) and (A-32) yields the condition

$$1 + \phi(n)(n-1) \leq \frac{\hat{r}_2 - \mathbb{E}_\theta[c(\theta \mid \underline{\theta}) \mid \xi^{**}]}{\mathbb{E}_\theta[\Pi(\theta \mid \theta^*) \mid \xi^{**}] - D} \leq \big(1-\phi(n)\big)^{-1}.$$

Next, if all other physicians have chosen some lower diagnostic cut-off $\xi^1_{-i} = \xi^*$ then the profit for a physician who deviates to a higher lower cut-off is expressed as

$$\begin{aligned}\mathbb{E}_{\xi,\theta}[\Pi_i(\hat{\xi}^l \mid \xi^l_{-i} = \xi^*, \hat{\xi} \leq \xi^*)] = \\ &\frac{1}{n}\int_{\underline{\xi}}^{\xi^*} \mathbb{E}_\theta[\Pi(\theta \mid \overline{\theta}) \mid \xi] d\Gamma(\xi) + \left(\frac{1-\phi(n)}{n}\right)\int_{\xi^*}^{\hat{\xi}^l} \mathbb{E}_\theta[\Pi(\theta \mid \overline{\theta}) \mid \xi] d\Gamma(\xi) \\ &+ \int_{\hat{\xi}^l}^{\xi^{**}} \big\{\mathbb{E}_\theta[\Pi(\theta \mid \theta^*) \mid \xi] - D\big\} d\Gamma(\xi) + \frac{1}{n}\int_{\xi^{**}}^{\overline{\xi}} \mathbb{E}_\theta[\Pi(\theta \mid \underline{\theta}) \mid \xi] d\Gamma(\xi).\end{aligned}$$

As with a deviation in the treatment cut-off (Proposition 4) the deviation in the lower diagnostic cut-off results in a loss in demand when patients prefer the diagnostic cut-off chosen by the other physicians. Equilibrium requires that a physician will not be better off from any deviation; i.e., the following is a necessary condition for equilibrium:

$$\frac{d}{d\hat{\xi}^l}\big\{\mathbb{E}_{\xi,\theta}[\Pi_i(\hat{\xi}^l \mid \xi^l_{-i} = \xi^*, \hat{\xi}^l \geq \xi^*)]\big\}\Big|_{\hat{\xi}^l = \xi^*} \leq 0 \tag{A-33}$$

Evaluating (A-33) yields the condition

$$\big(1-\phi(n)\big)\mathbb{E}_\theta[\Pi(\theta \mid \underline{\theta}) \mid \xi^*] \geq \mathbb{E}_\theta[\Pi(\theta \mid \theta^*) \mid \xi^*] - D. \tag{A-34}$$

Moreover, the proof for lemma 2 shows that

$$\big(1+\phi(n)(n-1)\big)\big[\mathbb{E}_\theta[\Pi(\theta \mid \theta^*) \mid \xi^*] - D\big] \leq \mathbb{E}_\theta[\Pi(\theta \mid \overline{\theta}) \mid \xi^*]. \tag{A-35}$$

Combining (A-34) and (A-35) yields the condition

$$1+\phi(n)(n-1) \leq \frac{\hat{r}_1 - \mathbb{E}_\theta[c(\theta \mid \overline{\theta}) \mid \xi^*]}{\mathbb{E}_\theta[\Pi(\theta \mid \theta^*) \mid \xi^*] - D} \leq \big(1-\phi(n)\big)^{-1}.$$

Lastly, the final expression of the Proposition follows directly from Proposition 4.

□

APPENDIX B. A PROFIT-MAXIMIZING MONOPOLY INSURER

If an insurer is operating in a perfectly competitive insurance market, then maximizing profit is the same as maximizing social surplus; i.e, the premium is driven down to marginal cost and the insurer maximizes consumer surplus. However, if the insurance market is monopolistically competitive, then the premium will be something above marginal cost and there will be a loss in consumer surplus. Because of the monotonicity of the difference in costs the insurer will still choose some cut-off (possibly a boundary) such that all patients having types below the cut-off are treated with T_1 and all types above are treated with T_2. Thus the profit-maximizing insurer's objective is to choose the premium level and cut-off type such that its profit is maximized subject to participation constraints for the patients and physicians. To explore how the insurer's cut-off is affected by its objective I take the extreme position and assume the insurer is a profit-maximizing monopoly insurer facing a monopolist physician.[29] In this case the insurer's optimization program can be expressed as

$$\max_{P,\theta^*} P - \int_{\underline{\theta}}^{\theta^*} r_1 dF(\theta) - \int_{\theta^*}^{\overline{\theta}} r_2 dF(\theta),$$

subject to

$$\begin{aligned}
&U(Y-P) - L(\theta) + \mathbb{E}_\theta[\psi(\theta \mid \hat{\theta})] \geq U(Y) - L(\theta),\\
&\mathbb{E}_\theta[\Pi(\theta \mid \theta^*)] = \int_{\underline{\theta}}^{\theta^*} \big(r_1 - c_1(\theta)\big)dF(\theta) + \int_{\theta^*}^{\overline{\theta}} \big(r_2 - c_2(\theta)\big)dF(\theta) \geq 0,\\
&r_1 - c_1(\theta^*) = r_2 - c_2(\theta^*), \text{ and}\\
&\theta^* \in \Theta.
\end{aligned}$$

The first two constraints are the participation constraints for the patients and physician, respectively. The third constraint reflects the relationship between r_1 and r_2 implied by the physician's optimization program.

[29] Strictly speaking I also assume that the insurer has all of the bargaining power such that the physician must accept whatever payment rule the insurer offers. Endowing the physician with some bargaining power will not alter the premium or cut-off chosen by the insurer, but it will result in a transfer of surplus from the insurer to the physician.

As reported by Proposition 2, when both of the physician payments used to induce the insurer's preferred cut-off are are positive then it must be the case that $\int_{\underline{\theta}}^{\theta^*} r_1 dF(\theta) + \int_{\theta^*}^{\overline{\theta}} r_2 dF(\theta) = \mathbb{E}_\theta[c(\theta \mid \theta^*)]$. Plugging $\mathbb{E}_\theta[c(\theta \mid \theta^*)]$ into the insurer's objective function reveals that its first-order conditions are the same as those for the social surplus-maximizing insurer. The difference between the two insurers, however, comes from the fact that the profit-maximizing insurer will extract the patients' surplus through the premium whereas the social surplus-maximizing insurer leaves the surplus with the patients. If the patients receive a positive surplus at the social optimum, this implies that the profit-maximizing insurer's premium is higher. As a consequence of the fact that a higher premium increases the patients' marginal utility and $c_1(\theta^E) < c_2(\theta^E)$, the profit-maximizing insurer will shift the cut-off higher to take advantage of the lower cost of treatment T_1. This is similar to a monopolist who reduces output and charges a higher price when costs increase. The following Proposition reports this result.

Proposition 9. *A profit-maximizing insurer will either select cut-off $\theta^* > \theta^E$ when the physician's participation constraint binds or $\theta^* > \theta^{SB}$ otherwise and induce the use of treatment T_1 on more patients than an insurer who maximizes total social welfare.*

Proof. There are two cases to consider. In the first case the physician's participation constraint is binding and the physician is left with zero profit. The optimal payment rule for a given θ^* is the rule identified in Proposition 2:

$$r_1^* = \mathbb{E}_\theta[c(\theta \mid \theta^E)] + \left[1 - F(\theta^E)\right]\left[c_1(\theta^E) - c_2(\theta^E)\right] \text{ and}$$
$$r_2^* = \mathbb{E}_\theta[c(\theta \mid \theta^E)] + F(\theta^E)\left[c_2(\theta^E) - c_1(\theta^E)\right].$$

Given these rules it is clear that $\int_{\underline{\theta}}^{\theta^*} r_1 dF(\theta) + \int_{\theta^*}^{\overline{\theta}} r_2 dF(\theta) = \int_{\underline{\theta}}^{\theta^*} c_1(\theta) dF(\theta) + \int_{\theta^*}^{\overline{\theta}} c_2(\theta) dF(\theta) = \mathbb{E}_\theta[c(\theta \mid \theta^*)]$. Plugging this into the insurer's optimization program allows it to be expressed as

$$\max_{P,\theta^*} P - \mathbb{E}_\theta[c(\theta \mid \theta^*)],$$

subject to $U(Y-P) - L(\theta) + \mathbb{E}_\theta[\psi(\theta \mid \hat{\theta})] \geq U(Y) - L(\theta)$ and $\theta^* \in \Theta$. Ignoring the boundary condition, the first-order conditions are

$$1 - \lambda U'(Y-P) = 0, \text{ and}$$
$$\lambda\psi(\theta^*, T_1) - c_1(\theta^*) = \lambda\psi(\theta^*, T_2) - c_2(\theta^*).$$

Rearranging the FOCs reveals that the premium and cut-off must satisfy the same relationship as with a social surplus-maximizing insurer:

$$\psi(\theta^*, T_1) - U'(Y-P)c_1(\theta^*) = \psi(\theta^*, T_2) - U'(Y-P)c_2(\theta^*).$$

The social surplus-maximizing insurer will set the premium to $P^{FB} = \mathbb{E}_\theta[c(\theta \mid \theta^E)]$ and leave all of the surplus with the patients. However, as long as patients receive some positive surplus, then the profit-maximizing insurer can extract that surplus by setting $P^* > P^{FB}$. Because $U(\cdot)$ is concave $U'(Y - P^*) > U'(Y - P^{FB})$. Furthermore, the concavity of the insurer's problem and the fact that $c_1(\theta^E) < c_2(\theta^E)$ together imply $\theta^*\ \theta^* > \theta^{FB}$.

When either the optimal r_1 or r_2 required to induce the desired cut-off θ^* is negative, then the physician must be left with some positive profit. In these cases the first-order condition will be identical to (A-8) and (A-9) in Proposition 3. However, the profit-maximizing insurer will be

able to charge a higher premium than the second-best premium when the patients are left with some positive surplus. Thus $P^* > P^{SB}$ resulting in $\theta^* > \theta^{SB}$. $\square$