PHYSICIAN INCENTIVES AND TREATMENT CHOICE*

KEVIN E. PFLUM†

ABSTRACT. I analyze the impact of physician competition for patients on treatment selection and an insurer's ability to induce its preferences through a supply-side payment mechanism. Informed patients choose the physician whose treatment practice best fits their preferences, aligning physician incentives with patient preferences. An insurer's ability to counter these incentives is not monotonic in how informed patients are, however. When demand is either perfectly inelastic to treatment practices because patients are completely uninformed or a sufficient proportion of the market is informed relative to the number of physicians in the market, then an insurer can induce its preferences through supply-side payment rules. Otherwise, more intensive policy levers such as utilization review must be employed. Programs that increase patient information can therefore improve efficiency despite generating stronger incentive to treat according to patient preferences. I also explore how noisy signals of illness type and diagnostic testing further complicate the insurer's problem.

1. INTRODUCTION

As consumers increasingly turn to online reviews for a variety of products it is no surprise that there are now dozens of websites providing reviews of physicians as well.[1] In addition to these third-party sites, some insurers are providing their own review systems to help enrollees select physicians based on the personal experiences of other enrollees. For example, WellPoint has partnered with Zagat—an established provider of restaurant and hotel reviews—to provide patient reviews of physicians for their 35 million enrollees.[2] Physician reviews have to some extent always been part of the business as patients have long shared their physician experiences with friends and family members; however, the wealth and availability of physician information provided by these new review systems is likely to have a much larger impact on how physicians compete for patients.

Date: Forthcoming in Journal of Economics and Management Strategy. Final Version: January 14, 2014.

*I am grateful to Paul Pecorino, Paan Jindapon, Paula Cordero Salas, Paul J. Healy, Suhui Li, Thomas Buchmueller and participants of the Ohio State University theory workshop, the 2012 Annual Conference of the Southern Economic Association, and the 2013 International Industrial Organization Conference for helpful discussions and comments.

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

[1]Examples include consumerreports.org, ratemds.com, vitals.com and drscore.com.

[2]"Zagat Gets Into Doctor Ratings," (http://blogs.wsj.com/health/2007/10/22/zagat-gets-into-doctor-ratings-business/, accessed July 23rd, 2013).

Of course physician competition for patients has not developed only as a consequence of patient reviews as there has always been some patients who seek second opinions or switch providers when they are sufficiently unhappy with their physician's treatment recommendations causing physicians to factor in their patients' preferences when proposing a treatment (Givens, 1957; McCarthy, 1985; Macpherson et al., 2001; Berry, 2007). Nonetheless, there is concern in the medical community that providing more information about physician treatment practices may exacerbate moral hazard by increasing the pressure to provide the treatments patients want. Exemplifying this concern, 55 per cent of the respondents to a recent South Carolina Medical Association survey reported ordering tests they felt were inappropriate because of pressure to avoid receiving a poor review from patients (Falkenberg, 2013). Moreover, these concerns are particularly salient now given that, pursuant to the Affordable Care Act of 2010, the Centers for Medicare and Medicaid Services (CMS) is preparing to make even more information available to prospective patients by publishing its own physician performance data that includes patient ratings and information on how well physicians' medical interventions succeed.

In light of these efforts to provide more information regarding physician treatment practices to patients, the objective of this study is two-fold: Explore how competitive pressure among physicians due to informed patients may impact physician treatment selection and identify how and when supply-side payment rules can overcome these incentives to induce physicians to follow an insurer's preferred treatment practice. These objectives are accomplished using a stylized model of treatment and physician selection which captures a couple critical features of the market. First, physician treatment selection is modeled as a choice among discrete treatments in which the cost and benefit of each treatment is dependent on the patients' illness severity or "type" so that it is more efficient to use one treatment over the other for each type while the more efficient treatment varies with the type.[3] For example, an intensive medical intervention, such as surgery, may be a more appropriate treatment choice for some men diagnosed with prostate cancer while a less intensive treatment option, such as active surveillance or radiation, is more appropriate for others.

Second, patients select physicians based on their treatment practices. That is, 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—reflecting the argument made by Mendel

[3]The discrete nature of treatment choice is in contrast to much of the literature examining optimal insurance and physician agency that frequently 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 Choné and Ma (2011).

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. Instead, patients prefer the treatment that leaves them with the highest expected benefit so select a physician whose treatment practice accomplishes this when informed about physician treatment practices. In consequence, physicians can gain market share by following treatment practices that more closely match patient preferences vis-à-vis other physicians. However, the amount of market share that a physician can capture is dependent on the proportion of patients who are informed. 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.

The principal finding of this study is that an insurer's ability to induce its preferences through supply-side payments critically depends on how informed patients are relative to the number of physicians competing for patients. Specifically, it is a simple matter for an insurer to induce its preferred treatment practice by making a physician internalize the social gain from the treatments when patients are completely uninformed and there is no competition via treatment choice. Accomplishing this when some patients are informed, however, requires that the physician earn relatively more when she treats marginal patients with their less preferred treatment in order to counteract the physicians' incentive to attract patients by following a practice that patients prefer. When a sufficient proportion of patients are informed (i.e., demand is sufficiently elastic to treatment practices) the insurer will be able to induce its preferred treatment practice in equilibrium. However, when demand is insufficiently elastic (but not perfectly inelastic) relative to the number of physicians in the market an insurer may not be able to simultaneously prevent physicians from increasing profit by gaining market share from utilizing the patients' preferred treatment more and prevent physicians from increasing profit by utilizing the more profitable but less preferred treatment.

The intuition behind these finding is as follows. A physician that chooses a treatment practice that is more preferred by patients compared to the other physicians' practices stands to gain a proportionally large market share as she takes a small proportion of informed patients from a large number of competing physicians. To counteract this incentive the insurer must make sure that the physician gives up a sufficiently large amount of profit on each patient she treats with the less preferred (by the insurer) treatment to overcome her increase in profit obtained by increasing market share. Accomplishing this, however, generates an incentive to instead treat more patients with the now more profitable (and less preferred by patients) treatment as the physician stands to lose only a small number of informed patients. In consequence, when too few patients are informed (but not all are uninformed) the insurer cannot prevent both deviations as they are mutually exclusive. As more patients become informed, however, the loss in patients from selecting a less desired treatment practice increases to the point where both deviations become unprofitable. As a result, if information on physician treatment practices is more

readily available, then an insurer will be able to better exercise control over physicians. The availability of diagnostic testing suffers from a similar problem if patients are also responsive to physician testing practices.However, diagnostic testing also complicates the insurer's problem as a physician's private gain from testing, which is increasing in the demand elasticity (proportion of informed patients), may be too large to overcome through the payment mechanism. In consequence, even when an insurer can induce its treatment practices, policies that increase patient information could also result in over-testing.

Although there is a large literature that has examined how insurers can exert control over a physician's treatment choice,[4] this paper shares features with Dranove (1988), Chernew, Encinosa and Hirth (2000), and Liu and Ma (2013) in particular. 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 (2013) also model the physician's treatment decision as a discrete choice between multiple treatment options.[5] Chernew et al. (2000) are primarily interested in the relationship between patients and the insurer while Liu and Ma (2013) 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 consider the effect of diagnostic testing on the insurer's payment rules. Moreover, none of these studies consider the impact of physician competition on a physician's treatment choice.

There are, however, a few related studies examining the effect of physician competition and insurance. In a seminal paper on optimal insurance, Ma and McGuire (1997) consider the effect of physician competition on effort 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 treatment quantity and effort are substitutes. Otherwise, when quantity and effort are

[4]See Gaynor (1994) and McGuire (2000) for detailed review of the literature on physician agency and physician induced demand (PID). 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.

[5]Chandra and Staiger (2007) also model the physician's treatment decision as a discrete choice between two treatment options. However, the authors' focus is on developing a Roy model of patient treatment choice with productivity spillovers to explain some of the regional variation in treatment intensities and not on the agency problem of inducing a particular treatment practice. They also do not consider the impact of competition.

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, 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 and increase physician effort. In contrast, in the current model competition aligns the physicians' incentives with the patients' preferences. In consequence, competition can prevent the insurer from inducing its preferred treatment practice, though the effect is not monotonic in the strength of competition as the insurer will not be able to exert control when the response or elasticity to physician treatment practices is too low (but not completely inelastic) for the number of physicians in the market. This result bears some resemblance to a result by Ellis (1998) in which providers over-provide 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-à-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 payment rules that allow an insurer to induce its preferences for both a monopoly physician and for the case of competing physicians. The effect of competition is analyzed in Section 4. Section 5 introduces diagnostic testing and examines how it limits an insurer to induce its preferences. Finally, Section 6 ends with some concluding remarks. All proofs are in Appendix A.

2. The 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 severity or 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.[6] Section 5 relaxes 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 must choose between two mutually exclusive 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 $\mathcal{C}^2$.[7] The cost of treatment reflects all of the physician's opportunity costs of providing the treatment. Furthermore, the increase in cost by illness severity 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.[8,9] Moreover, as the insurer does not observe the patients' illness severity the treatment is *ex ante* noncontractible, but *ex post* verifiable.[10]

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 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 overly 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 $H(\cdot \mid \theta, T_k)$;

[6]Instead of knowing their illness severity patients could receive a signal which is correlated with their true severity. The effect this has on the results is discussed further at the end of Section 3 and explored in Appendix C.

[7]The model is generalized to M treatments in Appendix D.

[8]Note that if physicians are altruistic, then the opportunity cost of one treatment is the foregone benefit of the other treatment in addition to the explicit costs of care.

[9]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 alternative 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.

[10]One may be concerned that if an insurer learns from the physician which treatment was selected, then the physician could have incentive to misreport. However, 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 will also hold here if we assume that the patient is responsible for some arbitrarily small coinsurance. To focus on how competitive pressure impacts treatment selection I simply assume the treatment is *ex post* verifiable.

i.e., a patient's benefit from treatment depends on his illness type and treatment. A patient's expected benefit from treatment takes the form:

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

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 is 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 follow its preferred treatment practice, which will differ from the patients' preferences. To ensure that the insurer and physician choice problems are well behaved I impose the following conditions for the benefit and cost of treatment:

A1: $d\{\psi(\theta, T_1) - \psi(\theta, T_2)\}/d\theta \leq 0$ for all $\theta \in \Theta$, and
A2: $d\{c_1(\theta) - c_2(\theta)\}/d\theta > 0$ for all $\theta \in \Theta$.

These are regularity conditions establishing that patients' preferences over treatments as well as the social value of the treatments are both monotonic over types, which guarantees that there is a unique social optimum.[11] Furthermore, let θ^P solve $\psi(\theta^P, T_1) = \psi(\theta^P, T_2)$ then A1 indicates that all types below θ^P will prefer treatment T_1 and all types above θ^P will prefer treatment T_2. A2 also guarantees that a physician's optimization problem will be well behaved and have a unique optimum.[12] Treatment T_1 could be thought of as the null or non-intensive treatment similar to Chandra and Staiger (2007), in which the physician monitors the patient but does not perform an explicit clinical intervention while T_2 is an intensive medical intervention. It is straightforward to extend the model to M treatments where again T_1 can represent a null treatment while the physician has available more than one additional medical intervention. This extension is performed in the Appendix.

[11] Each condition is essentially a discrete treatment analog to the assumption of concavity and convexity for the expected benefit and cost of treatment, respectively.

[12] To prove sufficiency in the insurer's second best optimization program (Proposition 3) it must be the case that $c_2''(\cdot)$ is not substantially larger than $c_1''(\cdot)$. As this is the only case in which the restriction is needed I do not impose it. See the proof for Proposition 3 for more details.

A couple remarks can be made regarding these assumptions. First, note that the results of the model would hold if instead of A2 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 (1998a), Jack (2005), Chandra and Staiger (2007) and Choné and Ma (2011) as A1 is sufficient to generate a unique optimum for the physician that also differs from the patients' preferred treatment practice.[13] Second, as patients are not exposed to the total cost of treatment some patients will prefer to be treated with more of one treatment than is socially optimal, but not necessarily all patients.

The patients, physician, and insurer play a simple game. The insurer chooses its payments 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 mechanism and the number of physicians in the market, physicians simultaneously choose (and commit to) their treatment practice given the distribution of illness severities and the proportion of informed patients (as described next) where a treatment practice prescribes the treatment that each type of patient will receive.[14]

A patient's choice of physician 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 (e.g., location, operating hours) 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.[15] Some proportion of patients may be informed about physician treatment practices, however. If so, then a physician who follows a practice that these patients prefer compared to the other physicians can expect to draw a higher market share by attracting these informed patients. It could also be the case that all patients are informed but only of the treatment practices for some subset of physicians in the market—perhaps by word of mouth from friends and family members—and can only choose from this subset.[16]

To capture the notion that not all patients are informed about all physician treatment practices, let $\phi(n)$, where n is the number of physicians in the market and $\phi : \mathbb{N} \to [0, 1]$, represent the proportion of the market that is informed about any particular physician's treatment style

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

[14] Commitment to one's treatment practice is important given the one-shot nature of the game. In practice informed patients select physicians based on the treatment selection histories of the physicians and this dynamic functions as a commitment device since any deviation today will affect demand tomorrow.

[15] 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.

[16] Note that this interpretation implies that patients are simply unaware of physicians for whom they have no information so they do not have a prior regarding their treatment practices nor do they have a prior for the number of physicians that are in the market.

when there are n physicians in the market. Note that both $\phi(\cdot)$ and n are exogenous to the model.[17] As the number of physicians increases, $\phi(n)$ may be constant in the case that ϕ simply represents the portion of patients that are informed or it may either increase or decrease with n as changing the number of physicians in the market alters the likelihood that patients learn about any specific physician via word of mouth (Satterthwaite, 1979; Pauly and Satterthwaite, 1981).[18] Patients who are not informed of any physician's treatment practice are randomly matched with a physician.

Observe that $\phi(n)$ affects the elasticity of demand with respect to treatment practice. For example, when no patients are informed ($\phi(n) = 0$), then demand is perfectly inelastic with respect to treatment practice and when all patients are informed ($\phi(n) = 1$) demand is perfectly elastic as even a small deviation towards patients' preferred practice will attract all patients. In this way, $\phi(n)$ alters the competitiveness of the market and can be used to analyze how the physician's treatment selection and insurer's reimbursements are affected by the informedness of the market.

The socially optimal outcome is 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 treat patients with T_1 when the patients' illness severity is below θ^{FB} and use T_2 when the illness severity is above θ^{FB}, where θ^{FB} solves*

$$\psi(\theta^{FB}, T_1) - U'(Y - P^{FB})c_1(\theta^{FB}) = \psi(\theta^{FB}, T_2) - U'(Y - P^{FB})c_2(\theta^{FB}). \tag{3}$$

If there is no $\theta \in \Theta$ solving (3), then $\theta^{FB} = \overline{\theta}$ when the left-hand-side of (3) is greater than the right-hand-side for all $\theta \in \Theta$ and $\theta^{FB} = \underline{\theta}$ when the right-hand-side of (3) is greater than the left-hand-side for all $\theta \in \Theta$. The first-best premium is

$$P^{FB} = \int_{\underline{\theta}}^{\theta^{FB}} c_1(\theta)dF(\theta) + \int_{\theta^{FB}}^{\overline{\theta}} c_2(\theta)dF(\theta).$$

The intuition behind the first-best outcome is straightforward. The cut-off θ^{FB}, 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. The social value of treatment is defined as the expected benefit of treatment net the social cost, which is the private cost of treatment

[17] This setup is similar to the model developed by Varian (1980) in which a certain exogenous proportion of consumers are informed about the distribution of prices in the market while the uninformed consumers randomly choose a store.

[18] See Rochaix (1989) for a model of consumer search in the market for physicians.

weighted by the patients' marginal utility of consumption. The monotonic difference in the expected benefits and costs of each treatment result in a single cut-off in which one treatment is optimal for all types below that cut-off and the other is optimal for all types above the cut-off. When the social value of treatment is higher for all types then only that treatment should be used. Lastly, because patients are risk-averse while physicians are risk-neutral, first-best requires that physicians do not earn positive economic profit and the premium simply covers the expected cost of treatment. Note also that first-best is independent of the number of physicians and the informedness of the market.

Moving forward, let $\theta^{FB} \in (\underline{\theta}, \overline{\theta})$ so that both treatments are optimal for some type allowing us to concentrate on the physician's decision to use one treatment over the other; and, let $c_1(\theta^{FB}) < c_2(\theta^{FB})$ so that more patients prefer to be treated with T_2 than is socially optimal.[19] 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 a choice of treatment cut-off and premium:

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

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

Figure 1 illustrates the characteristics of the model by identifying the socially efficient treatment cut-off θ^{FB} and the patients' preferred treatment cut-off θ^P. The figure includes both the patients' expected benefits and 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. An important characteristic of this model is that full insurance does not cause patients to demand inefficient levels of treatment (i.e., higher intensity for all types); but rather, it causes the marginal types between θ^P and θ^{FB} 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.

3. Inducing a Treatment Practice

Although the main objective is to identify how competition may impact the ability of an insurer to induce first-best, I am also more generally interested in identifying when an insurer can induce its *preferred* treatment practice (which may not be the social optimum) and how

[19] This latter assumption can be reversed, in which case patients will prefer T_1 over T_2 more than is socially optimal. The results will change accordingly, but the intuition behind the results remains the same.

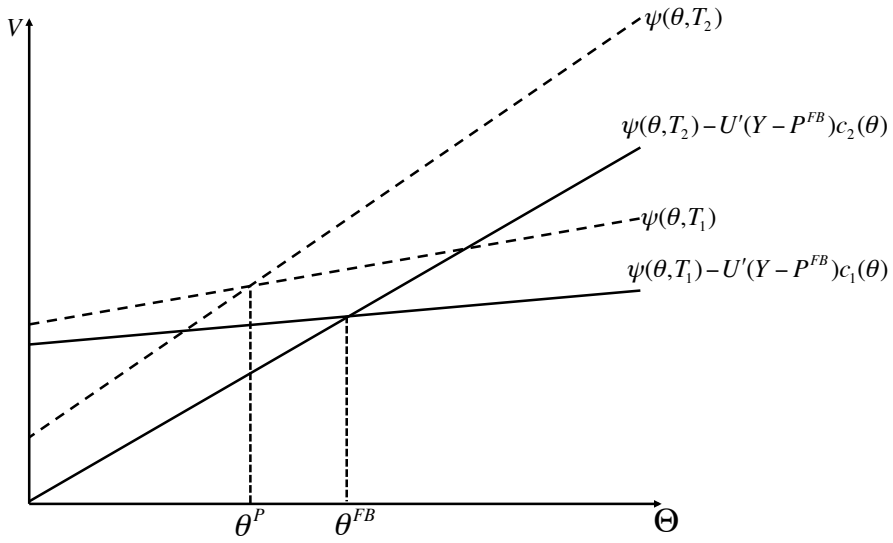


FIGURE 1. Illustration of the conflict between the preferences of the social planner and patients using linear cost and benefit functions. The dashed lines represent a patient's utility for each treatment and the solid lines represent the social value of each treatment. The private and socially optimal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

altering the competitiveness of the market impacts physician treatment choice and the insurer's ability to induce its preferred practice.[20] I begin by examining the problem when patients are uninformed making demand perfectly inelastic to treatment practices ($\phi(n) = 0$) and then build from that to examine how informing patients ($\phi(n) > 0$) impacts the physician and insurer problems.

3.1. *Uninformed Patients*

The insurer, acting as a Stackelberg leader, chooses the payment rule that induces the physician to optimally choose its preferred cut-off and treat all patients below the cut-off with T_1 and all patients above with T_2. Optimally an insurer would like to reveal the patient's illness severity through its payment mechanism and not have to pay more than the cost of treatment. However, without the ability to observe the physician's opportunity costs the only incentive compatible payment mechanism is a fixed reimbursement for the treatment performed.[21] To see this, suppose the insurer provides a menu of reimbursements for each treatment $\{r_1(\theta), r_2(\theta)\}_{\theta\in\Theta}$. Such a menu of payments is incentive compatible if and only if $r_k(\theta) - c_k(\theta) \geq r_k(\hat{\theta}) - c_k(\theta)$ and $r_k(\hat{\theta}) - c_k(\hat{\theta}) \geq r_k(\theta) - c_k(\hat{\theta})$ for all $\theta, \hat{\theta} \in \Theta$ and $k \in \{1, 2\}$. However, together these

[20]The profit-maximizing monopolist insurer's preferred treatment practice is derived in the Appendix.

[21]If the explicit costs of treatment are observable, but constant as in Liu and Ma (2013), then they would not reveal a patient's type anyway and the only incentive compatible payment would also be a fixed reimbursement.

conditions imply that $r_k(\theta) = r_k(\hat{\theta})$. Therefore, the insurer's payment mechanism simply consists of a reimbursement for each treatment similar to the prospective payment system used by Medicare.[22,23] As demand is perfectly inelastic to treatment practices, physicians are effectively monopolists who can choose the cut-off type that maximizes their profit without concern for how the choice impacts demand. When a physician is paid reimbursements r_1 and r_2 for treatments T_1 and T_2, respectively, her expected profit can be expressed as

$$\mathbb{E}\Pi = \int_{\underline{\theta}}^{\overline{\theta}} \left[I_{\{\tau=T_1\}}(\theta)\big(r_1 - c_1(\theta)\big) + \big(1 - I_{\{\tau=T_1\}}(\theta)\big)\big(r_2 - c_2(\theta)\big)\right] dF(\theta), \tag{4}$$

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 monotonic in the illness types. To see this, suppose there exists a $\hat{\theta} \in \Theta$ such that $r_1 - c_1(\hat{\theta}) = r_2 - c_2(\hat{\theta})$. Then for any type $\theta < \hat{\theta}$ it must be the case from A2 that profit is maximized using treatment T_1 while T_2 maximizes profit for any type $\theta > \hat{\theta}$. 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. The cut-off type $\hat{\theta}$ thus serves as the physician's choice variable and her optimization program can be expressed as

$$\max_{\hat{\theta}} \mathbb{E}\Pi(\hat{\theta}; r_1, r_2) = \max_{\hat{\theta}} \left\{ \int_{\underline{\theta}}^{\hat{\theta}} \{r_1 - c_1(\theta)\} dF(\theta) + \int_{\hat{\theta}}^{\overline{\theta}} \{r_2 - c_2(\theta)\} dF(\theta) \right\}.$$

Given the insurer sets reimbursements r_1 and r_2 it is clear from the argument above that choosing the reimbursements so that $r_1 - c_1(\theta^*) = r_2 - c_2(\theta^*)$ is necessary for inducing some θ^* as the treatment cut-off (this is also the first-order condition to the physician's optimization program) (sufficiency requires $\mathbb{E}\Pi \geq 0$). This condition indicates that the difference in payments (i.e., $\Delta r = r_2 - r_1 = c_2(\theta^*) - c_1(\theta^*)$) is what induces a cut-off and not the payment levels, *per se*. In consequence, any adjustment to Δr will alter a physician's treatment practice by causing her to alter her indifference type. Notably, this kind of change in treatment practice following a change in Δr has been documented by the economics and medical literature. For example, Gruber, Kim and Mayzlin (1999) observed that a reduction in the fee differential between natural and cesarean delivery by Medicaid resulted in fewer cesareans for Medicaid patients. Gruber et al. also found that the difference in the natural versus cesarean fee differential between privately insured and Medicaid patients explained the majority of the difference in

[22] 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 some particular diagnoses. The clearest example of this is the separate DRGs for natural and cesarean delivery.

[23] Type-dependent payments would be possible if physicians also provided a level of quality that impacted the benefit and cost of care *and* demand was responsive to the provided quality level (Chalkley and Malcomson, 1998b).

cesarean rates between the populations. Similarly, Weight, Klein and Jones (2008) found that chemical castration rates declined as physicians performed more surgical castrations (orchiectomy) for treatment of prostate cancer following a reduction in the reimbursement for chemical castration.

Whether the insurer wants to induce the social optimum or is trying to maximize profit it will want to leave the physician with zero rents in order to minimize costs, in which case the levels of r_1 and r_2 are also important. The following Proposition reports the unique payment rule that induces a monopolist physician to choose treatment cut-off θ^* while leaving the physician with zero economic profit.

Proposition 2. *Let $\{r_1^*, r_2^*\}$ be the insurer's reimbursements for treatments T_1 and T_2, respectively. The insurer can induce any $\theta^* \in \Theta$ and hold the physician to zero economic profit if and only if*

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

The payments that induce θ^* and leaves the physician with zero rents can be described as taking the expected cost of treating a patient and adjusting that by the difference in the cost of the treatments at the optimal cut-off times the probability that the patient should receive the alternative treatment. This adjustment process lowers r_1 below the expected cost of treatment proportional to the density of patients should be treated with T_2 and raises r_2 above the expected cost of treatment proportional to the density of patients that should be treated with T_1. Intuitively, the reimbursement for the treatment that will be used on the largest proportion of patients will be closer to the expected cost of care since the expected cost of care is more heavily weighted by that treatment.

Proposition 2 indicates that a simple payment rule is sufficient to induce some desired treatment practice while preventing the physician from earning rents. Of course that rule is dependent on the physician's opportunity cost of care, which in practice could vary considerably both across physicians and across regions.[24] Another potential limitation of the simple payment rule reported by Proposition 2 is that the payments must be sufficiently low that the physician incurs a loss for treating high types. This follows because the physician is held to zero economic profit requiring that she not earn rents on the highest type $\overline{\theta}$, otherwise she earns rents on all types.

[24]For example, Epstein and Nicholson (2009) find that within market variation in the rate of cesarean delivery is considerably larger than between market variation, likely reflecting large differences in physicians' opportunity costs for each treatment (which would include differences in their beliefs regarding the benefits of each treatment when they are partially altruistic).

As the lowest types are the most profitable they effectively subsidize the high types. Ensuring that the physician treats the high types (i.e., does not dump high cost patients) may be as straightforward as telling patients to report when a physician refuses to treat them and penalizing that physician. However, if physicians find other ways to exclude the high cost patients through wait lists or by skimping on services required by the high severity patients, then the insurer will be unable to induce its preferred treatment cut-off and leave the physician with zero economic profit. This problem can be avoided by charging the physician an "access fee" to have the right to treat the insurer's patients.[25] The reimbursements r_1 and r_2 can then be increased so that the physician does not incur an economic loss from treating any type (thus preventing dumping) while keeping Δr the same. The access fee can then be used to extract the excess rents.

If the inclusion of an access fee is not feasible and physicians cannot incur a loss on any type, then r_2 must be set to $c_2(\overline{\theta})$ while r_1 is set so that $r_1 - c_1(\theta^*) = c_2(\overline{\theta}) - c_2(\theta^*)$. As r_1 is increasing in θ^*, the higher r_2 will optimally be offset by choosing a lower cut-off θ^* so that r_1 does not increase as much as r_2. As a result, treatment T_2 will be used more compared to first-best. The following proposition identifies the socially optimal treatment cut-off when physicians' have an *ex post* nonnegative profit constraint to prevent dumping of high cost patients.

Proposition 3. *Let $\{r_1^*, r_2^*\}$ be the insurer's reimbursements for treatments T_1 and T_2, respectively. When physicians cannot incur a loss on any type $\theta \in \Theta$ the socially optimal treatment θ^* involves treating more types with T_2 compared to first-best; i.e., $\theta^* < \theta^{FB}$ where θ^* solves*

$$\begin{aligned}\big[\psi(\theta^*, T_1) - \psi(\theta^*, T_2) - U'(Y-P)\big(c_1(\theta^*) - c_2(\theta^*)\big)\big] f(\theta^*) \\ = U'(Y-P)\big[c_1'(\theta^*) - c_2'(\theta^*)\big] F(\theta^*).\end{aligned} \tag{5}$$

Eq. (5) has a simple interpretation. The left-hand-side represents the marginal change in surplus that results from increasing the cut-off θ^*. The marginal change is the difference in surplus at the cut-off multiplied by the density of patients, $f(\theta^*)$, that are effected by the change. Increasing the cut-off type requires increasing r_1 by $\big[c_1'(\theta^*) - c_2'(\theta^*)\big]$ since r_2 is fixed at $c_2(\overline{\theta})$ and the now higher r_1 must be paid for all types below θ^*. Thus, the right-hand-side represents the marginal social cost of increasing the cut-off type and the optimal treatment cut-off equates the benefit with the cost.

3.2. *Informed Patients*

Competition for patients 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

[25] Ma and McGuire (1997) encounter a similar limitation when quantity and effort are substitutes and there is competition between physicians.

cesarean delivery more or less frequently than patients prefer or think appropriate, then patients may substitute away from her.[26] 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 payments must be designed to prevent competitive deviations that do not 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 (including mixed strategy equilibria) necessarily results in an inefficient level of treatment. Furthermore, we can focus on the case in which the insurer wants to induce some $\theta^* > \theta^P$ since $\theta^{FB} > \theta^P$ and a profit maximizing insurer will not want to induce a $\theta^* < \theta^P$.[27]

To show that a payment rule induces some symmetric equilibrium we just need to consider the impact of unilateral deviations 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^*. In consequence, she will lose demand from informed marginal types between θ^* and $\hat{\theta}$. Although the physician loses demand, this deviation is profitable whenever the cost of using T_1 is substantially lower so 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 more preferred by marginal types 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 increases 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 higher overall 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

[26] 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`.

[27] Proposition 9, in the Appendix, shows that a profit maximizing, monopoly or monopolistically competitive insurer will choose a cut-off higher than θ^{FB}, thus even further from θ^P. In general, competing insurers will not try to induce a cut-off below θ^P as that is a treatment practice that is less preferred by patients and more costly. In consequence, an insurer can always increase profit by inducing at least θ^P.

the necessary and sufficient conditions for reimbursements that will prevent these profitable deviations at some $\theta^* \in (\theta^P, \overline{\theta})$.

Proposition 4. *Given reimbursements* $\{r_1, r_2\}$*,* $n \geq 2$ *physicians and market informedness* $\phi(n) \in [0,1)$*, a cut-off* $\theta^* \in (\theta^P, \overline{\theta})$*, can be induced in equilibrium if and only if* $\mathbb{E}_\theta[\Pi(\theta; r_1, r_2)] \geq 0$ *and the reimbursements satisfy:*

$$\text{(6a)} \qquad r_2 - c_2(\theta) \geq \big(1 - \phi(n)\big)\big(r_1 - c_1(\theta)\big) \qquad \forall \theta \geq \theta^*,$$

$$\text{(6b)} \qquad r_1 - c_1(\theta) \geq \big(1 + \phi(n)(n-1)\big)\big(r_2 - c_2(\theta)\big) \qquad \forall \theta \leq \theta^*.$$

Proposition 4 reports that the profit for each treatment is bounded from below. Specifically, if the profit from using treatment T_2 is too low relative to using T_1 for the cut-off type θ^* or higher, then the physician 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 lower bound, (6a), on profit using T_2 relative to T_1. On the other hand, if the profit from treatment T_2 is sufficiently high on lower types, then the physician can increase her demand by lowering her cut-off type and attracting an additional $\frac{1}{n}\phi(n)(n-1)$ of the marginal types. This deviation generates the lower bound, (6b), on profit using T_1 relative to T_2, which represents the proportional gain in demand when the physician deviates to a lower (more preferred by patients) cut-off.

Evaluating both (6a) and (6b) at θ^* shows that the conditions only provide bounds on the physician's relative profits at the desired cut-off, which could potentially be satisfied by multiple cut-offs given reimbursements $\{r_1, r_2\}$.[28] The equilibrium can be refined, however, by focussing on the cut-off type which is not Pareto dominated by another type within the interval. To illustrate, observe that when $n \geq 2$ and $\phi(n) > 0$ the lower bound is greater than one thus $r_1 - c_1(\theta^*) > r_2 - c_2(\theta^*)$ for any θ^* satisfying (6a) and (6b). Furthermore, when $r_1 - c_1(\theta) > r_2 - c_2(\theta)$ the physicians' profit increases when the cut-off is raised and more patients are treated with T_1. This implies that their profit is highest at the lowest bound and a single undominated equilibrium cut-off is pinned down by choosing r_1 and r_2 so that $r_1 - c_1(\theta^*) = \min\big\{\big(1 + \phi(n)(n-1)\big), (1 - \phi(n))^{-1}\big\} \cdot \big(r_2 - c_2(\theta^*)\big)$ and $\mathbb{E}_\theta[\Pi(\theta; r_1, r_2)] \geq 0$. This refinement is summarized in the following corollary.

Corollary 1. *Given reimbursements* $\{r_1, r_2\}$*,* $n \geq 2$ *physicians, and market informedness* $\phi(n) \in [0,1)$*, a cut-off* $\theta^* \in (\theta^P, \overline{\theta})$ *can be induced in equilibrium if* $r_1 - c_1(\theta^*) = \min\{\big(1 + \phi(n)(n-1)\big), (1 - \phi(n))^{-1}\} \cdot (r_2 - c_2(\theta^*))$ *and* $\mathbb{E}_\theta[\Pi(\theta; r_1, r_2)] \geq 0$*.*

Proposition 4 and Corollary 1 both apply to situations in which the market is not perfectly informed. In this case the reimbursements can be designed so that one treatment partially or

[28] Whenever $c_1'(\cdot)/c_2'(\cdot) \geq \max\{(1 - \phi(n))^{-1}, (1 - \phi(n)(n-1))\}$ it is sufficient to check that (6a) and (6b) are satisfied at θ^* only.

completely subsidizes the other; i.e.,

$$\int_{\underline{\theta}}^{\theta^*} \big(r_1 - c_1(\theta)\big) dF(\theta) \geq - \int_{\theta^*}^{\overline{\theta}} \big(r_2 - c_2(\theta)\big) dF(\theta) > 0.$$

However, this is not possible when the market is perfectly informed because a physician who incurs a loss from treatment T_2 can raise her cut-off to $\overline{\theta}$ and effectively dump all of the high-cost patients without actually refusing treatment. In consequence, the low-cost treatment cannot subsidize the high-cost treatment and physicians must either earn positive economic profit, or the insurer must use an access fee to extract the rents.[29] The following corollary to reports this special case.

Corollary 2. *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 $\theta^* \in (\theta^P, \overline{\theta})$ with some reimbursements r_1 and r_2 only if $\int_{\theta^*}^{\overline{\theta}} [r_2 - c_2(\theta)] d\theta \geq 0$.*

Before concluding this section, I return to the assumption that patients know their own type and its impact on the reimbursements. This assumption is important because it determines which patients are affected by and respond to physician treatment practices. This follows because when all physicians choose the cut-off θ^* and one deviates to a lower cut-off $\hat{\theta}$ only the marginal types between $\hat{\theta}$ and θ^* benefit by selecting that physician. In other words, only a small fraction of types are in play when a physician selects a different cut-off vis-à-vis the other physicians.

Instead of knowing their type, however, patients may only receive a signal (symptoms) that provide noisy information about their true illness severity and the noisiness of the signal will alter the set of patients that are marginal. For example a patient may receive a signal that indicates his type could be between $\hat{\theta}$ and θ^* making him a marginal patient even if his true type ultimately lies outside of this interval. Noisier signals increase the set of patients who are marginal until the signal is sufficiently noisy that it does not eliminate *any* of the illness severities—even if it indicates a patient is more or less likely to be severely ill—and all patients are marginal.

Following the proof for Proposition 4 it is straightforward to show that informedness of the market is irrelevant in this case and the the monopoly payment rule reported in Proposition 3 will induce any $\theta^* > \theta^P$ for all $\phi(n) \in [0, 1]$. The intuition is as follows. All informed patients will be attracted to a physician who deviates an arbitrarily small amount towards the patients' preferred cut-off since all types are marginal.[30] Physicians have an incentive to deviate

[29] If an access fee is not possible and physicians earn rents, then the findings of Proposition 3 are applicable and the socially optimal cut-off will be below the first-best cut-off.

[30] This is similar to how an arbitrarily small decrease in price will attract all consumers in a model of Bertrand competition for perfect substitutes.

closer to patients' preferred cut-off as long as the physician earns positive profit. To prevent this behavior, the reimbursements must be set so that physicians achieve their maximum profit at θ^* and the maximum profit is zero. The latter criteria is required so that a small loss in profit from deviating a little more cannot be overcome by the large increase in demand the deviation will create. This is exactly what the monopoly reimbursements reported in Proposition 2 accomplishes.[31] This case may better represent patient choice for a family physician where the choice of physician is typically made prior to falling ill and learning one's type. On the other hand, specific patients are more likely to be impacted by physician treatment practices when patients have some information about their illness type or their prognosis with various treatments such as would be the case with the chronically ill or those requiring the services of certain specialists and the reimbursements will need to be tailored to the degree of competition in the market.

4. The Impact of Competition Via Treatment Selection

Proposition 4 indicates that competition via treatment selection primarily affects the relative difference in the payments needed to induce physicians to choose some cut-off θ^*. When the market is completely uninformed and does not respond to physician treatment practices, inducing a particular practice requires payments that make the physician just indifferent at the desired cut-off. However, when more of the market becomes informed about physician treatment practices a physician must earn relatively more profit utilizing treatment T_1 at the desired cut-off to counter the incentive to increase demand by selecting a cut-off that is more preferred by patients generating condition (6a). On the other hand, the relative profit from using T_1 cannot be too high or the physician will instead have incentive to over treat with T_1 and this limit generates condition (6b).

Selecting reimbursements that fall within these bounds, however, is not always possible. When physicians must earn *ex post* profit (e.g., to avoid dumping) or the optimal payments are such that the physician receives nonnegative profit for either treatment at θ^*,[32] then the bounds provided by (6a) and (6b) require $1 + \phi(n)(n-1) \leq \big(1 - \phi(n)\big)^{-1}$. Otherwise there are no payments that can simultaneously prevent both an upward and downward deviation that is profitable. Both bounds are increasing in $\phi(n)$, but holding $\phi(n)$ fixed, only the lower bound is increasing in the number of physicians indicating that the lower bound may exceed the upper when the market is insufficiently informed for the number of physicians in the market. This follows because even a low demand elasticity can generate a significant increase in demand

[31] This case is formally analyzed in the Appendix.

[32] Observe that (6a) and (6a) imply $sign[r_1 - c_1(\theta^*)] = sign[r_2 - c_2(\theta^*)]$ so when it is profitable to use one treatment on type θ^* it's also profitable with the other. If physician profit must be nonnegative to prevent dumping then this condition will be satisfied; however, even without this restriction physician profit at θ^* may be nonnegative.

for the physician that selects a cut-off that is more preferred by patients when there are a large number of physicians while a physician that selects a less preferred cut-off would lose little demand.[33] In consequence, while the insurer must set the payments so that the profit differential between treating the marginal types with T_1 over T_2 is sufficiently large that the physician does not want to treat more patients with T_2, such a large difference will encourage the physician to treat more patients with T_1 than desired when the demand elasticity is sufficiently low. The following proposition reports the level of informedness that will prevent the insurer from inducing its preferred treatment cut-off.

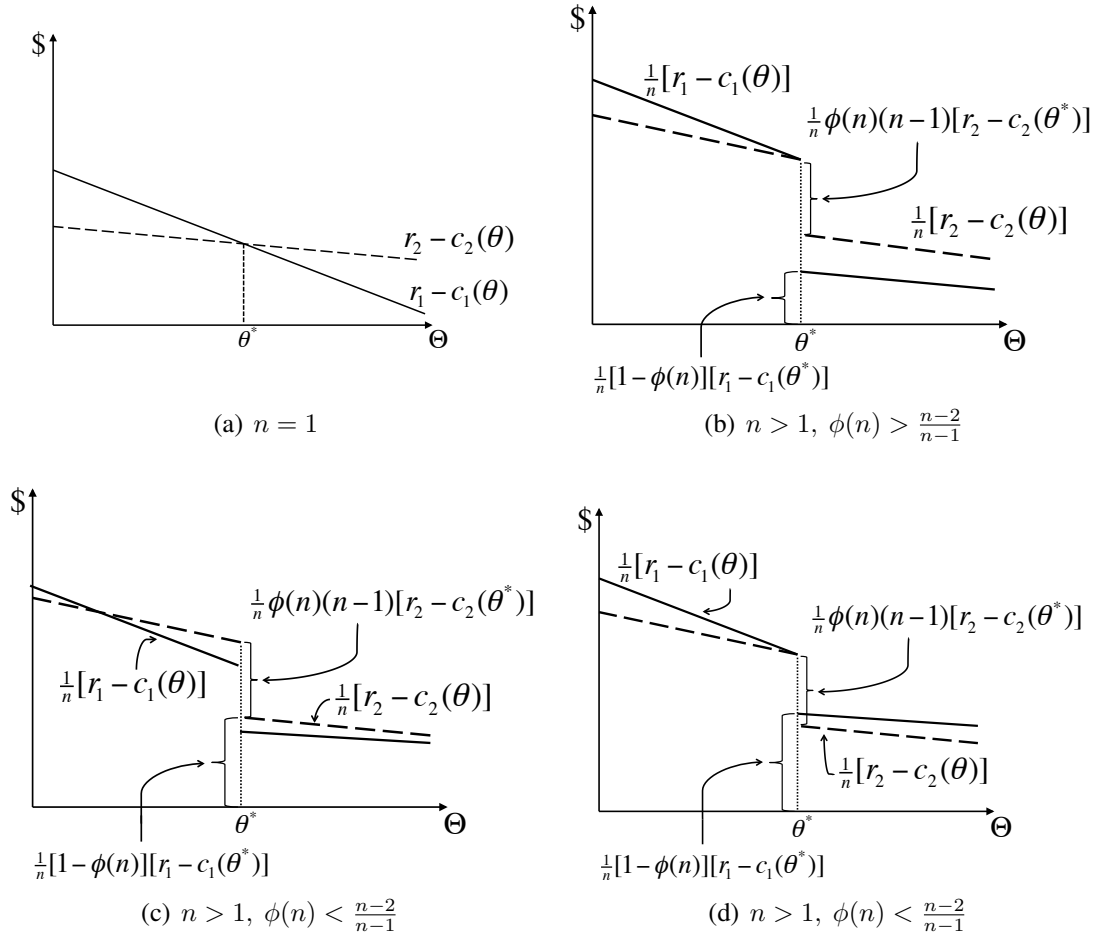
Proposition 5. *If* $r_i - c_i(\theta^*) \geq 0,\ i \in \{1,2\}$ *and* $0 < \phi(n) < \frac{n-2}{n-1}$*, then an insurer cannot induce* $\theta^* \in (\theta^P, \overline{\theta})$ *using only reimbursements* $\{r_1, r_2\}$.

Interestingly, Proposition 5 indicates that although an insurer needs demand to be sufficiently elastic to physician treatment practices given the number of physicians in the market, the insurer will also be able to induce its preferences when demand is perfectly inelastic to physician treatment practices (i.e., completely uninformed). The driver of this non monotonic effect of information is the asymmetric nature of the demand response following a change in treatment cut-off. When demand is perfectly inelastic, the reimbursements need only be set so that the physician's profit is the same for each treatment at the desired cut-off. However, even if the proportion of the market that is informed is arbitrarily small, but positive, the physician will gain $n-1$ more patients from choosing a more preferred cut-off than she will lose by choosing a less preferred cut-off. A higher number of physicians increases this asymmetry, requiring demand to be more elastic if the penalty incurred from utilizing T_1 more and losing patients is to be sufficient to prevent the physician from making such a deviation.[34]

To provide further intuition, Figure 2 uses linear cost functions to illustrate the insurer's problem under different scenarios. Panels 2(a) and 2(b) present a physician's profits when the payments induce θ^* in equilibrium. In both cases the payments are set so that the physician is indifferent at θ^*. Panel 2(a) presents the monopoly case in which the physician earns the same economic profit at θ^* for each treatment and a change in treatment cut-off does not alter her demand. Panel 2(b) illustrates the discontinuous jump in profit that occurs when a physician chooses a cut-off that differs from the other physicians. For example, if the physician chooses to use T_2 on types below θ^*, then her demand will increase by $\frac{1}{n}\phi(n)(n-1)$; however, her profit

[33] 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.

[34] For example, when there are only two physicians in the market the gain and loss in demand from deviating to a different cut-off are equivalent and the insurer will be able to induce its preferred cut-off for any $\phi(n) \in [0,1]$ while an insurer will need demand in a market of ten physicians to be much more elastic with $\phi \geq 8/9$ if it is to induce its preferences.



(a) $n = 1$

(b) $n > 1,\ \phi(n) > \frac{n-2}{n-1}$

(c) $n > 1,\ \phi(n) < \frac{n-2}{n-1}$

(d) $n > 1,\ \phi(n) < \frac{n-2}{n-1}$

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 selecting treatment cut-off θ^*.

from using T_1 is still higher on these lower types so she has no incentive to deviate. Similarly, her profit is higher using T_2 on types above θ^* as the loss in demand from using T_1 results in a large loss in profit.

Panels 2(c) and 2(d) illustrate the insurer's problem when $0 < \phi(n) < \frac{n-2}{n-1}$. In panel 2(c) the physician has incentive to deviate to a $\hat{\theta} < \theta^*$ and increase market share. The insurer could increase r_1 to counter this incentive, but as panel 2(d) illustrates, this creates an incentive to instead choose a less preferred cut-off as the loss in demand is insufficient to counter the increase in profit from using T_1 on more types.

Having established how physicians respond to the insurer's payments and the degree of competition in the market, we can now analyze how changes in the payments and informedness of

the market alter the physician treatment decision and the insurer's payments. These comparative statics are reported in the following proposition.

Proposition 6. *(Comparative Statics) Given reimbursements $\{r_1, r_2\}$, the Pareto dominant equilibrium cut-off θ^* satisfies the following comparative statics:*

When $r_i - c_i(\theta^*) \geq 0$ $i \in \{1, 2\}$, then:

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

Otherwise:

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

The optimal payment rule satisfies the following comparative statics:

When $r_i - c_i(\theta^*) \geq 0$ $i \in \{1, 2\}$, then:

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

Otherwise:

$$\frac{dr_2^*}{d\phi(n)} > 0 > \frac{dr_1^*}{d\phi(n)} \text{ and } \frac{dr_2^*}{dn} \lesseqgtr 0 \lesseqgtr \frac{dr_1^*}{dn} \text{ as } \phi'(n) \lesseqgtr 0.$$

The comparative statics are divided into two sets and each set is conditional on whether treating type θ^* is profitable. The comparative statics for $d\theta^*/d\phi(n)$ and $d\theta^*/dn$ identify how the equilibrium cut-off changes following an increase in the market informedness or the number of physicians while the reimbursements remained fixed.[35] The intuition for $d\theta^*/d\phi(n) < 0$ follows from the fact that as the market becomes more informed and treating a patient with illness severity θ^* is profitable, there is a larger incentive to deviate towards patients' preferred treatment cut-off to further increase profit. Similarly, increasing the number of physicians in the market decreases the physicians' market shares, amplifying the incentive to increase market share by deviating towards patients' preferred treatment cut-off. In the data this would resemble physician induced demand since the more intensive treatment is used following an increase in the number of physicians. However, it's important to observe that the driver of the increased utilization are patient preferences contrary to PID. Moreover, increasing the number of physicians could decrease the probability that a patient is knowledgeable about the treatment characteristics of any one physician; consequently, the sign for $d\theta^*/dn$ depends on which effect dominates. When physicians incur a loss from treating θ^* they have an incentive to treat fewer high severity patients since they are unprofitable and accomplishes this by selecting a less attractive treatment plan.

[35] This latter circumstance has been used in the literature to identify the presence of physician induced demand (e.g. Fuchs, 1978; Auster and Oaxaca, 1981; Pauly and Satterthwaite, 1981).

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 there is an increase in either the market informedness or the number of physicians. When $r_i - c_i(\theta^*) \geq 0$ and $\phi(n)$ increases, the incentive to deviate towards patients' more preferred treatment practice also increases; therefore, the payments must be adjusted so that it is relatively more profitable to treat the marginal types with T_1 to counteract this incentive. 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 knowledge of any one physician's treatment practice. In contrast, when physicians incur a loss from treating θ^* they must be incentivized more to use the patients' preferred treatment and avoid implicit dumping following an increase in $\phi(n)$.

A couple additional remarks can be made regarding the comparative statics. First, that the same payments will generate different treatment practices in markets having different demand elasticities may provide an additional explanation for some of the unexplained regional variation in treatment practices (Phelps, 2000; Skinner, 2012). Second, variation in physician opportunity costs can also cause physicians to follow different treatment practices. To some degree Medicare accounts for this by applying what it calls a geographical practice cost index (GPCI) adjuster to physician reimbursements. However, this adjuster is applied to the entire menu of reimbursements as it only accounts for variations in the costs of practicing medicine and not for differences in the competitiveness of a market. A more effective adjustment will need to account for cost differences across markets, differences in the relative opportunity costs of each treatments,[36] as well as differences in the competitiveness of the markets where total competitiveness depends on the number of physicians and the informedness of patients.

5. DIAGNOSTIC TESTING

Diagnostic testing is an important part of the treatment process as physicians frequently may not know with certainty which treatment is optimal for a particular patient. Moreover, there is concern that diagnostic testing could be 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 and to what extent testing can be controlled via the payment mechanism. There are many illnesses in which one or more diagnostic tests are always a necessary part of treatment such as an x-ray to identify the nature of a broken bone and how best to reset it. Testing in such situations is simply part of the cost of treatment, however, so is adequately captured by the no-diagnostic testing regime. The kind of testing this section is concerned with is that which is not always medically necessary but may be used to help a physician determine the optimal treatment.

[36]The relative opportunity costs could differ between treatments across markets because of differences in tort laws, for example.

To identify how this type of diagnostic testing alters the payment rule and ultimately how competition impacts the amount of testing performed, the following modification is made to the model. Instead of observing a patient's illness type θ_i directly, the physician and patient observe a signal (i.e., symptoms) $\xi_i \in [\underline{\xi}, \overline{\xi}]$ for patient i's illness where the signals are independently drawn from cumulative distribution Γ and 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 $\xi_1 > \xi_2$ so that higher signals indicate that the patient is likely to be more severely ill. The function $\phi(n)$ continues to represent the proportion of the market that is informed about a particular physician's treatment and testing style.

Physicians have available a diagnostic test which reveals a patient's type at a cost $D \geq 0$.[37] The value or benefit of the test comes from the fact that by revealing a patient's type, the socially optimal treatment can be chosen. However, this benefit may not exceed the cost of the test for all signals, in which case testing is not always socially optimal.[38] The following proposition establishes the first-best testing and treatment practice given this trade-off.

Proposition 7. *The first-best diagnostic testing regime is as follows. A physician does not conduct the diagnostic test and treats patients with T_1 when $\xi < \xi_L^{FB}$ and T_2 when $\xi > \xi_U^{FB}$; and conducts the diagnostic test whenever $\xi \in [\xi_L^{FB}, \xi_U^{FB}]$ and follows the treatment practice reported in Proposition 1 based on the realized value of θ where ξ_L^{FB} and ξ_U^{FB} solve, respectively:*

$$\int_{\theta^{FB}}^{\overline{\theta}} \left[\big(\psi(\theta, T_2) - \psi(\theta, T_1)\big) - U'(Y - P^{FB})\big(c_2(\theta) - c_1(\theta)\big)\right] dG(\theta \mid \xi_L^{FB}) = U'(Y - P^{FB})D \tag{7}$$

$$\int_{\underline{\theta}}^{\theta^{FB}} \left[\big(\psi(\theta, T_1) - \psi(\theta, T_2)\big) - U'(Y - P^{FB})\big(c_1(\theta) - c_2(\theta)\big)\right] dG(\theta \mid \xi_U^{FB}) = U'(Y - P^{FB})D \tag{8}$$

and

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

If there is no ξ_L^{FB} solving (7), then $\xi_L^{FB} = \underline{\xi}$; and, if there is no ξ_U^{FB} solving (8), then $\xi_U^{FB} = \overline{\xi}$.

Proposition 7 identifies two diagnostic cut-offs. The lower cut-off, ξ_L^{FB} is the signal, below which the social cost of the diagnostic test exceeds the benefit, while the upper cut-off, ξ_U^{FB} is the signal, above which the social cost of the diagnostic test exceeds the benefit. In both

[37] The cost to the physician 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.

[38] Indeed, the cost of the diagnostic test could be so high that it is never efficient to perform the test. We will ignore this uninteresting case and assume that the cost of the test is such that it is efficient to conduct the test for at least some signals of illness severity.

cases the social cost is the cost of the test weighted by the patient's marginal utility on other expenditures while the benefit is the expected difference in total social value for selecting the correct treatment. The proposition informs us that the value of performing a diagnostic test is not monotonic in the treatment severity since the value is decreasing in both tails of the signal distribution.

As the diagnostic test allows a physician to choose the optimal treatment regardless of how extreme the signal is, patients will always prefer more diagnostic testing to less when they are not exposed to the cost of the test. In consequence, competing physicians gain from additional diagnostic testing by taking market share from physicians who test less—even if they incur a cost for performing the test. Furthermore, since a given signal does not specify the exact illness severity a larger set of patients will be sensitive to differences in physician treatment practices, which creates a larger incentive to follow a treatment practice that is more preferred by patients. The reimbursements must be designed, of course, so that they overcome both the incentive to over test and the now stronger incentive to follow a treatment practice that is preferred by patients.

As the insurer wants to induce multiple testing cut-offs in addition to the single treatment cut-off, it will need to utilize additional payment instruments. Specifically, the insurer can offer reimbursements $\hat{r}_1$ and $\hat{r}_2$ for treatments T_1 and T_2, respectively, conditional on the physician not conducting the diagnostic test and r_1 and r_2 conditional on the physician conducting the diagnostic test.[39] Additionally, it may be necessary to pay the physician a fixed payment $R > 0$ so that her participation constraint is satisfied. With these policy instruments, inducing some diagnostic cut-offs ξ_L^* and ξ_U^* in equilibrium follows a similar intuition to that of inducing the treatment cut-off: A physician that conducts the diagnostic test for more extreme signals compared to the other physicians will take some market share and a physician that only conducts the diagnostic test for less extreme signals compared to the other physicians will lose market share. Proposition 8 provides the necessary and sufficient conditions required to simultaneously induce a treatment practice and diagnostic regime.

Proposition 8. *Given reimbursements* $\{r_1, r_2, \hat{r}_1, \hat{r}_2\}$*, fixed payment* $R \geq 0$*, and market informedness* $\phi(n) \in [0, 1)$*, the diagnostic cut-offs* $\xi_L^*, \xi_U^* \in [\underline{\xi}, \overline{\xi}]$ *and treatment cut-off* $\theta^* > \theta^P$ *represent an equilibrium if and only if the payments satisfy* $E_{\theta,\xi}[\Pi(\theta^*, \xi_L^*, \xi_U^*; r_1, r_2, \hat{r}_1, \hat{r}_2, R)] \geq$

[39] In practice the insurer can reimburse $\hat{r}_1$ and $\hat{r}_2$ for the treatments and provide additional reimbursements $r_1 - \hat{r_1}$ and $r_2 - \hat{r_2}$ for treatments T_1 and T_2 when the diagnostic test is conducted.

0, $r_1 \geq \hat{r}_1$, $r_2 \geq \hat{r}_2$, *and the conditions*

$$r_1 - c_1(\theta^*) = r_2 - c_2(\theta^*), \tag{9}$$

$$E_\theta[\Pi(\theta^*) \mid \xi_L^* \leq \xi \leq \xi_U^*] = D, \tag{10}$$

$$1 + \phi(n)(n-1) \leq \frac{\mathbb{E}_\theta\big[\Pi(\overline{\theta}) \mid \xi_L^*\big]}{\mathbb{E}_\theta[\Pi(\theta^*) \mid \xi_L^*] - D} \leq \big(1 - \phi(n)\big)^{-1}, \tag{11}$$

$$\big(1 - \phi(n)\big)^{-1} \leq \frac{\mathbb{E}_\theta\big[\Pi(\underline{\theta}) \mid \xi_U^*\big]}{\mathbb{E}_\theta[\Pi(\theta^*) \mid \xi_U^*] - D} \leq 1 + \phi(n)(n-1) \tag{12}$$

where $\mathbb{E}_\theta[\Pi(\theta) \mid \xi] = \int_{\underline{\theta}}^{\theta} \big(r_1 - c_1(\tilde{\theta})\big) dF(\tilde{\theta} \mid \xi) + \int_{\theta}^{\overline{\theta}} \big(r_2 - c_2(\tilde{\theta})\big) dF(\tilde{\theta} \mid \xi)$.

Proposition 8 shows that the insurer's problem is not only much more complicated in the presence of diagnostic testing, but that if patients are responsive to testing practices, then an insurer will not be able to simultaneously induce both upper and lower diagnostic cut-offs. To understand these constraints, first consider what happens when the signal falls within the range in which the physicians perform the diagnostic test. For example, suppose all of the physicians perform the test when presented the same set of signals $[\xi_L^*, \xi_U^*]$. In this case, a physician can attract an additional $\frac{1}{n}\phi(n)(n-1)[\Gamma(\xi_U^*) - \Gamma(\xi_L^*)]$ patients by selecting a treatment cut-off that is more preferred. This follows because any patient receiving such a signal will not know their true type *ex ante* so could benefit by selecting the physician with the more preferred treatment practice. Physicians have incentive then to select a treatment cut-off that is closer to patients' preferred cut-off compared to the other physicians chosen cut-off as long as it is profitable to do so. Accordingly, the reimbursements must be set so that the physician achieves her maximum profit at the desired θ^*—yielding condition (9). Moreover, the maximum expected profit must be zero so that a small loss in profit cannot be overcome by the large increase in demand that results from selecting a more preferred cut-off—yielding (10).

Conditions (11) and (12) provide bounds for the reimbursements conditional on not conducting the test in a similar manner to conditions (6a) and (6b) of the no-testing case. The numerator of the middle term in (11) represents the physician's profit from not administering the test and always choosing T_1 when the diagnostic signal is ξ_L^* while the denominator represents the expected profit from administering the test at the same signal ξ_L^*. Condition (12) 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_U^*. Competition creates a wedge between the diagnostic and no-diagnostic profits at the desired cut-off similar to the wedge created at the treatment cut-off since physicians can gain market share or effectively dump costly patients vis-à-vis the diagnostic practices of other physicians. The bounds of the two diagnostic tests are generally mutually exclusive, however, because lowering the lower cut-off and raising the upper cut-off both *increase* demand while

increasing the lower cut-off and lowering the upper cut-off *decrease* it. Only when patients do not respond to testing differences ($\phi(n) = 0$) can both conditions be satisfied simultaneously.

Proposition 8 also reports that the reimbursements must satisfy $r_1 > \hat{r}_1$ and $r_2 > \hat{r}_2$. This constraint follows because, although an insurer can differentially reimburse the physician for the treatment chosen based on whether or not she also conducted a diagnostic test, in practice a physician can choose not to reveal she has administered the test if withholding that information is beneficial to the physician.[40] 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.[41] This limits an insurer when the private benefit from revealing the patient's type is sufficiently large that it overcomes the cost of conducting the test. In this case the insurer will not be able to induce one or both of the desired treatment cut-offs. In consequence, the test must be sufficiently costly if the insurer is to have control over the physicians' testing practice. The following corollary to Proposition 8 reports the cost necessary to induce a testing regime.

Corollary 3. *If there is no cost to patients from the diagnostic test, then the insurer can induce treatment cut-off $\theta^* > \theta^P$ and diagnostic cut-offs $\xi_L^*, \xi_U^* \in [\underline{\xi}, \overline{\xi}]$ with treatment reimbursements $\{r_1, r_2, \hat{r}_1, \hat{r}_2\}$ where $r_1 \geq \hat{r}_1$, $r_2 \geq \hat{r}_2$ only if*

$$D \geq \mathbb{E}_\theta\big[\Pi(\theta^*) \mid \xi_L^*\big] - \frac{\mathbb{E}_\theta\big[\Pi(\overline{\theta}) \mid \xi_L^*\big]}{\big(1 + \phi(n)(n-1)\big)}, \text{ and} \tag{13}$$

$$D \geq \mathbb{E}_\theta\big[\Pi(\theta^*) \mid \xi_U^*\big] - \frac{\mathbb{E}_\theta\big[\Pi(\underline{\theta}) \mid \xi_U^*\big]}{\big(1 + \phi(n)(n-1)\big)}. \tag{14}$$

Both (13) and (14) follow directly from conditions (11) and (12) in Proposition 8. When a physician chooses to conduct the test on more types compared to the other physicians she gains in two ways. First, the physician gains by increasing her market share at the expense of the other physicians similar to when she selects a treatment practice that is more aligned with patient preferences; and second, the physician gains from being able to choose the profit maximizing treatment for the patients' illness type. The physician's private cost of testing must be sufficiently large to overcome these two gains if the insurer is to induce its preferred testing regime. Because the benefit to the physician from testing more is increasing in the proportion of the market that responds to physician testing differences, programs that increase the transparency of physician testing differences could generate higher levels of diagnostic testing.

[40] This is particularly true when patients do not incur a cost from the test since they will always prefer more testing over less.

[41] Ma and Riordan (2002) allow an insurer to pay a physician not to treat patients. This is similar to the case in the current model in which T_1 represents a null treatment since for some types the physician's economic profit will be higher when she chooses to not treat the patient. The difference with diagnostic testing is that by learning a patient's type, the physician is able to choose the treatment that maximizes her profit, even if the null treatment is revealed to be optimal.

Lastly, if patients instead incur the cost from diagnostic testing, which would 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 the cost $D > 0$ from the diagnostic test, then at lower diagnostic cut-off ξ_L^* patients will prefer to be tested less whenever $\mathbb{E}_\theta[\psi(\theta \mid \bar{\theta}) \mid \xi_L^*] > \mathbb{E}_\theta[\psi(\theta \mid \theta^*) \mid \xi_L^*] - D$.[42] In this case, a physician gains market share by increasing her 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 D is sufficiently large that $\mathbb{E}_\theta[\psi(\theta \mid \underline{\theta}) \mid \xi_U^*] > \mathbb{E}_\theta[\psi(\theta \mid \theta^*) \mid \xi_U^*] - D$, then the insurer must compensate the physicians for performing the diagnostic test on high types in order to overcome the competitive incentive to reduce testing. However, in either case it is simple matter for the insurer to compensate a physician more for performing the test.

6. Conclusions

This paper has examined how competition for patients via treatment selection impacts physician treatment choice and the limits of what an insurer can do to induce its preferred treatment practice through supply-side payments. When patients do not respond to differences in physician treatment practices, inducing a specific practice simply requires payments that leave the physician indifferent at the desired cut-off. As patients become more informed about the differences in physician treatment practices increasing the elasticity of demand, however, physicians must earn relatively more profit utilizing the less costly treatment at the desired cut-off to counter the incentive to increase demand. Although the insurer must increase the physician's relative profit from utilizing the lower cost treatment, the additional profit cannot be too great or the physician will have incentive to over utilize the less costly treatment. Balancing the incentive to increase demand with the incentive to use the relatively more profitable treatment (that is less preferred by patients) requires that patients be sufficiently informed, otherwise supply-side payments will be insufficient to induce physicians to follow a specific treatment practice. These results indicate that although competitive pressures will further align physician incentives with patient preferences, increasing those pressures may ultimately benefit the insurer by generating conditions that allow it to exert control over the physician.

Inducing an insurer's preferred diagnostic testing regime faces similar limitations when patients are responsive to the differences in physician testing practices. However, because the incentives work in opposite directions, an insurer will not be able to induce a testing regime that has both a lower and upper cut-off. Moreover, physicians benefit from testing, even absent any change in demand, since revealing a patient's true type allows it to select the profit maximizing treatment. That patients may select physicians based on their testing practices only increases

[42] Where $\mathbb{E}_\theta[\psi(\theta \mid \theta^*) \mid \xi] = \int_{\underline{\theta}}^{\theta^*} \psi(\theta, T_1) dG(\theta \mid \xi) \int_{\theta^*}^{\overline{\theta}} \psi(\theta, T_2) dG(\theta \mid \xi)$.

the physicians' benefit. As a result, factors that increase the responsiveness of patients to physician testing differences can generate excessive testing that cannot be controlled via supply-side payments if the physician's private cost is too low.

When the insurer is unable to induce its preferences through supply-side payments additional instruments such as utilization review, which removes the agency problem altogether, will be required. Because utilization review is both expensive and commonly used by private insurers, it would be particularly worthwhile to analyze how the practice can be optimally utilized and the relationship between utilization review and competition. For example, insurers use networks and the threat of exclusion to increase their bargaining power vis-à-vis providers—particularly hospitals—but increasing the elasticity to physician treatment practices has the potential to shift the balance of bargaining power in the physicians' direction as physicians that select a treatment practice that are more preferred by patients will be more valuable as well.

NOTES

[41] Strictly speaking I also assume that the insurer has all of the bargaining power such that the physician must accept whatever payments the insurer offers subject to the physician's participation constraint. 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.

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.

_____ **and** _____, "Contracting for Health Services with Unmonitored Quality," *Economic Journal*, 1998, *108*, 1093–1110.

Chandra, Amitabh and Douglas O. Staiger, "Productivity Spillovers in Health Care: Evidence from the Treatment of Heart Attacks," *Journal of Political Economy*, 2007, *115* (1), 103–140.

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," *Annales and d' Économie et de Statistique*, 2011, *101-102*, 229–256.

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.

Epstein, Andrew J. and Sean Nicholson, "The formation and evolution of physician treatment styles: An application to cesarean sections," *Journal of Health Economics*, 2009, *28*, 1126–1140.

Falkenberg, Kai, "Why Rating Your Doctor Is Bad For Your Health," in "Forbes" 2013.

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.

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," *Journal of Economic Behavior & Organization*, 2013, *85*, 79–96.

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. 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.

Skinner, Jonathan, "Causes and Consequences of Regional Variations in Health Care," in Mark V. Pauly, Thomas McGuire, and Pedro B. Barros, eds., *Handbook of Health Economics*, Vol. 2 of *Handbook of Health Economics*, Elsevier, 2012, chapter 5, pp. 45–94.

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.

Varian, Hal, "A Model of Sales," *American Economic Review*, 1980, *70* (4), 651–659.

Weight, C.J., E.A. Klein, and J.S. Jones, "Androgen deprivation falls as orchiectomy rates rise after changes in reimbursement in the U.S. Medicare population," *Cancer*, 2008, *112* (10), 2135–2201.

APPENDIX A. MATHEMATICAL PROOFS

Proposition 1: 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 $k \in \{1, 2\}$. 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 $k \in \{1, 2\}$.

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

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

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

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

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

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).

By A1 – A2, when there exists some θ^{FB} such that $\psi(\theta^{FB}, T_1) - U'(Y-P)c_1(\theta^{FB}) = \psi(\theta^{FB}, T_2) - U'(Y-P)c_2(\theta^{FB})$ then surplus is higher using T_2 for types above θ^{FB} and T_1 for types below θ^{FB}. Where there does not exist such a θ^{FB} then one treatment will yield higher surplus for all types. If T_1 yields the highest surplus then $\theta^{FB} = \overline{\theta}$ and if T_2 yields the highest surplus then $\theta^{FB} = \underline{\theta}$. A3 ensures that one of the treatments always yields positive surplus.

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

Proposition 2: Whether an insurer is trying to maximize social surplus or individual profit it will have to pick treatment reimbursements $\{r_1, r_2\}$ that satisfy the physician's participation constraint at the desired cut-off θ^*:

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

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

Given reimbursements $\{r_1, r_2\}$ the physician's optimization program is defined as

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

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

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

Eq. (A-5) represents a necessary condition and pins down r_1 relative to r_2. By expressing r_1 as a function of r_2, the physician's participation constraint can be expressed as

$$\int_{\underline{\theta}}^{\theta^*} [r_2 + c_1(\theta^*) - c_2(\theta^*) - c_1(\theta)]dF(\theta) + \int_{\theta^*}^{\overline{\theta}} [r_2 - c_2(\theta)]dF(\theta) \geq 0. \tag{A-6}$$

Of course whether the insurer is trying to maximize social surplus or own profit it will want to hold the physician to zero economic profit so that (A-6) binds. Eq. (A-6), can then be used to pin down r_2 relative to θ^* resulting in the payments identified in the proposition.

Lastly we need to verify that $r_1 \geq 0$ and $r_2 \geq 0$. First note that $r_2 > r_1$ so that it is sufficient to check that $r_1 \geq 0$. Observe that

$$\begin{aligned}
\mathbb{E}_\theta\big[c(\theta \mid \theta^*)\big] &= \int_{\underline{\theta}}^{\theta^*} c_1(\theta)dF(\theta) + \int_{\theta^*}^{\overline{\theta}} c_2(\theta)dF(\theta) \\
&> \big[1 - F(\theta^*)\big]c_2(\theta^*) \\
&> \big[1 - F(\theta^*)\big]\big[c_2(\theta^*) - c_1(\theta^*)\big].
\end{aligned}$$

Therefore $r_1 > 0$ for all $\theta^* \in \Theta$.

Proposition 3: If the physician cannot incur a loss from treating any type then the insurer's payments are limited by the cost of treating the highest type $c_2(\overline{\theta})$. To minimize the total cost it must be the case that the physician earns zero economic profit at $\overline{\theta}$ and $r_2 = c_2(\overline{\theta})$. With r_2 fixed r_1 must be set so that $r_1 - c_1(\theta^*) = r_2 - c_2(\theta^*)$ therefore

$$r_1 = c_2(\overline{\theta}) - \big(c_2(\theta^*) - c_1(\theta^*)\big). \tag{A-7}$$

Using r_1 and r_2 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

$$\begin{aligned}
P &= \int_{\underline{\theta}}^{\theta^*} r_1 dF(\theta) + \int_{\theta^*}^{\overline{\theta}} r_2 dF(\theta) \\
&= \big[c_2(\overline{\theta}) - \big(c_2(\theta^*) - c_1(\theta^*)\big)\big]F(\theta^*) + \big[1 - F(\theta^*)\big]c_2(\overline{\theta}).
\end{aligned}$$

Taking the first-order conditions yields

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

As $F(\cdot)$ satisfies the monotone hazard rate property the insurer's problem is concave in θ and the θ^* as long as $c_2''(\cdot)$ is not significantly greater than $c_1''(\cdot)$. This follows because $\frac{d}{d\theta}\big(\psi(\theta, T_1) - \psi(\theta, T_2) - U'(Y-P)[c_1(\theta) - c_2(\theta)]\big) \leq 0$ and $\frac{d}{d\theta}\left(\frac{F(\theta)}{f(\theta)}\right)\big[c_2'(\theta) - c_1'(\theta)\big] \leq 0$.

Lastly, it must be the case that $\theta^* < \theta^{FB}$ since the insurer's problem is concave in θ and the first-order condition, (A-8), evaluated at θ^{FB} is positive.

Proposition 4 : 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 $\theta^* > \theta^P$ 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

$$\begin{aligned}\mathbb{E}_\theta[\Pi_i(\hat{\theta} \mid \theta_{-i} = \theta^*, \theta^P \leq \theta^* \leq \hat{\theta})] = \\ \frac{1}{n}\int_{\underline{\theta}}^{\theta^*}\big(r_1 - c_1(\theta)\big)dF(\theta) + \big(1-\phi(n)\big)\frac{1}{n}\int_{\theta^*}^{\hat{\theta}}\big(r_1 - c_1(\theta)\big)dF(\theta) \\ + \frac{1}{n}\int_{\hat{\theta}}^{\overline{\theta}}\big(r_2 - c_2(\theta)\big)dF(\theta).\end{aligned} \tag{A-9}$$

Eq. (A-9) 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. Necessity requires that an upward deviation is not profitable at any $\hat{\theta} \geq \theta^*$:

$$\frac{d}{d\hat{\theta}}\big\{\mathbb{E}_\theta[\Pi_i(\hat{\theta} \mid \theta_{-i} = \theta^*, \theta^P \leq \theta^* \leq \hat{\theta})]\big\}\Big|_{\hat{\theta} \geq \theta^*} \leq 0. \tag{A-10}$$

Evaluating (A-10) yields the condition

$$\big(1-\phi(n)\big)\big(r_1 - c_1(\hat{\theta})\big) \leq r_2 - c_2(\hat{\theta})\ \forall \hat{\theta} \geq \theta^*. \tag{A-11}$$

Observe that when $\phi(n) = 0$, this condition is equivalent to the cut-off condition for the monopolist physician and that (A-10) evaluated at θ^* is sufficient as $c_1'(\cdot) \geq c_2'(\cdot)$ implies that the derivative holds for all $\hat{\theta} > \theta^*$.

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

$$\begin{aligned}\mathbb{E}_\theta[\Pi_i(\hat{\theta} \mid \theta_{-i} = \theta^*, \theta^P \leq \hat{\theta} \leq \theta^*)] = \\ \frac{1}{n}\int_{\underline{\theta}}^{\hat{\theta}}\big(r_1 - c_1(\theta)\big)dF(\theta) + \left[\phi(n)\left(\frac{n-1}{n}\right) + \frac{1}{n}\right]\int_{\hat{\theta}}^{\theta^*}\big(r_2 - c_2(\theta)\big)dF(\theta) \\ + \frac{1}{n}\int_{\theta^*}^{\overline{\theta}}\big(r_2 - c_2(\theta)\big)dF(\theta).\end{aligned} \tag{A-12}$$

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, necessity requires that it is not profitable at any $\hat{\theta} < \theta^*$. The change in profit from a downward deviation at $\hat{\theta} \leq \theta^*$ is

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

Evaluating (A-13) yields the condition

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

Again, when $\phi(n) = 0$ this condition is equivalent to the cut-off condition for the monopolist physician and (A-13) evaluated at θ^* is sufficient as $c_1'(\cdot) \geq c_2'(\cdot)$ implies that the derivative holds for all $\hat{\theta} < \theta^*$.

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 $\hat{\theta} < \theta^P$. Necessity requires that that

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

Evaluating (A-16) yields the condition $r_1 - c_1(\hat{\theta}) \geq \big(1 - \phi(n)\big)\big(r_2 - c_2(\hat{\theta})\big)\ \forall \hat{\theta} \leq \theta^P$, which is true whenever condition (A-14) is true. Lastly, I have shown that conditions (A-11) 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^* is nonnegative establishes sufficiency.

Proposition 5: When physician profit must be nonnegative ex post, or the profit is positive for both treatments at θ^* ((6a) and (6b) together imply $sign[r_1 - c_1(\theta^*)] = sign[r_2 - c_2(\theta^*)]$) then reimbursements $\{r_1, r_2\}$ can satisfy conditions (6a) and (6b) only if $1 + \phi(n)(n-1) \leq \big(1-\phi(n)\big)^{-1}$. Rearranging this inequality shows 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 $1 - \phi(n) \leq \frac{r_1 - c_1(\theta^P)}{r_2 - c_2(\theta^P)} \leq \big(1 - \phi(n)\big)^{-1}$. When $r_1 - c_1(\theta^*) < 0$ then both conditions can be simultaneously satisfied if $1 + \phi(n)(n-1) > \big(1 - \phi(n)\big)^{-1}$. If $1 + \phi(n)(n-1) \leq \big(1 - \phi(n)\big)^{-1}$, however, then the payments can be increased so that the physicians' profit is positive at θ^* (allowing the physician to extract positive) profit using only reimbursements $\{r_1, r_2\}$.

Proposition 6: Given reimbursements $\{r_1, r_2\}$, the Pareto dominant equilibrium is the θ where

$$r_1 - c_1(\theta^*) = \min\{1 + \phi(n)(n-1), (1 - \phi(n))^{-1}\} \cdot (r_2 - c_2(\theta^*)). \tag{A-17}$$

When $1 + \phi(n)(n-1) < (1 - \phi(n))^{-1}$ then $r_1 - c_1(\theta^*) > 0$ and $r_2 - c_2(\theta^*) > 0$. Because $c_1'(\cdot) > c_2'(\theta)$ anything that increases the right-hand-side of (A-17) lowers θ and anything that lowers the left-hand-side increases the equilibrium θ. The right-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. When $1 + \phi(n)(n-1) > (1 - \phi(n))^{-1}$ then $r_1 - c_1(\theta^*) < 0$ and $r_2 - c_2(\theta^*) < 0$ generating the opposite effect on ϕ and n.

When $1 + \phi(n)(n-1) < (1 - \phi(n))^{-1}$ then from the insurer's maximization program r_2^* must satisfy

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

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 ϕ 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.

Similarly, when $1+\phi(n)(n-1) > (1-\phi(n))^{-1}$ then from the insurer's maximization program r_2^* must satisfy

$$r_2^*\big[\big[(1-\phi(n))^{-1}-1\big]F(\theta^*)+1\big] = \mathbb{E}_\theta[c(\theta \mid \theta^*)] + F(\theta^*)\big[c_2(\theta^*)(1-\phi(n))^{-1} - c_1(\theta^*)\big]. \tag{A-21}$$

Totally differentiating (A-21) with respect to ϕ and n yields the results.

Proposition 7: 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 testing and 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 point-wise optimization, the FOCs for σ_1 and σ_2 yield:

$$\int_\Theta \big[\psi(\theta,T_j) - \lambda c_j(\theta)\big] dG(\theta \mid \xi) = 0 \text{ for } j \in \{1,2\}. \tag{A-22}$$

The FOC for σ_3 is:

$$\int_\Theta \left[\sum_k \tau_k(\theta)\psi(\theta,T_k) - \lambda \sum_k \tau_k(\theta)[c_k(\theta)+D]\right] dG(\theta \mid \xi) = 0, \tag{A-23}$$

and the FOC for P is

$$-U'(Y-P) - \lambda = 0, \tag{A-24}$$

where λ is the Lagrangian multiplier. The FOCs are all independent of the σ_j thus we can choose the diagnostic regime 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-22) and (A-23). Eq. (A-23) assumes its maximal value when $\tau_1(\theta) = 1$ for all $\theta < \theta^{FB}$, $\tau_1(\theta) = 0$ otherwise, and $\tau_2(\theta) = 1 - \tau_1(\theta)$ for all $\theta \in \Theta$.

The lower diagnostic cut-off, ξ_L^{FB}, represents the signal with which the FOCs for σ_1 and σ_3 are equal:

(A-25)

$$\int_{\underline{\theta}}^{\overline{\theta}} [\psi(\theta, T_1) - \lambda c_1(\theta)] dG(\theta \mid \xi_L^{FB}) = \int_{\underline{\theta}}^{\theta^{FB}} [\psi(\theta, T_1) - \lambda c_1(\theta)] dG(\theta \mid \xi_L^{FB}) + \int_{\theta^{FB}}^{\overline{\theta}} [\psi(\theta, T_2) - \lambda c_2(\theta)] dG(\theta \mid \xi_L^{FB}) - \lambda D,$$

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

(A-26)

$$\int_{\underline{\theta}}^{\overline{\theta}} [\psi(\theta, T_2) - \lambda c_2(\theta)] dG(\theta \mid \xi_U^{FB}) = \int_{\underline{\theta}}^{\theta^{FB}} [\psi(\theta, T_1) - \lambda c_1(\theta)] dG(\theta \mid \xi_U^{FB}) + \int_{\theta^{FB}}^{\overline{\theta}} [\psi(\theta, T_2) - \lambda c_2(\theta)] dG(\theta \mid \xi_U^{FB}) - \lambda D.$$

If there is no ξ_L^{FB} satisfying (A-25), then the test should be conducted for all lower signals down to $\underline{\xi}$; and if there is no ξ_U^{FB} satisfying (A-26), then the test should be conducted for all lower signals up to $\overline{\xi}$. Both of these cases follow because it is assumed that there exist some signals for which it is efficient to conduct the test. If this assumption is not made then it can be the case that $\xi_L^{FB} > \xi_U^{FB}$, in which case it would be optimal to never conduct the diagnostic test.

Lastly, the first-best premium is simply the expected cost of treatment including the diagnostic test.

Proposition 8: First, $E_\theta[\Pi(\theta^*, \xi_L^*, \xi_U^*; r_1, r_2, \hat{r}_1, \hat{r}_2)] \geq 0$ is a standard individual rationality constraint which can always be satisfied with any reimbursements when a lump-sum payment $R > 0$ is included in the contract. 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 an incentive compatibility issue as the physician can always administer the test without reporting doing so to receive the higher payment.

Conditional on conducting the diagnostic test, patients will prefer physicians that select the treatment cut-off that is closest to their preferred cut-off. However, patients do not know whether they are a marginal type so any patient receiving a diagnostic signal $\xi \in [\xi_L^*, \xi_U^*]$ will select the physician with the most preferred treatment practice. Therefore, a physician that follows the same testing regime as the other physicians, but selects a less preferred treatment

cut-off will have expected profit:

$$\mathbb{E}_{\xi,\theta}[\Pi_i(\hat{\theta},\xi_L^*,\xi_U^* \mid \hat{\theta} > \theta_{-i} = \theta^*)] =$$
$$\frac{1}{n}\int_{\underline{\xi}}^{\xi_L^*} \mathbb{E}_\theta[\Pi(\overline{\theta}) \mid \xi] d\Gamma(\xi) + \left(\frac{1-\phi(n)}{n}\right)\int_{\xi_L^*}^{\hat{\xi}_U^*} \big\{\mathbb{E}_\theta[\Pi(\hat{\theta}) \mid \xi] - D\big\} d\Gamma(\xi)$$
$$+ \frac{1}{n}\int_{\xi_U^*}^{\overline{\xi}} \mathbb{E}_\theta[\Pi(\underline{\theta}) \mid \xi] d\Gamma(\xi) + R,$$

where $\mathbb{E}_\theta[\Pi(\theta) \mid \xi] = \int_{\underline{\theta}}^{\theta} \big(r_1 - c_1(\tilde{\theta})\big) dF(\tilde{\theta} \mid \xi) + \int_{\theta}^{\overline{\theta}} \big(r_2 - c_2(\tilde{\theta})\big) dF(\tilde{\theta} \mid \xi)$.

Similarly, the expected profit of a physician who follows the same testing regime as the other physicians, but selects a more preferred treatment cut-off will have expected profit

$$\mathbb{E}_{\xi,\theta}[\Pi_i(\hat{\theta},\xi_L^*,\xi_U^* \mid \hat{\theta} < \theta_{-i} = \theta^*)] =$$
$$\frac{1}{n}\int_{\underline{\xi}}^{\xi_L^*} \mathbb{E}_\theta[\Pi(\overline{\theta}) \mid \xi] d\Gamma(\xi) + \left[\phi(n)\left(\frac{n-1}{n}\right) + \frac{1}{n}\right]\int_{\xi_L^*}^{\hat{\xi}_U^*} \big\{\mathbb{E}_\theta[\Pi(\hat{\theta}) \mid \xi] - D\big\} d\Gamma(\xi)$$
$$+ \frac{1}{n}\int_{\xi_U^*}^{\overline{\xi}} \mathbb{E}_\theta[\Pi(\underline{\theta}) \mid \xi] d\Gamma(\xi) + R.$$

Preventing the first deviation requires

$$\left(\frac{1-\phi(n)}{n}\right)\int_{\xi_L^*}^{\hat{\xi}_U^*} \big\{\mathbb{E}_\theta[\Pi(\hat{\theta}) \mid \xi] - D\big\} d\Gamma(\xi) \leq \int_{\xi_L^*}^{\hat{\xi}_U^*} \big\{\mathbb{E}_\theta[\Pi(\theta^*) \mid \xi] - D\big\} d\Gamma(\xi) \tag{A-27}$$

while preventing the second deviation requires

$$\left[\phi(n)\left(\frac{n-1}{n}\right) + \frac{1}{n}\right]\int_{\xi_L^*}^{\hat{\xi}_U^*} \big\{\mathbb{E}_\theta[\Pi(\hat{\theta}) \mid \xi] - D\big\} d\Gamma(\xi)$$
$$\leq \frac{1}{n}\int_{\xi_L^*}^{\hat{\xi}_U^*} \big\{\mathbb{E}_\theta[\Pi(\theta^*) \mid \xi] - D\big\} d\Gamma(\xi). \tag{A-28}$$

From the physician's FOC it is clear that her profit is maximized at the treatment cut-off θ^* if and only if $r_1 - c_1(\theta^*) = r_2 - c_2(\theta^*)$. Therefore (A-27) is satisfied as long as $\int_{\xi_L^*}^{\hat{\xi}_U^*} \mathbb{E}_\theta[\Pi(\theta^*) \mid \xi] d\Gamma(\xi) \geq D$ and (A-28) is satisfied as long as $\int_{\xi_L^*}^{\hat{\xi}_U^*} \mathbb{E}_\theta[\Pi(\theta^*) \mid \xi] d\Gamma(\xi) \leq D$ yielding the requirement

$$\int_{\xi_L^*}^{\hat{\xi}_U^*} \mathbb{E}_\theta[\Pi(\theta^*) \mid \xi] d\Gamma(\xi) = \mathbb{E}_\theta[\Pi(\theta^*) \mid \xi_L^* \leq \xi \leq \xi_U^*] = D.$$

To show conditions (11) and (12) are necessary start by assuming that all physicians choose treatment cut-off θ^* and lower diagnostic cut-off ξ_L^*. If all other physicians have also chosen some upper diagnostic cut-off $\xi_{-i} = \xi_U^*$ then the profit for a physician who deviates to a lower

upper cut-off is expressed as

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

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:

$$\frac{d}{d\hat{\xi}^u}\big\{\mathbb{E}_{\xi,\theta}[\Pi_i(\theta^*, \xi_L^*, \hat{\xi} \mid \xi_{-i} = \xi_U^*, \hat{\xi} \leq \xi_U^*)]\big\}\Big|_{\hat{\xi}=\xi_U^*} \geq 0 \tag{A-29}$$

The derivative is positive because the physician is deviating downward. Furthermore, the expression need only hold at ξ_U^* because the benefit of testing is diminishing with higher ξ so if the derivative is positive at ξ_U^* it is positive at all $\xi < \xi_U^*$. Evaluating (A-29) yields the condition

$$\mathbb{E}_\theta[\Pi(\theta^*) \mid \xi_U^*] - D \geq \big(1-\phi(n)\big)\mathbb{E}_\theta[\Pi(\underline{\theta}) \mid \xi_U^*]. \tag{A-30}$$

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

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

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 since patients prefer the higher diagnostic cut-off chosen by the deviating physician. Equilibrium requires that a physician will not be better off from any deviation; i.e., the following is a necessary condition for equilibrium:

$$\frac{d}{d\hat{\xi}^u}\big\{\mathbb{E}_{\xi,\theta}[\Pi_i(\theta^*, \xi_L^*, \hat{\xi} \mid \xi_{-i} = \xi_U^*, \hat{\xi} \geq \xi_U^*)]\big\}\Big|_{\hat{\xi}=\xi_U^*} \leq 0 \tag{A-31}$$

The derivative is negative because the physician is deviating upward. Furthermore, the expression need only hold at ξ_U^* because the benefit of testing is diminishing with higher ξ so if the derivative is negative at ξ_U^* it is negative at all $\xi > \xi_U^*$. Evaluating (A-31) yields the condition

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

Combining (A-30) and (A-32) yields the condition

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

Note that the inequalities flipped from what is indicated by (A-30) and (A-32) because (10) implies that $\mathbb{E}_\theta[\Pi(\theta^*) \mid \xi_U^*] - D < 0$.

The proof for the lower diagnostic cut-off follows similarly however $\mathbb{E}_\theta[\Pi(\theta^*) \mid \xi_L^*] - D > 0$ so there is no reversal in the signs.

Corollary 3: The corollary follows immediately from rearranging (A-32) and the analogous condition for the lower diagnostic condition. These conditions can fail to be satisfied with sufficiently low testing cost D because the expected profit from treatment without conducting the test is always lower than the expected profit from treatment after conducting the test since the physician will gain the ability to choose the optimal treatment. Allowing for higher reimbursement conditional on conducting the test only amplifies the difference in expected cost.

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 giving insurers some market power, then the premium will be greater than marginal cost and there will be a loss in consumer surplus. Because of the monotonicity of the difference in costs the insurer still chooses 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.[43] In this case the insurer's optimization program can be expressed as

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

[43]Strictly speaking I also assume that the insurer has all of the bargaining power such that the physician must accept whatever payments the insurer offers subject to the physician's participation constraint. 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.

subject to

$$U(Y-P)+\int_{\underline{\theta}}^{\theta^*}\psi(\theta,T_1)dF(\theta)+\int_{\theta^*}^{\overline{\theta}}\psi(\theta,T_2)dF(\theta)\geq U(Y)$$

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

$r_1-c_1(\theta^*)=r_2-c_2(\theta^*)$, and

$\theta^*\in\Theta$.

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.

As reported by Proposition 2, leaving the physician with zero economic profit requires $\int_{\underline{\theta}}^{\theta^*} r_1 dF(\theta)+\int_{\theta^*}^{\overline{\theta}} r_2 dF(\theta)=\mathbb{E}_\theta[c(\theta\mid\theta^*)]$. 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 extracts the patients' surplus through the premium whereas the social surplus-maximizing insurer leaves the surplus with the patients. If the patients had positive surplus at the social optimum, then this implies that the profit-maximizing insurer's premium is greater. As a consequence of the fact that a higher premium increases the patients' marginal utility (by lowering the patient's utility from alternative expenditures) the profit-maximizing insurer will increase the treatment cut-off 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 select cut-off $\theta^* > \theta^{FB}$ inducing the use of treatment T_1 on more patients than an insurer who maximizes total social welfare.*

Proof. The optimal reimbursements for a given θ^* are the reimbursements identified in Proposition 2:

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

Plugging these payments 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 $U(Y-P)+\int_{\underline{\theta}}^{\theta^*}\psi(\theta,T_1)dF(\theta)+\int_{\theta^*}^{\overline{\theta}}\psi(\theta,T_2)dF(\theta)\geq U(Y)$ 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 sets the premium to $P^{FB} = \mathbb{E}_\theta[c(\theta \mid \theta^E)]$ and leaves all of the surplus with the patients. However, as long as patients receive some positive surplus, then the profit-maximizing insurer will extract that surplus and $P^* > \mathbb{E}_\theta[c(\theta \mid \theta^E)]$. As $U(\cdot)$ is concave $U'(Y - P^*) > U'(Y - P^{FB})$. Furthermore, the concavity of the insurer's problem and the fact that $c_1(\theta^{FB}) < c_2(\theta^{FB})$ and assumptions A1 and A2 together all imply $\theta^* > \theta^{FB}$. □

APPENDIX C. PATIENT UNCERTAINTY IN ILLNESS SEVERITY

Instead of knowing his own type, θ, a patient may instead have a private signal in the form of symptoms indicating how severely ill he may be. The main model in the paper illustrates the physician treatment problem when the patients' signal precisely indicates their illness severity. Nevertheless, if the symptoms are sufficiently correlated with the patient's true illness severity such that signals reduce the possible illness severities to a subset of Θ, then the insights of the main model continue to hold and the payments must be adjusted to reflect the degree of competition in the market. If, however, the symptoms do not provide patients with enough additional information about their illness severity such that any illness severity is possible given a particular signal, then all patients are effectively marginal and the degree of competition is irrelevant.

The impact on the reimbursements is as follows. When the signal precisely identifies the illness severity then a deviating physician will either gain or lose only those patients who benefit from the deviation. For example, if the physician deviates towards a more preferred cut-off $\hat{\theta} < \theta^*$, then that physician will gain $\left[F(\theta^*) - F(\hat{\theta})\right]\phi(n)(n-1)/n$ patients. A physician will make such a deviation if the profit gained from the additional patients exceeds the loss in using the less profitable treatment. As the signal becomes noisier in the sense that the set of possible illness severities increases for given signal, the physician that deviates towards a more attractive cut-off will attract those patients that could benefit from the deviation based on their signal; however, some of those patients will have an illness severity that falls outside of $(\hat{\theta}, \theta^*)$ and will receive the same treatment that a non-deviating physician would provide. The deviation is more profitable for the deviating physician, though, because she experiences no loss in profit from the patients whose true type is outside of the set $(\hat{\theta}, \theta^*)$ while increasing her demand from these types.

Taken to the extreme, when the signal does not reduce the set of possible illness severities, a deviating physician will attract all patients when selecting a more preferred cut-off and lose all patients when selecting a less preferred treatment cutoff (subject to the market informedness $\phi(n)$). Preventing physicians from deviating to a less preferred treatment option is straightforward since the physician will lose demand. However, preventing a deviation towards a more preferred treatment option is more difficult because an arbitrarily small change will result in almost no loss in profit on existing patients while allowing the physician to attract an addition $\phi(n)(n-1)/n$ patients. In consequence it will have to be the case that the physicians' economic profit is both at its maximum and zero at θ^*, which occurs when $\{r_1, r_2\}$ satisfy Proposition 2. In this way any arbitrarily small change in treatment cut-off will leave the physician with negative economic profit.

To formally see how the patients' imperfect information impacts physician treatment choice the following addition is made to the model. Instead of knowing their type, patients receive a private signal η of their illness severity where η is drawn from the same set as the illness severities; i.e., $\eta \in [\underline{\theta}, \overline{\theta}] = \Theta$. The distribution of signals G is known by patients and physicians

so that for a given signal patients know the possible illness severities; however, the signal is private information. The conditional distribution of illness severities $F(\theta \mid \eta)$ is continuous and piecewise differentiable. For a given signal η, the expected illness severity is $\mathbb{E}[\theta \mid \eta] \equiv \int_{\underline{\theta}}^{\overline{\theta}} \theta dF(\theta \mid \eta)$.

Let $\underline{\eta}(\theta) \equiv \inf_{\eta \in \Theta}\{1 - F(\theta \mid \eta) = 0\}$; that is, $\underline{\eta}(\theta)$ is the signal with which θ is the lower support of the distribution of illness severities that can generate the signal. If the infimum does not exist then $\underline{\eta}(\theta) = \underline{\theta}$. By this definition, if a physician chooses cut-off $\hat{\theta}$ where all other physicians choose θ^* and $\theta^P < \hat{\theta} < \theta^*$, then all patients receiving a signal very close to (but greater than) $\underline{\eta}(\hat{\theta})$ will prefer to visit that physician over the others; however, the realized illness severity of these patients can be something much lower. Next, let $\overline{\eta}(\theta) \equiv \sup_{\theta \in \Theta}\{F\big(\theta \mid \eta)\big) = 0\}$; that is, $\overline{\eta}(\theta)$ is the signal with which θ is the upper support of the distribution of illness severities that can generate the signal. If the supremum does not exist then $\overline{\eta}(\theta) = \overline{\theta}$. By this definition, if a physician chooses cut-off $\hat{\theta}$ where all other physicians choose θ^* and $\theta^P < \theta^* < \hat{\theta}$, then all patients receiving a signal very close to (but less than) $\overline{\eta}(\hat{\theta})$ will prefer to visit the other physicians over the one choosing a higher cut-off; however, the realized illness severity of these patients can be higher.

Observe that if η provides insufficient information to reduce to the set of possible severities then $\underline{\eta}(\theta) = \underline{\theta}$ and $\overline{\eta}(\theta) = \overline{\theta}$ for all $\theta \in \Theta$ and $d\underline{\eta}(\theta)/d\theta = d\overline{\eta}(\theta)/d\theta = 0$. The piecewise differentiability of the conditional distribution function guarantees that $\underline{\eta}(\cdot)$ and $\overline{\eta}(\cdot)$ are piecewise continuously differentiable as well. Using the functions $\underline{\eta}(\cdot)$ and $\overline{\eta}(\cdot)$ Proposition 4 becomes

Proposition 10. *Given reimbursements $\{r_1, r_2\}$, fixed payment $R \geq 0$, $n \geq 2$ physicians, and market informedness $\phi(n) \in [0, 1)$ a cut-off θ^*, where $\theta^P < \theta^* < \overline{\theta}$, represents an equilibrium if and only if $\mathbb{E}_\theta[\Pi(\theta^* \mid \underline{\eta}(\theta^*) \leq \eta \leq \overline{\eta}(\theta^*)] = 0$, $\mathbb{E}_\theta[\Pi(\theta^*; r_1, r_2)] \geq 0$ and the payment rule satisfies:*

$$\int_{\theta^*}^{\overline{\theta}} [r_2 - c_2(\theta)] dF(\theta \mid \overline{\eta}(\theta^*)) \frac{d\overline{\eta}(\theta^*)}{d\theta} g(\overline{\eta}(\theta^*)) \geq$$

$$\big(1 - \phi(n)\big) \left[\begin{array}{l} \int_{\overline{\eta}(\theta^*)}^{\overline{\theta}} \big[r_2 - c_2(\theta)\big] f(\theta \mid \overline{\eta}(\theta^*)) \frac{d\overline{\eta}(\theta^*)}{d\theta} g(\overline{\eta}(\theta^*)) \\ \qquad + \int_{\underline{\eta}(\theta^*)}^{\overline{\eta}(\theta^*)} \big[r_1 - c_1(\theta^*) - (r_2 - c_2(\theta^*))\big]) f(\theta^* \mid \eta) dG \end{array} \right]$$

and

$$\int_{\underline{\theta}}^{\theta^*} [r_1 - c_1(\theta)] dF(\theta \mid \underline{\eta}(\theta^*)) \frac{d\underline{\eta}(\theta^*)}{d\theta} g(\underline{\eta}(\theta^*)) \geq$$

$$\big(1 + \phi(n)(n-1)\big) \left[\begin{array}{l} \int_{\underline{\theta}}^{\underline{\eta}(\theta^*)} \big[r_1 - c_1(\theta)\big] f(\theta \mid \overline{\eta}(\theta^*)) \frac{d\underline{\eta}(\theta^*)}{d\theta} g(\underline{\eta}(\theta^*)) \\ \qquad - \int_{\underline{\eta}(\theta^*)}^{\overline{\eta}(\theta^*)} \big[r_1 - c_1(\theta^*) - (r_2 - c_2(\theta^*))\big]) f(\theta^* \mid \eta) dG \end{array} \right],$$

where

$$\mathbb{E}_\theta[\Pi(\theta^* \mid \eta \in \big(\underline{\eta}(\theta^*), \overline{\eta}(\theta^*)\big)] = \int_{\underline{\eta}(\theta^*)}^{\overline{\eta}(\theta^*)} \left\{ \int_{\underline{\theta}}^{\theta^*} [r_1 - c_1(\theta)] dF(\theta \mid \eta) + \int_{\hat{\theta}}^{\overline{\theta}} [r_2 - c_2(\theta)] dF(\theta \mid \eta) \right\} d\Gamma.$$

Proof. The condition $\mathbb{E}_\theta[\Pi(\theta^*; r_1, r_2)] \geq 0$ is simply a participation constraint while the others are incentive constraints. A fixed payment R may be necessary because of this constraint and the added constraint $\mathbb{E}_\theta[\Pi(\theta^* \mid \underline{\eta}(\theta^*) \leq \eta \leq \overline{\eta}(\theta^*)] = 0$. 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 $\Pi(\theta^*) \equiv \Pi(\theta^* \mid \theta_{-j} = \theta^*) \geq \Pi(\theta \mid \theta_{-j} = \theta^*)$ for all $\theta \neq \theta^*$ and for all physicians j.

A physician may deviate upwards or downwards and either lose or gain patients, respectively. The expected profit of a physician who deviates to a higher cut-off given the other physicians choose cut-off θ^* is

(C-1)

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

Eq. (C-1) shows that a physician who deviates upward will lose $\big[G(\overline{\eta}(\hat{\theta})) - G(\underline{\eta}(\theta^*))\big]\phi(n)/n$ patients, but will now treat $\big[F(\hat{\theta} \mid \underline{\eta}(\theta^*) \leq \eta \leq \overline{\eta}(\hat{\theta})) - F(\theta^* \mid \underline{\eta}(\theta^*) \leq \eta \leq \overline{\eta}(\hat{\theta}))\big](1-\phi(n))/n$ with T_1 instead of T_2. It is clear that when $\eta = \theta$ (i.e., the signal reveals the true illness severity), then (C-1) is equivalent to (A-9) in Proposition 4.

The following is a necessary condition for an equilibrium at θ^*:

$$\text{(C-2)} \qquad \frac{d}{d\hat{\theta}} \big\{ \mathbb{E}_\theta[\Pi_i(\hat{\theta} \mid \theta_{-i} = \theta^*, \theta^P \leq \theta^* \leq \hat{\theta})] \big\} \Big|_{\hat{\theta} = \theta^*} \leq 0.$$

Evaluating (C-2) yields the condition

(C-3)

$$\begin{aligned}
&\int_{\theta^*}^{\overline{\theta}} [r_2 - c_2(\theta)] dF(\theta \mid \overline{\eta}(\theta^*)) \frac{d\overline{\eta}(\theta^*)}{d\theta} g(\overline{\eta}(\theta^*)) \geq \\
&\qquad \big(1 - \phi(n)\big) \left[\begin{array}{l} \int_{\overline{\eta}(\theta^*)}^{\overline{\theta}} \big[r_2 - c_2(\theta)\big] f(\theta \mid \overline{\eta}(\theta^*)) \frac{d\overline{\eta}(\theta^*)}{d\theta} g(\overline{\eta}(\theta^*)) \\ \qquad + \int_{\underline{\eta}(\theta^*)}^{\overline{\eta}(\theta^*)} \big[r_1 - c_1(\theta^*) - (r_2 - c_2(\theta^*))\big]) f(\theta^* \mid \eta) dG \end{array} \right]
\end{aligned}$$

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

(C-4)

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

A physician who deviates downward can thus expect to attract an additional $\phi(n)(1 - n^{-1})\big[G(\overline{\eta}(\theta^*)) - G(\underline{\eta}(\hat{\theta}))\big]$ patients since she chooses a more preferred treatment for types $\hat{\theta}$ to θ^*. It is clear that when $\eta = \theta$, then (C-4) is equivalent to (A-12) in Proposition 4.

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.,

$$\text{(C-5)} \qquad \frac{d}{d\hat{\theta}}\big\{\mathbb{E}_\theta[\Pi_i(\hat{\theta} \mid \theta_{-i} = \theta^*, \theta^P \leq \hat{\theta} \leq \theta^*)]\big\}\Big|_{\hat{\theta}=\theta^*} \geq 0.$$

Evaluating (C-5) yields the condition

(C-6)

$$\int_{\underline{\theta}}^{\theta^*}[r_1 - c_1(\theta)]dF(\theta \mid \underline{\eta}(\theta^*))\frac{d\underline{\eta}(\theta^*)}{d\theta}g(\underline{\eta}(\theta^*)) \geq$$
$$\big(1+\phi(n)(n-1)\big)\left[\begin{array}{l}\int_{\underline{\theta}}^{\underline{\eta}(\theta^*)}\big[r_1 - c_1(\theta)\big]f(\theta \mid \overline{\eta}(\theta^*))\frac{d\underline{\eta}(\theta^*)}{d\theta}g(\underline{\eta}(\theta^*)) \\ \qquad\qquad - \int_{\underline{\eta}(\theta^*)}^{\overline{\eta}(\theta^*)}\big[r_1 - c_1(\theta^*) - (r_2 - c_2(\theta^*))\big]\big)f(\theta^* \mid \eta)dG\end{array}\right]$$

Conditions (C-3) and (C-6) are both necessary. However, they only establish that the deviation profit is maximized at the treatment cut-off θ^*. A physician that selects a cut-off arbitrarily close to θ^* still experiences a discontinuous jump in her profit. Therefore to ensure that such a jump is not profitable, it must also be the case that

$$\big(1-\phi(n)\big)\int_{\underline{\eta}(\theta^*)}^{\overline{\eta}(\theta^*)}\left[\int_{\underline{\theta}}^{\theta^*}[r_1 - c_1(\theta)]dF(\theta \mid \eta) + \int_{\theta^*}^{\overline{\theta}}[r_2 - c_2(\theta)]dF(\theta \mid \eta)\right]dG(\eta)$$
$$\leq \int_{\underline{\eta}(\theta^*)}^{\overline{\eta}(\hat{\theta})}\left[\int_{\underline{\theta}}^{\theta^*}[r_1 - c_1(\theta)]dF(\theta \mid \eta) + \int_{\theta^*}^{\overline{\theta}}[r_2 - c_2(\theta)]dF(\theta \mid \eta)\right]dG(\eta)$$

and

$$\big(1+\phi(n)(n-1)\big)\int_{\underline{\eta}(\theta^*)}^{\overline{\eta}(\theta^*)}\left[\int_{\underline{\theta}}^{\theta^*}[r_1-c_1(\theta)]dF(\theta\mid\eta)+\int_{\theta^*}^{\overline{\theta}}[r_2-c_2(\theta)]dF(\theta\mid\eta)\right]dG(\eta)$$
$$\leq\int_{\underline{\eta}(\theta^*)}^{\overline{\eta}(\hat{\theta})}\left[\int_{\underline{\theta}}^{\theta^*}[r_1-c_1(\theta)]dF(\theta\mid\eta)+\int_{\theta^*}^{\overline{\theta}}[r_2-c_2(\theta)]dF(\theta\mid\eta)\right]dG(\eta),$$

which together yield the condition $\mathbb{E}_\theta[\Pi(\theta^*\mid\underline{\eta}(\theta^*)\leq\eta\leq\overline{\eta}(\theta^*)]=0$. □

Note that when $\underline{\eta}(\theta)=\underline{\theta}$ and ,$\overline{\eta}(\theta)=\overline{\theta}$ for all $\theta\in\Theta$ we have $\frac{d\underline{\eta}(\theta^*)}{d\theta}=\frac{d\overline{\eta}(\theta^*)}{d\theta}=0$. In this case conditions (C-3) and (C-6) reduce to $r_1-c_1(\theta^*)=r_1-c_1(\theta^*)$, the monopoly payment rule. Furthermore, $\mathbb{E}_\theta[\Pi(\theta^*\mid\underline{\eta}(\theta^*)\leq\eta\leq\overline{\eta}(\theta^*)]=0$ is equivalent to $\int_{\underline{\theta}}^{\theta^*}[r_1-c_1(\theta)]dF(\theta)+\int_{\hat{\theta}}^{\overline{\theta}}[r_2-c_2(\theta)]dF(\theta)=0$ and both conditions are achieved using the monopoly payments defined in Proposition 2.

APPENDIX D. MULTIPLE TREATMENTS

The basic model considers only two treatments in order to focus on physicians' choice of one treatment over another. However, the model can be easily extended to consider the physician's decision over multiple treatments including a null treatment in which no clinical intervention is performed by the physician. Similar to the two treatment regime we must impose some conditions on the cost and value of alternative treatments. For example, additional treatments are only relevant to the problem if they are socially optimal for some illness severity. Otherwise an insurer can easily induce a physician to never choose such an alternative by not reimbursing the physician. The second assumption that must be imposed is that of concavity over treatments, which naturally follows from the fact that if one treatment is superior to another for a more severely ill patient, then it will continue to be superior for an even more severely ill patient. To generalize the model to M treatments, assumptions A1 and A2 can be restated as:

A1: $d\{\psi(\theta,T_{m-1})-\psi(\theta,T_m)\}/d\theta\leq 0\;\forall\theta\in\Theta,\forall m\in 2,\ldots,M$,
A2: $d\{c_{m-1}(\theta)-c_m(\theta)\}/d\theta\geq 0\;\forall\theta\in\Theta,\forall m\in 2,\ldots,M$, and
A3: For every $m\in\{1,\ldots,M\}$ there is a type $\theta\in\Theta$ such that T_m is socially optimal.

Assumptions A1 and A2 impose concavity on the insurer's problem and convexity to the firm's cost function, ensuring that there is an optimal treatment plan. Assumption A3 states that each of the treatments will be optimal for some set of types ruling out the possibility that there are some treatments that are never optimal. This is without loss of generality as the insurer can always not provide a reimbursement for those treatments that are never socially optimal.

Let θ_m^* represent the desired cut-off type when choosing between any two treatments T_{m-1} and T_m where the insurer prefers physicians use treatment T_{m-1} for types below θ_m^* and treatment T_m for types at or above θ_m^* because of A1 and A2. Together, A1 – A3 imply that there are $M-1$ treatment cut-offs where $\theta_2^*<\theta_3^*<\ldots<\theta_M^*$. Lastly, to keep the analysis brief, assume $1+\phi(n)(n-1)<(1-\phi(n))^{-1}$. Given this setup Proposition 4 can be restated as follows.

Proposition 11. *Given reimbursements $\{r_1,\ldots,r_M\}$, $n\geq 2$ physicians, and patient response $\phi(n)\in[0,1)$, cut-offs $\{\theta_2^*,\ldots,\theta_M^*\}$, where $\theta_m^P<\theta_m^*<\overline{\theta}$ for all $m\in\{1,\ldots,M\}$, represents*

an equilibrium if and only if $\mathbb{E}_\theta[\Pi(\theta; r_1, \ldots, r_M)] \geq 0$ *and the reimbursements satisfy:*

$$\text{(D-1)} \qquad 1 + \phi(n)(n-1) \leq \frac{r_{m-1} - c_{m-1}(\theta_m^*)}{r_m - c_m(\theta_m^*)} \leq \big(1 - \phi(n)\big)^{-1} \; \forall m \in \{2, \ldots, M\}.$$

The derivation of (D-1) follows directly from the proof for Proposition 4. Proposition 11 reports $2 \times (M-1) + 1$ constraints that the M payments must satisfy. However, because choosing the cut-off $\hat{\theta}_m$ that solves

$$\text{(D-2)} \qquad \big[1 + \phi(n)(n-1)\big]\big(r_m - c_m(\hat{\theta}_m)\big) = r_{m-1} - c_{m-1}(\hat{\theta}_m)$$

is Pareto dominant for the physicians, the insurer need only ensure that the $\{r_1, \ldots, r_M\}$ satisfy the $M-1$ constraints defined by (D-2) and an Mth participation constraint: $\mathbb{E}_\theta[\Pi(\theta; r_1, \ldots, r_M)] \geq 0$ in order to induce a particular treatment plan $\{\theta_2^*, \ldots, \theta_M^*\}$. Hence, as long as $1 + \phi(n)(n-1) \leq (1 - \phi(n))^{-1}$ an insurer will be able to induce its preferred treatment practice consisting of M treatments.