Health Insurance, Treatment Plan, and Delegation to Altruistic Physician

Ting Liu
Department of Economics
Michigan State Univeristy
110 Marshall-Adams Hall
East Lansing, MI 48824
tingliu@msu.edu

Ching-to Albert Ma
Department of Economics
Boston University
270 Bay State Road
Boston, MA 02215
ma@bu.edu

October 2011

Abstract

We study delegating a consumer's treatment plan decisions to an altruistic physician. The physician's degree of altruism is his private information. The consumer's illness severity will be learned by the physician, and also become his private information. Treatments are discrete choices, and can be combined to form treatment plans. We distinguish between two commitment regimes. In the first, the physician commits to treatment decisions at the time a payment contract is accepted. In the second, the physician does not commit to treatment decisions at that time, and can wait until he learns the patient's illness to do so. In the commitment game, the first best is implemented by a single payment contract to all types of the altruistic physician. In the noncommitment game, the first best is not implementable. All but the most altruistic physician earn positive profits, and treatment decisions are distorted from the first best.

Acknowledgment: We thank David Bardey, Pedro Barros, Chiara Canta, Yuk-fai Fong, Michael Luca, and Henry Mak for comments. We also thank David Martimont for discussing various issues with us. Many seminar and conference participants gave us comments and suggestions. Part of this research was done while the second author was visiting the Universidad Carlos III de Madrid and Tufts University; the hospitality at the two Economics Departments is gratefully acknowledged.

1 Introduction

Physicians have different practice styles. Researchers of the Dartmouth Atlas[1] project find that the fee-for-service Medicare spending on enrollees varies as much as 2.5 times across localities, even after controlling for differences in local prices, and the age, race and underlying health of the population. Much of this unwarranted and substantial variation in medical spending is attributed to physicians' practice styles.

Since physicians in the fee-for-service Medicare program face similar financial incentives, the variation in practice styles may reflect their views on what is best for patients (as well as profits). For example, in Fort Myers, Florida, 2.3 times as many Medicare enrollees per capita receive knee replacements as do recipients living in the neighboring Miami region. Doctors in Fort Myers are more likely to recommend surgical management of osteoarthritis of the knee while those in Miami favor medical management. Across the country, there are similar variations in surgical rates for such conditions as chronic angina, lower back pain, arthritis of the knee or hip, and early stage cancer of the prostate. (Wennberg et al. 2008). Surprisingly, patients' health outcomes in high-spending areas are no better than in low-spending areas.

In this paper, we study how an insurer can reduce the unnecessary cost due to practice-style variations by designing payment contracts for heterogenous physicians. We address the following questions: what is the efficient treatment plan when there are multiple treatment options? What is the optimal payment contract? Under what conditions can the efficient treatment plan be implemented? If the efficient plan is not implemented, what are the distortions? Finally, how are insurance premiums affected?

We model heterogeneous physicians by means of partial physician altruism; the physician's preferences depend on the patient's utility and his profit. Different physicians have different degrees of altruism. Since Arrow (1963) observed the importance of altruistic physicians for the health market, the altruistic-physician assumption has been adopted often.[2] Most papers in the literature have assumed that the degree of altruism

---

[1] See http://www.dartmouthatlas.org.

[2] A sample of papers using the assumption that physicians are altruistic includes Chalkley and Malcomson (1998), Choné and Ma (2010), Dranove and Spier (2003), Dusheiko et al. (2006), Ellis and McGuire (1986, 1990), Jack (2005), Ma (1998), Ma and Riordan (2002), Newhouse (1970), Rochaix (1989), and Rogerson (1994).

is given and known.[3] We go beyond the fixed-altruism assumption and allow many physician types, this being the physician's private information.

An altruistic physician may trade off his own profit against the consumer's utility. This formal construct does permit an ultra altruistic physician to run a financial loss to subsidize treatments. This, however, is unrealistic. Being an economic agent, a physician must face some financial constraints, so we assume that a physician must on average earn a minimum profit. We do allow a physician to sustain some financial loss sometimes, but he must expect to earn a minimum profit. We normalize this minimum expected profit to zero.[4]

The physician practice-style issue rests on an environment in which many treatment options for an illness are available. We model multiple treatment options in the simplest way, and let there be two treatments. A less costly treatment succeeds in eliminating a patient's illness disutility with a lower probability. A second treatment is more costly, but succeeds with a higher probability. Also, we propose a more realistic scenario, in which treatments can be combined. For example, a high-cost treatment may be used after a low-cost treatment fails to eradicate the illness. The physician decides on treatment plans, which are sequences of treatments.

Our main findings are the following. First, the first-best treatment plan prescribes a conservative approach under a cost-convexity assumption, which says that the higher the success probability, the higher is the cost per unit success probability. If the severity is low, then no treatment is used; if it is of medium value, a low-cost treatment will be used; if it is high, then the low-cost treatment will be used, followed by the high-cost treatment if necessary. In other words, the consumer should never take the high-cost treatment before trying the low-cost treatment.

Second, the first best can be implemented when the physician can commit to treatment plans at the same time the payment contract is accepted. Furthermore, it is implemented by a *single* contract. Complicated menus of contracts are not needed. This result is surprising because in principal-agent models, information

---

[3] All the papers in footnote 2 use the known altruism assumption except Choné and Ma (2010) and Jack (2005).

[4] Our results remain the same when the minimum profit is strictly positive.

asymmetry often generates information rent, and results in distortion. We find that the physician's private information about his preferences does not generate any information rent under treatment plan commitment.

The commitment issue is best explained by first describing the extensive form game. In Stage 1, an insurer offers an insurance contract to the consumer, and a payment contract to the physician. In Stage 2, nature determines the physician's degree of altruism, which is privately known to the physician. In Stage 3, the physician and the consumer decide whether to accept the contract. The physician also decides on a practice style which is a rule for prescribing a treatment plan for any illness severity. In Stage 4, nature determines the patient's illness severity. The physician learns the illness severity and follows the treatment plan decided earlier.

The first best can be implemented by a contract designed as if the physician is the least altruistic type. Suppose the least altruistic physician puts 10% weight on consumer's utility. The insurer should offers a contract with a 10% cost share and a transfer equal to 10% of the expected first-best cost. The 10% altruistic physician will fully internalize the social costs and benefits when bearing 10% of the cost. A lump-sum transfer equal to 10% of the expected cost in the first best allows the least altruistic physician to break even.

Why can this contract still implement the first best when the physician puts, say, a 50% weight on the consumer's utility? If the physician accepts the contract and implements the first best, he also breaks even. He would have liked to offer more generous treatments because he was more altruistic. But if he had done so, he would not break even. The transfer is so low—only 10% of the expected first-best cost—that more generous treatment plans would put the 50% physician in the red. The nonnegative expected profit constraint is so binding that the 50% physician must follow the strategy of the least altruistic physician.

We then study a game without commitment. The first two stages of the game are the same as before. But now in Stage 3, the physician only decides on whether to accept the payment contract. The treatment decision is postponed to Stage 4, after he learns the consumer's illness severity. In this game, the single contract specified in the game with commitment fails to implement the first best. The 50% altruistic physician will reject a 10% cost-share contract because he rationally anticipates offering treatments more generous than the first best. The low transfer corresponding to the 10% cost-share contract would not allow him to break even.

If the insurer has to retain a physician with high degrees of altruism, contracts with higher cost shares will have to be offered. In fact, a menu of incentive-compatible payment contracts will be offered. Information rent will have to be given up: all but the most altruistic physician will earn strictly positive profits. To limit information rent, distortions from first-best treatment plans will be implemented, and the insurance premium for the consumer may be higher.

The contrast between equilibria with commitment and without commitment is striking. It highlights the importance of the timing of treatment plan decision and the social value of treatment plan commitment. The single contract can implement the first best only when the doctor's treatment plan decision is made together with contract acceptance decision, and the doctor can commit to the predetermined plan when seeing patients. If the physician is able to follow such a treatment plan, he will accept the single contract. Although the physician just breaks even, he enjoys the utility derived from patients' well-being while the efficient treatments are carried out. Rejecting the single contract results in a zero profit and a lower utility from the consumer's well-being because the consumer's illness is left untreated.

Our result is normative in nature. It introduces a new and more efficient practice of medicine. If a treatment decision can be decided upon when the financial constraint is relevant, treatment efficiency can be attained. A sort of "bottomline medicine" principle is being advocated whereby resources and medical treatments should always be considered together. The policy implication is that the insurer should encourage doctors to formulate their treatment plans at the point of contract acceptance, and give doctors incentives to carry out the plan when seeing patients. For example, when offering the single contract, the insurer can announce that he will only renew contract with physicians who can break even. Although a physician may bear a financial loss from treating some patients even when he follows his treatment plan, he must break even on average. Being in red all the time indicates that the doctor has overused the more expensive treatment. The insurer can induce the doctor to stick to his treatment plan by threatening to terminate the contract if the doctor loses money consistently.

One often thinks that commitment is powerful, and this has been shown time and again in economic models. Yet, in our model, a physician earns a zero profit when he is able to commit to a treatment plan, but a positive profit when he is unable to do so. In other words, a physician's commitment ability is being exploited by the insurer. We do not suggest that physicians have an incentive to give up treatment plan commitment. In fact, physicians in our model are altruistic and their preferences are *not* based on profits alone. Furthermore, a physician's total utility may be higher when he is able to commit to treatment plans: although he is not making a profit, the higher utility from delivering first-best treatments may more than compensate.

This paper uses the assumption that economic agents have nonmonetary motives. This, by now, is quite common, as evidenced in recent papers by Akerlof and Kranton (2005), Bénabou and Tirole (2003), Besley and Ghatak (2005), Delfgaauw and Dur (2007, 2008), Francois (2000), Murdock (2002), and Prendergast (2007, 2008). In our model, the agent is the physician who cares about his own financial gains and the patient's welfare. Our paper differs from these works in that the physician's degree of altruism is unknown (see also footnotes 2 and 3 above).

Unknown altruism in the health market has been considered before by Jack (2005) and Choné and Ma (2010). Nevertheless, our paper differs in many ways. Jack's model considers choices of noncontractible quality by a provider, and lets the physician suffer some financial losses. We do not consider quality, and impose a nonnegative expected profit constraint. In Choné and Ma, health care quantities are contractible, and the physician possesses private information about both altruism and patient illness severity before accepting a contract. In Jack (2005) and Choné and Ma (2010), there are no equilibria in which the first best is implemented.

The literature on physician payment is large. An earlier survey is McGuire (2000), and a more recent one is Léger (2008). Despite the prevalence of multiple treatment options, most existing works either do not model treatment plans (Pauly (1968), Zeckhauser (1970), Choné and Ma (2010)), or allow patients to take only one treatment (Ma and Riordan (2002)). One exception is Chernew, Encinosa and Hirth (2000) who allow the patient to choose one treatment out of many options. Different from all these works, our model has multiple treatment options and considers treatment sequences. To the best of our knowledge, this is the first attempt to do that.

The rest of the paper is organized as follows. Section 2 presents the model and the first best. Section 3 studies the two delegation games. Section 4 discusses related issues and policy implications. Section 5 draws conclusions. All proofs are in an Appendix.

2 The model and the first best

A risk-averse consumer has income Y and suffers from an illness. The loss due to illness is described by a random variable ℓ on a support $[0, \bar{\ell}]$, with distribution and density functions $F(\ell)$ and $f(\ell) > 0$, respectively.[5] We assume that the upper support of the illness loss, $\bar{\ell}$, is sufficiently large. We let the consumer's utility function be separable in income and the loss from illness, and measure the disutility of illness by the loss, so the consumer's utility is $U(Y) - \ell$ when ℓ is the illness loss. The function U is strictly increasing, strictly concave, and the marginal utility at zero income is infinite ($U'(x) \to \infty$ as $x \to 0^+$).

The consumer's loss due to illness can be recovered by medical treatments. We assume that there are two treatments; in Section 4 we will discuss when more treatments are available. A treatment either recovers the loss ℓ or does not, and is defined by the probability of success and the cost. Treatment can be taken sequentially, so if a treatment does not succeed, a second treatment can be used. We assume that when a treatment fails once, it will fail again. In other words, the effectiveness of a treatment is perfectly correlated over trials. Given the binary structure, a treatment will never be used twice.

We call the two treatments, Treatment 1 and Treatment 2. Treatment 1 succeeds with probability θ_1 and costs c_1. Treatment 2 succeeds with probability θ_2 and costs c_2. These four parameters are strictly positive. Treatment 2 is more effective than Treatment 1 but also costs more, so we have $\theta_1 < \theta_2$ and $c_1 < c_2$. We make an assumption on the relative effectiveness of the treatments:

---

[5] Unlike most other models, we do not set up a probability of the consumer falling ill, upon which the loss occurs. Our model is slightly more general because we allow a large density around $\ell = 0$, so it can approximate models with a fixed probability of falling ill.

Assumption 1 (Cost Convexity) $\frac{c_1}{\theta_1} < \frac{c_2}{\theta_2}$.

Assumption 1 says that the cost per unit of success probability of Treatment 2 is higher than Treatment 1. This is a sort of convexity assumption on treatment costs; the cost per unit of success probability increases with the success probability. We will discuss what will happen if Assumption 1 is violated.

In this paper we consider *Treatment Protocols.* A treatment protocol describes a sequence of treatments. There are five treatment protocols:

Protocol 0: Do not use any treatment.

Protocol 1: Use Treatment 1 only.

Protocol 2: Use Treatment 2 only.

Protocol 3: Use Treatment 1, and then Treatment 2 if Treatment 1 fails.

Protocol 4: Use Treatment 2, and then Treatment 1 if Treatment 2 fails.

Again, because we have assumed that a treatment outcome is perfectly correlated across trials, Treatment Protocols do not include multiple trials of the same treatment. The *ex ante* success probabilities of Protocols 3 and 4 are, respectively, $\theta_3 \equiv \theta_1 + (1 - \theta_1)\theta_2$ and $\theta_4 \equiv \theta_2 + (1 - \theta_2)\theta_1$. These *ex ante* success probabilities are the same because each of Protocols 3 and 4 allows the consumer to try both treatments; each also offers a higher success probability than either Protocol 1 or Protocol 2.[6] The expected costs of Protocols 3 and 4 are, respectively, $c_3 \equiv c_1 + (1 - \theta_1)c_2$ and $c_4 \equiv c_2 + (1 - \theta_2)c_1$. By Assumption 1, Protocol 4 costs more than Protocol 3: $c_4 - c_3 = c_2\theta_1 - c_1\theta_2 > 0$.

Without any insurance, the consumer will decide on the treatment protocol after she learns her illness loss. For low values of ℓ, she may not get any treatment; for high values, she may. The consumer faces fluctuations in income since she has to bear treatment costs. The consumer can insure herself against income fluctuations due to illness by purchasing an insurance contract in a competitive insurance market. Insurers

---

[6] We abstract from time delays in treatments. We can certainly build into our model such delays. For example, if a treatment is used second, the success probability is reduced by a fraction.

are risk neutral, and they offer insurance contracts to maximize the consumer's expected utility subject to a breakeven constraint.[7]

2.1 First best

In the first best, illness loss ℓ is verifiable. An insurance contract can be made contingent on the value of ℓ. Due to risk aversion, the first best shields the consumer from all risks due to treatment costs. A first-best contract specifies a premium P and four treatment protocol functions $\tau_i : [0, \overline{\ell}] \longrightarrow [0,1]$, $i = 1,2,3,4$. The consumer pays P before the realization of ℓ, and will not incur any payment after ℓ is realized and when treatment is used. The function τ_i, $i = 1,2,3,4$, specifies the probability that Protocol i is to be used when the consumer's loss is ℓ. We have used the nontreatment Protocol 0 as default.

When the consumer suffers a loss ℓ and is treated by Protocol i, her expected payoff is $U(Y-P) - \ell + \theta_i \ell$. The first-best contract $(P, \tau_1, \tau_2, \tau_3, \tau_4)$ maximizes the consumer's expected utility

$$\int_0^{\overline{\ell}} [U(Y-P) - \ell + \sum\nolimits_{i=1}^{4} \tau_i(\ell)\theta_i \ell] \mathrm{d}F(\ell) \tag{1}$$

subject to the breakeven constraint

$$P = \int_0^{\overline{\ell}} \sum\nolimits_{i=1}^{4} \tau_i(\ell) c_i \mathrm{d}F(\ell) \tag{2}$$

and the boundary conditions

$$\sum\nolimits_{i=1}^{4} \tau_i(\ell) \leq 1 \qquad \text{and} \qquad 0 \leq \tau_i(\ell) \leq 1, \tag{3}$$

for each $\ell \in [0, \overline{\ell}]$ and $i = 1,2,3,4$. The utility function in (1) consists of the utility from the income less the premium, the utility loss ℓ, as well as the recovery prospects from the four treatment protocols. The breakeven constraint (2) ensures that any insurance firm offering the contract will make zero expected profit. The remaining constraints in (3) make sure that the treatment protocol probabilities are consistent.

First, we rank the relative cost effectiveness of the treatment protocols:

Lemma 1 *Under Assumption 1,* $\dfrac{c_1}{\theta_1} < \dfrac{c_3}{\theta_3} < \dfrac{c_4}{\theta_4} < \dfrac{c_2}{\theta_2}$.

---

[7] Also, we can replace the perfectly competitive insurance market by a public regulator.

According to Lemma 1, in terms of cost per unit of success probability, the ranking, in ascending order, is Protocol 1, Protocol 3, Protocol 4, and Protocol 2. Now, $\theta_3 = \theta_4 > \theta_2$, so both in terms of success probability and cost per unit of success probability, Protocols 2 and 4 are dominated by Protocol 3. In other words, Protocols 2 and 4 are less efficient than Protocol 3.[8]

Proposition 1 *In the first best, the consumer pays a premium P^*, and receives no treatment if her loss is lower than ℓ^*, Protocol 1 if her loss is between ℓ^* and ℓ^{**}, and Protocol 3 if her loss is higher than ℓ^{**}, where $\ell^* \equiv U'(Y-P^*)\dfrac{c_1}{\theta_1} < U'(Y-P^*)\dfrac{c_2}{\theta_2} \equiv \ell^{**}$. The premium is given by $P^* = c_1[1-F(\ell^*)]+(1-\theta_1)c_2[1-F(\ell^{**})]$.*

Proposition 1 presents two principles in the first best. First, the consumer is risk averse, so financial risks due to illness will be borne by the insurer. The premium is the consumer's only payment. Second, the consumer should never take the more effective and more costly treatment before trying the less effective and cheaper treatment first. By Lemma 1, Protocols 2 and 4 are inefficient, so they are never used. If the illness loss is very low, less than ℓ^*, it is not cost effective to use any treatment because the benefit $\ell\theta_i$ is lower than the cost adjusted by the marginal utility of income λc_i, where λ, the Lagrangean multiplier, equals $U'(Y - P^*)$.

When the illness loss is higher than ℓ^*, treatment should be used. Protocol 1 yields a net benefit of $\ell\theta_1 - \lambda c_1$, while Protocol 3 yields $\ell\theta_3 - \lambda c_3$. The difference between the net benefit of Protocols 1 and 3 is $(\theta_3 - \theta_1)\ell - \lambda(c_3 - c_1)$. Protocol 3 is simply Protocol 1 with Treatment 2 as an option, so the incremental success probability $\theta_3 - \theta_1$ is just θ_2, and the incremental cost $c_3 - c_1$ is c_2. The incremental benefit $(\theta_3 - \theta_1)\ell$ is increasing in ℓ, so when ℓ is higher than ℓ^{**}, the more expensive Protocol 3 is cost effective, while for ℓ lower than ℓ^{**}, the less expensive Protocol 1 is cost effective.

---

[8] We briefly comment on the case when the Cost Convexity assumption is violated. In that case, we have $\dfrac{c_1}{\theta_1} > \dfrac{c_2}{\theta_2}$. The ranking of cost per unit of success probability becomes $\dfrac{c_2}{\theta_2} < \dfrac{c_4}{\theta_4} < \dfrac{c_3}{\theta_3} < \dfrac{c_1}{\theta_1}$, so that Protocols 3 and 1 will be inefficient. Proposition 1 will be modified: Protocol 2 will be used for intermediate values of ℓ, while Protocol 4 for high values.

3 Altruistic physician and delegation

Suppose now the consumer's illness loss is not observed by the insurer. Although treatments prescribed by the physician are verifiable *ex post*, they are *ex ante* noncontractible. The physician will observe the illness loss and be delegated to make the treatment decision. In the delegation regime, an insurance company establishes a payment contract with the physician, and an insurance contract with the consumer. The physician's decision on a treatment plan can be interpreted as his practice style.

The insurance contract for the consumer consists of a premium P. We focus on physician payment and delegation, so we assume that the patient does not bear any financial risks *ex post.* In fact, this is what the first best prescribes. The payment contract for the physician is a two-part tariff, (S, T), where S is the physician's share of the incurred treatment cost, and T is a lump-sum or capitation payment.

The physician is risk neutral, and partially altruistic to the consumer. The physician learns about the consumer's illness loss ℓ after the payment contract has been accepted. This is a natural assumption in an insurance model because at the time the insurer offers contracts, the consumer is not yet sick. When the physician treats the consumer with Protocol i, his expected payoff consists of profit and the consumer's utility: $T - Sc_i + \alpha[U(Y-P) - \ell + \theta_i \ell]$. Both S and T are nonnegative, but we do not restrict S from being less than 1. The profit from using Protocol i is $T - Sc_i$; he receives the transfer T, and bears a cost Sc_i, with the balance of the cost paid for by the insurer. The parameter α measures the strength of the consumer's utility in the physician's preferences.

The altruism parameter α is a random variable, drawn on a strictly positive support $[\underline{\alpha}, \overline{\alpha}]$, with distribution and density functions, respectively, $G(\alpha)$ and $g(\alpha) > 0$. We assume that the hazard rate $\dfrac{G(\alpha)}{g(\alpha)}$ is increasing in α. The physician knows α, and this is his private information. We use the term a type-α physician for a physician with altruism parameter α. We assume that F and G are independent.

A higher value of α indicates a physician who cares more about the patient's welfare. The strength of the physician's trade-off between profit and patient utility is captured by the altruism parameter α. In making a decision based on this trade-off, the physician must respect an *ex ante* nonnegative profit constraint.

As in other agency models, we include a reservation utility constraint. If the altruistic physician does not accept the contract, he does not earn any profit, but does not treat the patient either, so his utility is $\alpha\int_0^{\overline{\ell}}[U(Y)-\ell]\mathrm{d}F(\ell)$, which is defined to be his reservation utility.

To better understand the equilibria when the physician's degree of altruism is unknown, we first show in the next subsection that the first best can be implemented when the physician's degree of altruism is known.

3.1 Known altruism

In this subsection, we assume that the altruism parameter α is common knowledge. The physician, with known altruism parameter α, is paid a lump-sum $T(\alpha)$, and bears a cost $S(\alpha)c_i$ when he uses Protocol i for the patient. Subject to the payment scheme, the physician makes treatment decisions for the patient.[9]

Suppose that, on observing the illness loss ℓ, the physician uses treatment Protocol i with probability $\tau_i(\ell)$. His expected utility is

$$\int_0^{\overline{\ell}}\left\{T(\alpha)-S(\alpha)\sum\nolimits_{i=1}^4\tau_i(\ell)c_i+\alpha[U(Y-P)-\ell+\sum\nolimits_{i=1}^4\tau_i(\ell)\theta_i\ell]\right\}\,\mathrm{d}F(\ell). \tag{4}$$

He chooses τ_i to maximize (4) subject to a nonnegative expected profit constraint

$$\int_0^{\overline{\ell}}\left\{T(\alpha)-S(\alpha)\sum\nolimits_{i=1}^4\tau_i(\ell)c_i\right\}\,\mathrm{d}F(\ell)\geq 0, \tag{5}$$

and a participation constraint:

$$\int_0^{\overline{\ell}}\left\{T(\alpha)-S(\alpha)\sum\nolimits_{i=1}^4\tau_i(\ell)c_i+\alpha[U(Y-P)-\ell+\sum\nolimits_{i=1}^4\tau_i(\ell)\theta_i\ell]\right\}\,\mathrm{d}F(\ell)\geq\alpha\int_0^{\overline{\ell}}[U(Y)-\ell]\mathrm{d}F(\ell)$$

which says that the utility from accepting the contract is higher than from refusing it. Now, the participation constraint never binds. Rewrite it as

$$\begin{gathered}\int_0^{\overline{\ell}}\left\{T(\alpha)-S(\alpha)\sum\nolimits_{i=1}^4\tau_i(\ell)c_i\right\}\,\mathrm{d}F(\ell)\geq\\ \alpha\int_0^{\overline{\ell}}[U(Y)-\ell]\mathrm{d}F(\ell)-\alpha\int_0^{\overline{\ell}}\left\{[U(Y-P)-\ell+\sum\nolimits_{i=1}^4\tau_i(\ell)\theta_i\ell]\right\}\,\mathrm{d}F(\ell).\end{gathered}$$

---

[9] Our result is consistent with the general principle that the first best is implemented when the agent only acquires private information after contracting. We do restrict contracts to be two-part tariffs. Furthermore, the agent is partially altruistic, which is not the case in the literature.

The right-hand side of this inequality is the patient's loss from the lack of insurance. Due to a competitive insurance market, this loss is never positive, so, in fact, minimum profit implies participation. In the games with asymmetric information, given the minimum profit constraint, the participation constraint remains slack for each type of the altruistic physician, so from now on, we will ignore it.

For a payment scheme that implements the first best for each $\alpha \in [\underline{\alpha}, \overline{\alpha}]$ we set

$$S(\alpha) \equiv \lambda\alpha \tag{6}$$

$$T(\alpha) \equiv \lambda\alpha \left[\int_{\ell^*}^{\ell^{**}} c_1 \, \mathrm{d}F(\ell) + \int_{\ell^{**}}^{\overline{\ell}} c_3 \, \mathrm{d}F(\ell) \right], \tag{7}$$

where $\lambda = U'(Y - P^*)$, and P^*, ℓ^*, and ℓ^{**} are the first-best premium and threshold loss levels defined in Proposition 1.

Lemma 2 *Given $S(\alpha)$ and $T(\alpha)$ defined in (6) and (7), the delegation scheme implements the first best.*

The cost share $S(\alpha) \equiv \lambda\alpha$ in Lemma 2 makes the physician internalize the consumer's treatment cost and benefit. The physician is partially altruistic, and values the patient's benefit according to $\alpha\theta_i\ell$. To align his preferences with the first best, he should be made to bear the cost at $\lambda\alpha c_i$, where λ, the marginal utility of income at first best, adjusts for the difference in the measurement between benefits (in utility) and cost (in money). This is exactly what $S(\alpha) \equiv \lambda\alpha$ does. Under this cost share, the physician's expected utility in (4) becomes

$$\begin{aligned} & \int_0^{\overline{\ell}} \left\{ T(\alpha) - S(\alpha) \sum\nolimits_{i=1}^4 \tau_i(\ell) c_i + \alpha[U(Y-P) - \ell + \sum\nolimits_{i=1}^4 \tau_i(\ell)\theta_i\ell] \right\} \mathrm{d}F(\ell) \\ = & \int_0^{\overline{\ell}} \left\{ \alpha \left[\sum\nolimits_{i=1}^4 \tau_i(\ell)\{\theta_i\ell - \lambda c_i\} \right] + T(\alpha) + \alpha[U(Y-P) - \ell] \right\} \mathrm{d}F(\ell), \end{aligned}$$

so the term inside the big square brackets is the benefit less cost. The transfer $T(\alpha)$ ensures that the physician makes a zero expected profit. The higher the value of α, the more altruistic is the physician, and he bears more costs *ex post* but receives a larger *ex ante* transfer.

The physician's behavior for the maximization of (4) subject to (5) assumes that he chooses the treatment protocols at the time of contract acceptance and before he observes the illness severity. This assumption is only made for convenience. When $S(\alpha)$ and $T(\alpha)$ are given by (6) and (7), the physician can also make the treatment decision *after* he observes ℓ. The treatment decisions will be exactly the same. Commitment is not an issue when altruism is known.

Finally, when Protocol 3 is executed, the physician will find it optimal to use Treatment 2 when Treatment 1 fails. When $\ell > \ell^{**}$ and Treatment 1 fails, the physician's decision to continue with Treatment 2 gives him a utility $\alpha\theta_2\ell - \alpha\lambda c_2 > 0$ by Proposition 1, so it is better than refusing to provide Treatment 2. In other words, the physician's decision is fully time consistent.

3.2 Unknown altruism

In this subsection, we study delegation games with unknown altruism. Equilibria hinge on whether the physician can commit to a predetermined treatment plan. We first present the game with commitment; the game without commitment follows. In each game, the insurer chooses contracts to maximize the consumer's expected utility, so we assume that the consumer will accept the contract. We have assumed that the consumer does not know the physician's type. One might wonder if there would be an incentive for a consumer to seek out a more altruistic physician, and we will discuss this issue in Section 4.

3.2.1 Equilibria in delegation with treatment plan commitment

We show the first best can be implemented by a single contract when the physician can commit to a treatment plan made at the point of contract acceptance. The extensive form of the game has four stages.

Stage 1: An insurer offers an insurance contract to the consumer and a payment contract to the physician.

Stage 2: Nature draws α from the distribution G. The physician learns α.

Stage 3: The physician decides whether to accept the payment contract, and the consumer decides whether to accept the insurance contract. The game ends if either party refuses to accept; otherwise, the physician also decides on how he will prescribe treatment protocols depending on illness loss.

Stage 4: Nature draws ℓ from the distribution F. The physician learns ℓ, and carries out treatment protocols according to the prescription rule decided in Stage 3. The physician will be paid according to the payment contract.

When altruism is unknown, type-α physician will mimic another type if the full menu of contracts defined in the regime of known altruism is offered. From Lemma 2, if a type-α physician selects $(S(\alpha), T(\alpha))$, he will choose the first-best treatment protocols and break even. However, the type-α physician can do better by exaggerating α and choosing a contract meant for type-α', $\alpha' > \alpha$. Under $(S(\alpha'), T(\alpha'))$, he can still implement the first best and break even, but will gain by being slightly less generous than offering first-best treatments. This deviation will result in a second-order loss in the consumer's expected utility but a first-order gain in the profit because $T(\alpha') > T(\alpha)$.

Our next result shows that, surprisingly, each type of physician can still be made to implement the first best even when the full menu of contracts defined in the regime of known altruism fails to do so. This is achieved by a very simple payment contract, namely $(S(\underline{\alpha}), T(\underline{\alpha}))$, defined in (6) and (7). This contract is designed as if the physician is the least altruistic type $\underline{\alpha}$.

A type-α physician's best response against $(S(\underline{\alpha}), T(\underline{\alpha}))$ is to select $\tau_i(\ell)$ to maximize

$$\int_0^{\bar{\ell}} \left\{ T(\underline{\alpha}) - S(\underline{\alpha}) \sum\nolimits_{i=1}^{4} \tau_i(\ell) c_i + \alpha [U(Y-P) - \mathrm{E}(\ell) + \sum\nolimits_{i=1}^{4} \tau_i(\ell) \theta_i \ell] \right\} \, \mathrm{d}F(\ell) \tag{8}$$

subject to

$$\int_0^{\bar{\ell}} \left\{ T(\underline{\alpha}) - S(\underline{\alpha}) \sum\nolimits_{i=1}^{4} \tau_i(\ell) c_i \right\} \, \mathrm{d}F(\ell) \geq 0. \tag{9}$$

A type-α physician's choice of treatment decision in Stage 3 is made contingent on possible illness loss. Anticipating that he will follow this treatment plan after observing the illness severity in Stage 4, the physician decides whether to accept the payment contract.

Lemma 3 *When given contract $(S(\underline{\alpha}), T(\underline{\alpha}))$ defined in (6) and (7), a type-α physician, $\alpha > \underline{\alpha}$, chooses the first-best treatment thresholds l^* and l^{**}.*

Lemma 3 reports a surprising result. Principal-agent models often show that information asymmetry generates information rent for the agent and results in distortion. Here, the physician's private information about both α and ℓ can be circumvented. Under the payment contract $(S(\underline{\alpha}), T(\underline{\alpha}))$, the best response of the type-$\underline{\alpha}$ physician is the first-best treatment protocol. His incentives have been aligned with the first best.

Now consider a more altruistic, type-α physician. He cares more about the consumer's utility than type-$\underline{\alpha}$, so he would like to be more generous, offering Protocol 1 at $\ell < \ell^*$, and Protocol 3 at $\ell < \ell^{**}$. Indeed, the first-order derivative of (8) with respect to τ_i is $\alpha\theta_i\left[\ell - \lambda\frac{\underline{\alpha}}{\alpha}\frac{c_i}{\theta_i}\right]$, which is greater than $\alpha\theta_i\left[\ell - \lambda\frac{c_i}{\theta_i}\right]$, the corresponding first-order derivative in the first best. The capitation payment, however, is $T(\underline{\alpha})$, which only compensates for the cost share $S(\underline{\alpha})$ when treatments are at the first best. The type-α suffers a loss if he follows a treatment plan more generous than the first best. The binding nonnegative profit constraint therefore stops the type-α physician from being more generous than a type-$\underline{\alpha}$ physician. Since he is able to commit to a treatment plan, it is a best response for the type-α physician to accept the contract $(S(\underline{\alpha}), T(\underline{\alpha}))$, and to implement the first best.[10] To summarize, we present (proof omitted)

Proposition 2 *In the equilibrium under delegation with treatment plan commitment, the insurer offers a single payment contract* $(S(\underline{\alpha}), T(\underline{\alpha}))$. *In equilibrium, each physician type accepts the contract and delivers the first-best treatment protocols to the consumer.*

The key to the first-best result stems from the requirement that treatment plans are made when the nonnegative expected profit consideration is still relevant. We could consider an alternative extensive form where the physician decides on treatment plans after he has accepted a contract, but before he observes ℓ (but fully anticipating that he will). This kind of commitment has no bite, and the equilibrium will be exactly the same as if commitment were impossible (as in the next subsection). This is because once the contract $(S(\underline{\alpha}), T(\underline{\alpha}))$ has been accepted, the treatment plan decision will be determined *only* by the cost share $S(\underline{\alpha})$ while the transfer $T(\underline{\alpha})$ is already received.

Proposition 2 highlights the social value of treatment plan commitment. If the physician determines his treatment plan when accepting the payment contract and sticks to it, the insurer can successfully induce all types of physicians to carry out the efficient treatment plan. Therefore, the unwarranted cost variation due to physicians' heterogeneous preferences can be reduced.

---

[10] We have assumed that the altruism parameter is in a strictly positive support $[\underline{\alpha}, \overline{\alpha}]$. If the support of α includes 0 (so that $\underline{\alpha} = 0$), our result will be modified slightly. Here, the first best can be approximated. Setting the payment contract at $(S(\alpha), T(\alpha))$, where $\alpha > 0$ and is arbitrarily close to 0, will implement the first best for all physician types higher than α. However, the contract $(S(0), T(0))$ will *not* implement the first best for any physician type.

The next subsection discusses the scenario when the physician lacks the ability to commit to a predetermined treatment plan.

3.2.2 Equilibria in delegation without treatment plan commitment

The first two stages of the game without treatment plan commitment are the same as game with commitment, except that a payment contract is now a menu. The last two stages are as follows:

Stage 3: The physician decides whether to accept the payment menu, and the consumer decides whether to accept the insurance contract. The game ends if either party refuses to accept; otherwise, the physician picks an item from the menu.

Stage 4: Nature draws ℓ from the distribution F. The physician learns ℓ, and decides on treatment protocols. The physician will be paid according to the payment contract that he has selected in Stage 3.

The key difference between games with and without treatment plan commitment is the timing of treatment decisions. Under delegation with treatment plan commitment, in Stage 3 the physician formulates a treatment protocol for each illness loss to be observed later. In Stage 4, he simply follows the plan decided earlier. Under delegation without treatment plan commitment, the physician makes his treatment decision *after* he has accepted the contract. In other words, in the game without treatment plan commitment, the physician makes the contract acceptance decision and treatment decisions sequentially. By contrast, in the game with treatment plan commitment, he makes the two decisions simultaneously.

Clearly, the single contract $(S(\underline{\alpha}), T(\underline{\alpha}))$ can no longer implement the first best. Anticipating using treatment plans more generous than the first best, physician types more altruistic than $\underline{\alpha}$ will reject this contract. This is because the transfer $T(\underline{\alpha})$ is so low that they cannot break even.

We derive the menu of optimal contracts by examining the physician's treatment protocol decisions in Stage 4. Suppose that a type-α physician has accepted a payment contract $(S(\alpha'), T(\alpha'))$ in Stage 3, and learns that the consumer's illness loss is ℓ. His decision is only affected by the cost-share parameter $S(\alpha')$, not the transfer $T(\alpha')$. Given ℓ and $S(\alpha')$, his payoff from choosing Protocol i with probability

τ_i is $-S(\alpha')\sum_{i=1}^{4}\tau_i c_i + \alpha\left[U(Y-P) - \ell + \sum_{i=1}^{4}\tau_i\theta_i\ell\right]$. The first-order derivative with respect to τ_i is $\alpha\theta_i\ell - S(\alpha')c_i$. As in the earlier analysis, the equilibrium treatment is characterized by two thresholds $\widehat{l}(\alpha';\alpha)$ and $\widehat{\widehat{l}}(\alpha';\alpha)$. The physician will never use the inefficient protocols. A consumer with ℓ smaller than $\widehat{l}(\alpha';\alpha)$ receives no treatment; ℓ between $\widehat{l}(\alpha';\alpha)$ and $\widehat{\widehat{l}}(\alpha';\alpha)$, Protocol 1; ℓ larger than $\widehat{\widehat{l}}(\alpha';\alpha)$, Protocol 3. The equilibrium in Stage 4 is completely characterized by the thresholds

$$\widehat{l}(\alpha';\alpha) = \frac{S(\alpha')}{\alpha}\frac{c_1}{\theta_1} \quad \text{and} \quad \widehat{\widehat{l}}(\alpha';\alpha) = \frac{S(\alpha')}{\alpha}\frac{c_2}{\theta_2}. \tag{10}$$

To save on notation, we write $\widehat{l}(\alpha;\alpha)$ and $\widehat{\widehat{l}}(\alpha;\alpha)$ as $\widehat{l}(\alpha)$ and $\widehat{\widehat{l}}(\alpha)$, respectively.

In contrast to delegation with treatment plan commitment, the equilibrium treatment decisions are to be made without any reference to the nonnegative profit requirement. In Stage 4, the physician does not have the option of rejecting a payment contract. The requirement of making a nonnegative expected profit has no bite here.

Next we study the physician's equilibrium choice of a payment contract in Stage 3. Suppose that the menu $\{(S(\alpha), T(\alpha))\}$ has been offered to the physician in Stage 2. We use a generalized version of the revelation principle (Myerson 1982). Define a type-α physician's expected payoff from selecting contract $(S(\alpha'), T(\alpha'))$ and the thresholds l' and l'' by

$$\begin{aligned} V(\alpha', l', l''; \alpha) \quad \equiv \quad & T(\alpha') - S(\alpha')\left[\int_{l'}^{l''} c_1 \mathrm{d}F(l) + \int_{l''}^{\bar{l}} c_3 \mathrm{d}F(l)\right] + \\ & \alpha\left[U(Y-P) - \mathrm{E}(l) + \int_{l'}^{l''} \theta_1 l \mathrm{d}F(l) + \int_{l''}^{\bar{l}} \theta_3 l \mathrm{d}F(l)\right]. \end{aligned}$$

We consider equilibria in which a type-α physician selects contract $(S(\alpha), T(\alpha))$, and adopts the thresholds $\widehat{l}(\alpha)$ and $\widehat{\widehat{l}}(\alpha)$. Clearly, for any choice of $(S(\alpha'), T(\alpha'))$ the thresholds that maximize V are $\widehat{l}(\alpha', \alpha)$ and $\widehat{\widehat{l}}(\alpha';\alpha)$, as in the continuation equilibrium (10). A menu of contracts is said to be incentive compatible if $V(\alpha, \widehat{l}(\alpha), \widehat{\widehat{l}}(\alpha); \alpha) \geq V(\alpha', l', l''; \alpha)$ for all α' and α, and all l' and l''. Given a menu $(S(\alpha), T(\alpha))$, define the type-α physician's maximum payoff by $W(\alpha) \equiv \max\limits_{\alpha', l', l''} V(\alpha', l', l''; \alpha)$.

Lemma 4 *A menu of contracts $\{(S(\alpha), T(\alpha))\}$, $\alpha \in [\underline{\alpha}, \overline{\alpha}]$, is incentive compatible only if W is convex,*

$$W'(\alpha) = U(Y-P) - E(l) + \int_{\widehat{l}(\alpha)}^{\widehat{\widehat{l}}(\alpha)} \theta_1 l dF(l) + \int_{\widehat{\widehat{l}}(\alpha)}^{\bar{l}} \theta_3 l dF(l), \tag{11}$$

and both $\widehat{l}(\alpha)$ and $\widehat{\widehat{l}}(\alpha)$ are decreasing in α.

According to Lemma 4, incentive compatibility requires that the physician's equilibrium utility be convex in the altruism parameter. The physician's equilibrium payoff, $W(\alpha)$, must rise at an increasing rate.[11] Furthermore, it says that the equilibrium thresholds must be decreasing so that a more altruistic physician prescribes more treatments. We write the continuation equilibrium condition (10) as

$$\widehat{l}(\alpha) = \frac{S(\alpha)}{\alpha}\frac{c_1}{\theta_1} \quad \text{and} \quad \widehat{\widehat{l}}(\alpha) = \frac{S(\alpha)}{\alpha}\frac{c_2}{\theta_2}, \tag{12}$$

so incentive compatibility requires the cost share to altruism parameter ratio, $S(\alpha)/\alpha$, be decreasing.

Next, we analyze the physician's nonnegative profit constraint. By selecting $(S(\alpha), T(\alpha))$, a type-α physician's expected profit is

$$\pi(\alpha) \equiv T(\alpha) - S(\alpha)\left[\int_{\widehat{l}(\alpha)}^{\widehat{\widehat{l}}(\alpha)} c_1 \mathrm{d}F(l) + \int_{\widehat{\widehat{l}}(\alpha)}^{\bar{l}} c_3 \mathrm{d}F(l)\right].$$

Substituting this expression into $W(\alpha) = V(\alpha, \widehat{l}(\alpha), \widehat{\widehat{l}}(\alpha); \alpha)$, we have $W(\alpha) = \pi(\alpha) + \alpha\ W'(\alpha)$, or

$$\pi(\alpha) = W(\alpha) - \alpha W'(\alpha). \tag{13}$$

Differentiating both sides of this equation, we have $\pi'(\alpha) = -\alpha W''(\alpha)$. The convexity of $W(\alpha)$ implies that $\pi(\alpha)$ is decreasing. The physician's nonnegative profit constraints are therefore simplified to $\pi(\overline{\alpha}) \geq 0$. In other words, if the most altruistic physician breaks even, so do all other physician types.

Lemma 5 *Incentive compatibility is equivalent to $S(\alpha)/\alpha$ being decreasing, and hence $\widehat{l}(\alpha)$ and $\widehat{\widehat{l}}(\alpha)$ decreasing. Nonnegative expected profit for the physician is equivalent to $\pi(\overline{\alpha}) \geq 0$.*

We continue with the derivation of the equilibrium contract menu. The insurer must break even given the continuation equilibrium after Stage 1. The total expected expenditure by the insurer equals the expected profit and treatment cost, averaged over all physician types. Hence, the premium P satisfies

$$P = \int_{\underline{\alpha}}^{\overline{\alpha}} \pi(\alpha)\mathrm{d}G(\alpha) + \int_{\underline{\alpha}}^{\overline{\alpha}} \left[\int_{\widehat{l}(\alpha)}^{\widehat{\widehat{l}}(\alpha)} c_1 \mathrm{d}F(l) + \int_{\widehat{\widehat{l}}(\alpha)}^{\bar{l}} c_3 \mathrm{d}F(l)\right] \mathrm{d}G(\alpha). \tag{14}$$

---

[11] Because U is a utility function of income, its sign can be positive or negative; hence, W and W' can be positive or negative. Indeed, signs of W and W' are irrelevant for incentive compatibility.

From $W(\alpha) \equiv W(\overline{\alpha}) - \int_{\alpha}^{\overline{\alpha}} W'(x)\mathrm{d}x$, we can substitute for W in the expression for π in (13):

$$\pi(\alpha) = W(\overline{\alpha}) - \int_{\alpha}^{\overline{\alpha}} W'(x)dx - \alpha W'(\alpha).$$

Then we use (11) in Lemma 4 to replace $W'(x)$. After integration by parts, we can substitute for $\pi(\alpha)$ and express (14) as

$$\begin{aligned} P \quad = \quad & \int_{\underline{\alpha}}^{\overline{\alpha}} \left[\int_{\widehat{l}(\alpha)}^{\widehat{\widehat{l}}(\alpha)} c_1 \mathrm{d}F(l) + \int_{\widehat{\widehat{l}}(\alpha)}^{\bar{l}} c_3 \mathrm{d}F(l) \right] \mathrm{d}G(\alpha) + W(\overline{\alpha}) \\ & - \int_{\underline{\alpha}}^{\overline{\alpha}} \left\{ \left(\frac{G(\alpha)}{g(\alpha)} + \alpha \right) \left(U(Y-P) - \mathrm{E}(l) + \int_{\widehat{l}(\alpha)}^{\widehat{\widehat{l}}(\alpha)} \theta_1 l \mathrm{d}F(l) + \int_{\widehat{\widehat{l}}(\alpha)}^{\bar{l}} \theta_3 l \mathrm{d}F(l) \right) \right\} \mathrm{d}G(\alpha). \qquad (15) \end{aligned}$$

The premium for the patient includes treatment costs and the physician's utility, which consists of the base utility $W(\overline{\alpha})$ less the consumer's utility multiplied by the physician's altruism parameter adjusted by the hazard rate $(G(\alpha)/g(\alpha) + \alpha)$.

From (13), we have

$$\begin{aligned} \pi(\overline{\alpha}) \quad & = \quad W(\overline{\alpha}) - \overline{\alpha} W'(\overline{\alpha}) \\ & = \quad W(\overline{\alpha}) - \overline{\alpha} \left[U(Y-P) - \mathrm{E}(l) + \int_{\widehat{l}(\overline{\alpha})}^{\widehat{\widehat{l}}(\overline{\alpha})} \theta_1 l \mathrm{d}F(l) + \int_{\widehat{\widehat{l}}(\overline{\alpha})}^{\bar{l}} \theta_3 l \mathrm{d}F(l) \right], \end{aligned}$$

so $\pi(\overline{\alpha}) \geq 0$ if and only if

$$W(\overline{\alpha}) \geq \overline{\alpha} \left[U(Y-P) - \mathrm{E}(l) + \int_{\widehat{l}(\overline{\alpha})}^{\widehat{\widehat{l}}(\overline{\alpha})} \theta_1 l \mathrm{d}F(l) + \int_{\widehat{\widehat{l}}(\overline{\alpha})}^{\bar{l}} \theta_3 l \mathrm{d}F(l) \right]. \qquad (16)$$

The equilibrium in Stage 4 also requires (12), which says that $\widehat{l}(\alpha)$ and $\widehat{\widehat{l}}(\alpha)$ follow a fixed ratio; this will be shown to be satisfied, so we will ignore this requirement for now.

The equilibrium allocation implemented by the insurer is the solution to the following program: choose P, $W(\overline{\alpha})$, $\widehat{l}(\alpha)$, and $\widehat{\widehat{l}}(\alpha)$ to maximize the consumer's expected utility

$$U(Y-P) - \mathrm{E}(l) + \int_{\underline{\alpha}}^{\overline{\alpha}} \left(\int_{\widehat{l}(\alpha)}^{\widehat{\widehat{l}}(\alpha)} \theta_1 l \mathrm{d}F(l) + \int_{\widehat{\widehat{l}}(\alpha)}^{\bar{l}} \theta_3 l \mathrm{d}F(l) \right) \mathrm{d}G(\alpha)$$

subject to the breakeven constraint (15), the physician nonnegative profit constraint (16), and $\widehat{l}(\alpha)$, $\widehat{\widehat{l}}(\alpha)$ both decreasing. Let μ denote the multiplier for the insurer's breakeven constraint (15). We present the characterization of the solution:

Proposition 3 *Under treatment plan noncommitment, the equilibrium thresholds and premium,* $\widehat{l}(\alpha)$*,* $\widehat{\widehat{l}}(\alpha)$*, and* P *are given by*

$$\begin{aligned} \widehat{l}(\alpha) &= \mu\frac{c_1}{\theta_1}\left[1+\left(\frac{G(\alpha)}{g(\alpha)}+\alpha\right)\mu\right]^{-1} && (17)\\ \widehat{\widehat{l}}(\alpha) &= \mu\frac{c_2}{\theta_2}\left[1+\left(\frac{G(\alpha)}{g(\alpha)}+\alpha\right)\mu\right]^{-1} && (18)\\ U'(Y-P) &= \mu. && (19) \end{aligned}$$

The type-$\overline{\alpha}$ *physician earns zero profit, and* $W(\overline{\alpha})$ *is given by (16) as an equality; all other physician types earn strictly positive profits.*

From the equilibrium thresholds in Proposition 3 and equation (12), we can find the cost share and transfer functions for the implementation. The cost share function is

$$S(\alpha) = \alpha\mu\left[1+\mu\left(\frac{G(\alpha)}{g(\alpha)}+\alpha\right)\right]^{-1} \tag{20}$$

and the transfer function is

$$T(\alpha) = W(\alpha) - \alpha W'(\alpha) + S(\alpha)\left[\int_{\widehat{l}(\alpha)}^{\widehat{\widehat{l}}(\alpha)} c_1 dF(l) + \int_{\widehat{\widehat{l}}(\alpha)}^{\bar{l}} c_3 dF(l)\right], \tag{21}$$

where $W'(\alpha)$ is determined by equation (11) and $W(\alpha)$ is obtained by integrating $W'(\alpha)$. The physician's implementation of Protocol 3 is time consistent. If $\ell > \widehat{\widehat{l}}(\alpha)$, his utility from continuing with Treatment 2 for the consumer is $\alpha\ell\theta_2 - S(\alpha)c_2$. From (20), this is $\alpha\ell\theta_2 - \alpha\mu\left[1+\mu\left(\frac{G(\alpha)}{g(\alpha)}+\alpha\right)\right]^{-1}c_2$, which is strictly positive by Proposition 3.

The determination of the equilibrium thresholds includes the term $\dfrac{G(\alpha)}{g(\alpha)}+\alpha$, and this is the key difference from the first best in Proposition 1. The first-best thresholds are determined by a straightforward cost-effectiveness principle. This has to be modified due to the missing information about the physician's degree of altruism. From (6) and (7) in Lemma 2 for the implementation of the first best when altruism is known, the cost share and transfer should increase proportionally with respect to α if α were known. This creates an incentive problem when α is unknown. A less altruistic physician benefits by claiming to be more altruistic, obtaining a higher transfer and cutting back on treatments.

The cost share and transfer in Proposition 3 must increase less than proportionally with α to deter a physician from exaggerating his degree of altruism. More altruistic physicians provide more treatments at the cost of receiving less transfers. The equilibrium cost shares and transfers involve the hazard rate, $\dfrac{G(\alpha)}{g(\alpha)}$, a standard, Myerson "virtual" adjustment due to private information. Furthermore, treatment benefits are valued by physicians, so the adjustment also includes the term α in addition to the virtual component.

Because the physician's profit is passed on to consumers, we have the following corollary:

Corollary 1 *The equilibrium premium P is higher than the first best premium P^*.*

The comparison between equilibrium thresholds in Proposition 3 and the first best is not straightforward. The first best is independent of the distribution of α, but the functions $\widehat{l}$ and $\widehat{\widehat{l}}$ have ranges that depend on the distribution as well as the support of α. One expects that for low values of α, equilibrium thresholds will be higher than first best, while for high values of α, they will be lower. That is, less altruistic physicians provide treatments less than the first best, and the opposite for more altruistic physicians. We can verify this with an example. Let the utility function be $U(Y) = \ln Y$, so $U'(Y) = 1/Y$. Suppose that l is uniformly distributed on [0, 1] while α is uniformly distributed on $[\underline{\alpha}, \underline{\alpha} + 1]$, $\underline{\alpha} > 0$. By Proposition 1, the first-best thresholds are

$$l^* = \frac{1}{Y - P^*}\frac{c_1}{\theta_1} \quad \text{and} \quad l^{**} = \frac{1}{Y - P^*}\frac{c_2}{\theta_2}. \tag{22}$$

Equation (19) reduces to $\mu = 1/(Y - P)$. The equilibrium thresholds in Proposition 3 are

$$\widehat{l}(\alpha) = \frac{1}{Y - P}\frac{c_1}{\theta_1}\left[1 + \frac{2\alpha - \underline{\alpha}}{Y - P}\right]^{-1} \quad \text{and} \quad \widehat{\widehat{l}}(\alpha) = \frac{1}{Y - P}\frac{c_2}{\theta_2}\left[1 + \frac{2\alpha - \underline{\alpha}}{Y - P}\right]^{-1}. \tag{23}$$

By Corollary 1, the premium P is larger than the first-best premium P^*. From (22) and (23), $\widehat{l}(\alpha)$ and $\widehat{\widehat{l}}(\alpha)$ are larger than the first-best thresholds for $\alpha < \dfrac{P - P^* + \underline{\alpha}}{2}$, and are smaller than or equal to the first-best thresholds otherwise. If the difference $P - P^*$ is between $\underline{\alpha}$ and $\underline{\alpha} + 2$, there exists a type-$\widetilde{\alpha}$ physician delivering first-best treatments. Physicians less altruistic than type-$\widetilde{\alpha}$ will provide less treatment than the first best whereas physicians more altruistic than type-$\widetilde{\alpha}$ will provide more.

Proposition 3 and Corollary 1 can be used to explain the wide variations of medical costs. Differences in physician practice styles are here captured by differences in physician altruism. The same illness will be treated differently depending on the attending physician's preferences. Such variations, however, can be avoided if altruistic physicians make treatment decisions when the full financial consequences are respected, as Proposition 2 shows.

4 Discussions and policies

4.1 More than two treatments

We have assumed that there are only two treatments available. Proposition 1 can be extended to an arbitrary number of treatments under Cost Convexity. For example, when there are three treatments, we can construct many treatment protocols by various treatment sequences. However, only three are efficient. These three are (i) use Treatment 1 only; (ii) use Treatment 1, and if it fails use Treatment 2; and (iii) begin with Treatment 1, if it fails use Treatment 2, and if that also fails, use Treatment 3. The intuition behind the inefficiency of protocols other than those in (i), (ii),and (iii) mimics that in Lemma 1. For example, the protocol of Treatment 2 and then Treatment 3 upon failure of Treatment 2 is dominated by the protocol in (iii). Adding Treatment 1 before Treatment 2 raises the total probability of success, and reduces the expected cost due to Cost Convexity, because Treatment 1 has the lowest cost-success probability ratio. Retaining the two-treatment assumption saves on notation, while relaxing it would not lead to qualitatively new results.

4.2 Searching for altruistic physicians

In Proposition 3, a physician provides more treatments when he is more altruistic. Therefore, *ex post*, a consumer prefers to be treated by a more altruistic physician. In Lemma 5, a physician reveals his type by selecting an item from the full cost-share-transfer menu. Typically, however, consumers may not be aware of the financial arrangement between the insurer and the physician, so a physician's altruism information may not be inferred.

In repeated interactions, without treatment plan commitment, consumers' incentive to search must exist. In our setup, after an initial treatment episode, if a consumer knows the illness severity, then she can update her belief about the physician's altruism. For example, suppose that the severity is moderate, but the physician does not recommend Treatment 2 after Treatment 1 has failed. Then the consumer will infer that the physician is not very altruistic.

Searching for more altruistic physicians is irrelevant when treatment plan commitment is possible. In the first-best equilibrium in Proposition 2, all physician types provide the same treatment. When search is relevant, it is associated with inefficiency and higher premium due to the lack of commitment in Proposition 3. Search exacerbates inefficiency. To attract consumers, physicians may offer more treatments even when their own degree of altruism is low. Clustering of consumers among altruistic physicians may likely increase the premium, too. A policy implication is that inefficient search can be avoided if treatment plan commitment is possible.

4.3 Selecting physicians

Physicians earn profits when there is a lack of treatment plan commitment. A way to limit profit is to reject some physician types. We have assumed that all types in $[\underline{\alpha}, \overline{\alpha}]$ must earn nonnegative profits. It is possible to relax this by allowing the insurer to retain only those with α between $\underline{\alpha}$ and $\overline{\alpha}' < \overline{\alpha}$. This can be implemented by reducing the transfer function T in (21) (say, by a constant). Those physicians with α larger than $\overline{\alpha}'$ will not accept any contract. All those who accept will make less profits, and the distortion can be reduced.

The cost of rejecting highly altruistic physician types comes in the form of rationing. We have considered contracts for one consumer and one physician. We implicitly have assumed that the aggregate supply equals aggregate demand. Rejecting some physician types reduces the physician supply. Even in a competitive insurance market, the premium may have to increase; otherwise, nonprice rationing results.

4.4 Comparison with classical moral hazard

In the classical moral hazard model, the consumer makes all the treatment decisions, and will be made to bear partial treatment costs, but is free to choose the order of treatments. We think that this will perform poorly. First, as we have argued before, this scenario is unrealistic. Consumers currently must rely on physicians for treatments; the complexity of modern medicine rules out independent consumer decisions. Second, consumer cost sharing may lead to more inefficient decisions. For example, if the consumer is to bear 20% of treatment costs, she may find Protocols 2 and 4 attractive, even though these are socially inefficient. In addition, this will impose financial risks to patients. An equilibrium in the classical, consumer cost-share model performs worse than the equilibrium in Proposition 2. Treatment plan commitment performs better than mitigating moral hazard through consumer cost sharing.

5 Concluding remarks

We study how an insurer can reduce the unnecessary cost due to practice-style variations by designing payment contracts for heterogenous physicians. Our model consists of two elements that are missing in the classical optimal insurance model. Treatments can be combined, and physicians are altruistic, with different degrees of altruism. We develop new principles from this setup. First, we show that the first-best treatment plan follows a conservative pattern. Second, we consider delegating treatment decisions to physicians, and show that the first best can be implemented only when a physician can commit to treatment plans at the time of contract acceptance. We offer various policy implications.

Treatment plans involve a time dimension, and it is natural that commitment plays a key role in the analysis. The physician committing to using particular plans may result in time-inconsistent decisions. But such commitment has social value; it reduces premium and inefficient search.

The treatment technology is richer than the usual, static health care quantity approach. This lets us rule out some treatment combinations as inefficient. However, our main results for delegation under treatment plan commitment and noncommitment should hold without any modification if the physician is choosing a quantity of services.

We acknowledge that our model abstracts from learning. Two issues naturally arise when learning is important. First, the likelihood of treatment success may itself be uncertain. A first treatment is often an experimentation for the physician to learn about treatment efficacy. The failure of a treatment may then update the likelihood that other treatments may be successful. Second, illness severity may be uncertain. A first treatment may reveal that the illness is more or less severe than initially thought. This new information will impact subsequent treatments. These issues are for further research.

Appendix

Proof of Lemma 1: Let $k_1 \equiv \frac{c_1}{\theta_1}$and $k_2 \equiv \frac{c_2}{\theta_2}$. By Assumption 1, $k_1 < k_2$. From the definitions of θ_3 and c_3, we substitute c_1 and c_2 by $k_1\theta_1$ and $k_2\theta_2$, respectively, and obtain

$$\frac{c_3}{\theta_3} = \left[\frac{\theta_1}{\theta_1 + (1-\theta_1)\theta_2}\right] k_1 + \left[\frac{(1-\theta_1)\theta_2}{\theta_1 + (1-\theta_1)\theta_2}\right] k_2,$$

which is a weighted average of k_1 and k_2, so $\frac{c_1}{\theta_1} < \frac{c_3}{\theta_3} < \frac{c_2}{\theta_2}$.

Because $\theta_3 = \theta_4$ and $c_3 < c_4$ by Assumption 1, we have $\frac{c_3}{\theta_3} < \frac{c_4}{\theta_4}$. It remains to show that $\frac{c_4}{\theta_4} < \frac{c_2}{\theta_2}$. By Assumption 1, $c_1\theta_2 < c_2\theta_1$. To both sides of this inequality we multiply by $(1-\theta_2)$ and then add $c_2\theta_2$. This results in $(c_2 + (1-\theta_2)c_1)\theta_2 < c_2(\theta_2 + (1-\theta_2)\theta_1)$. Since $c_4 = c_2 + (1-\theta_2)c_1$ and $\theta_4 = \theta_2 + (1-\theta_2)\theta_1$, we have $c_4\theta_2 < c_2\theta_4$, so $\frac{c_4}{\theta_4} < \frac{c_2}{\theta_2}$. ■

Proof of Proposition 1: Omit the boundary conditions. Use pointwise optimization, and form the Lagrangian for ℓ:

$$L = \int_0^{\bar{\ell}} [U(Y-P) - \ell + \sum_{i=1}^4 \tau_i(\ell)\theta_i\ell] \mathrm{d}F(\ell) + \lambda \left(P - \int_0^{\bar{\ell}} \sum_{i=1}^4 \tau_i(\ell)c_i \mathrm{d}F(\ell)\right)$$

where $\lambda > 0$ is the multiplier of the premium constraint. The first-order derivatives are

$$\frac{\partial L}{\partial P} = -U'(Y-P) + \lambda \tag{24}$$

$$\frac{\partial L}{\partial \tau_i} = f(\ell)\left(\theta_i\ell - \lambda c_i\right) = f(\ell)\theta_i\left(\ell - \lambda\frac{c_i}{\theta_i}\right), \quad i = 1,2,3,4. \tag{25}$$

The derivatives in (25) are independent of τ_i, so at each ℓ, the Protocol with the highest positive value of $\frac{\partial L}{\partial \tau_i}$ among $i = 1,2,3,4$ will be used. If all the derivatives are negative, then no treatment will be used.

First, $\tau_2(\ell) = \tau_4(\ell) = 0$ for all ℓ; the consumer never uses Protocols 2 and 4. Because $\frac{c_3}{\theta_3} < \frac{c_2}{\theta_2}$ by Lemma 1 and $\theta_2 < \theta_3$, $\frac{\partial L}{\partial \tau_2} < \frac{\partial L}{\partial \tau_3}$, $\forall \ell$. Therefore, we must have $\tau_2(\ell) = 0$, $\forall \ell$. Because $\frac{c_3}{\theta_3} < \frac{c_4}{\theta_4}$ by Lemma 1 and $\theta_3 = \theta_4$, $\frac{\partial L}{\partial \tau_4} < \frac{\partial L}{\partial \tau_3}$, $\forall \ell$. Therefore, we must have $\tau_4(\ell) = 0$, $\forall \ell$.

By Lemma 1, when $\ell < \lambda\frac{c_1}{\theta_1}$, the first-order derivatives $\frac{\partial L}{\partial \tau_i}$ are all negative. Define $\ell^* \equiv \lambda\frac{c_1}{\theta_1}$. From Lemma 1, when $\ell < \ell^*$, $\tau_i(\ell) = 0$, $i = 1,2,3,4$. Hence, the consumer does not use any treatment when $\ell < \ell^*$.

Next, from (25), we have

$$\begin{aligned}
&\frac{\partial L}{\partial \tau_3} - \frac{\partial L}{\partial \tau_1} \\
= \quad & \left[(\theta_3 - \theta_1)\ell - \lambda(c_3 - c_1)\right] f(l) \\
= \quad & (1-\theta_1)[\theta_2 \ell - \lambda c_2] f(l) \qquad (26)
\end{aligned}$$

Now define $\ell^{**} \equiv \dfrac{\lambda c_2}{\theta_2}$. (Because we assume that $\overline{\ell}$ is sufficiently large, we have $\ell^{**} < \overline{\ell}$, and it is well-defined.) The expression in (26) is positive if and only if $\ell > \ell^{**}$. Both $\dfrac{\partial L}{\partial \tau_3}$ and $\dfrac{\partial L}{\partial \tau_1}$are positive when $\ell > \ell^{*}$. Together, we have $\tau_1(\ell) = 1$ when $\ell^* \le \ell < \ell^{**}$, and $\tau_3(\ell) = 1$ when $\ell^{**} < \ell < \overline{\ell}$.

Setting the first-order derivative (24) to 0, we have $\lambda = U'(Y-P)$, so the values of ℓ^* and ℓ^{**} are those in the Proposition. Finally, the premium P^* is $[F(\ell^{**}) - F(\ell^*)]c_1 + [1 - F(\ell^{**})]c_3$, which simplifies to

$$P^* = c_1[1 - F(\ell^*)] + (1-\theta_1)c_2[1 - F(\ell^{**})]. \qquad (27)$$

There is a unique solution for P between 0 and Y. Let $g(P)$ denote the right-hand side of (27), where ℓ^* and ℓ^{**} are now regarded as functions of P. Since $U'(Y-P)$ increases in P, ℓ^* and ℓ^{**} increase in P. The function $g(P)$ is decreasing in P. The function $g(P)$ reaches the maximum at $P = 0$, and $g(0) = c_1\left[1 - F(\dfrac{U'(Y)c_1}{\theta_1})\right] + (1-\theta_1)c_2\left[1 - F(\dfrac{U'(Y)c_2}{\theta_2})\right] > 0$. The function $g(P)$ reaches the minimum at $P = Y$, and $g(Y) = 0$ because $U'(0) = +\infty$. We conclude that there is a unique solution for $P = g(P)$. ■

Proof of Lemma 2: First, given the contract $(S(\alpha), T(\alpha))$, the type-α physician chooses treatment protocols $\tau_i(\ell)$, $i = 1,2,3,4$, to maximize his expected utility

$$\int_0^{\overline{\ell}} \left\{ T(\alpha) - S(\alpha)\sum\nolimits_{i=1}^4 \tau_i(\ell)c_i + \alpha[U(Y-P) - \ell + \sum\nolimits_{i=1}^4 \tau_i(\ell)\theta_i\ell] \right\} \, \mathrm{d}F(\ell). \qquad (28)$$

The first-order derivative of (28) with respect to $\tau_i(\ell)$ is

$$\begin{aligned}
& \alpha f(l)\theta_i \left[\ell - \frac{S(\alpha)c_i}{\alpha\theta_i}\right] \\
= \quad & \alpha f(l)\theta_i \left(\ell - \frac{\lambda c_i}{\theta_i}\right) \qquad (29)
\end{aligned}$$

upon substitution $S(\alpha)$ by $\lambda\alpha$. The first-order derivative (29) is the first-order derivative (25) for the first best multiplied by α, a constant. We conclude that the type-α physician's optimal treatment decision is first best.

Given the contract, the physician's expected profit from his optimal, first-best treatment decision is

$$\begin{aligned} & T(\alpha) - S(\alpha)\left[\int_{\ell^*}^{\ell^{**}} c_1 \,\mathrm{d}F(\ell) + \int_{\ell^{**}}^{\bar{\ell}} c_3 \,\mathrm{d}F(\ell)\right] \\ = \; & \lambda\alpha\left[\int_{\ell^*}^{\ell^{**}} c_1 \,\mathrm{d}F(\ell) + \int_{\ell^{**}}^{\bar{\ell}} c_3 \,\mathrm{d}F(\ell)\right] - \lambda\alpha\left[\int_{\ell^*}^{\ell^{**}} c_1 \,\mathrm{d}F(\ell) + \int_{\ell^{**}}^{\bar{\ell}} c_3 \,\mathrm{d}F(\ell)\right] = 0, \end{aligned}$$

so constraint (5) is satisfied.

It remains to show that the insurer breaks even. The insurer receives the first-best premium P^* from the consumer. He pays the physician the transfer $T(\alpha)$, and $1 - S(\alpha)$ share of the cost to the physician. The insurer's expected profit is therefore

$$\begin{aligned} & P^* - T(\alpha) - (1 - S(\alpha))\left[\int_{\ell^*}^{\ell^{**}} c_1 \,\mathrm{d}F(\ell) + \int_{\ell^{**}}^{\bar{\ell}} c_3 \,\mathrm{d}F(\ell)\right] \\ = \; & P^* - \left[\int_{\ell^*}^{\ell^{**}} c_1 \,\mathrm{d}F(\ell) + \int_{\ell^{**}}^{\bar{\ell}} c_3 \,\mathrm{d}F(\ell)\right] - \left\{T(\alpha) - S(\alpha)\left[\int_{\ell^*}^{\ell^{**}} c_1 \,\mathrm{d}F(\ell) + \int_{\ell^{**}}^{\bar{\ell}} c_3 \,\mathrm{d}F(\ell)\right]\right\}. \quad (30) \end{aligned}$$

The insurer breaks even in the first-best contract, so $P^* - \left[\int_{\ell^*}^{\ell^{**}} c_1 \,\mathrm{d}F(\ell) + \int_{\ell^{**}}^{\bar{\ell}} c_3 \,\mathrm{d}F(\ell)\right] = 0$. The term inside the big curly brackets in (30) is the physician's profit and has been shown be to zero. Hence, the insurer makes zero expected profit. ■

Proof of Lemma 3: The Lagrangian for the constraint optimization program maximizing (8) subject to (9) is

$$L = \int_0^{\bar{\ell}} \left\{(1+\varphi)\left[T(\underline{\alpha}) - S(\underline{\alpha})\sum\nolimits_{i=1}^4 \tau_i(\ell)c_i\right] + \alpha\left(U(Y-P) - \mathrm{E}(\ell) + \sum\nolimits_{i=1}^4 \tau_i(\ell)\theta_i\ell\right)\right\} \mathrm{d}F(\ell),$$

where $\varphi \geq 0$ is the multiplier for the nonnegative expected profit constraint. From pointwise optimization, the first-order derivative with respect to $\tau_i(\ell)$ is:

$$\frac{\partial L}{\partial \tau_i} = \alpha f(l)\theta_i\left[l - \frac{\lambda\underline{\alpha}(1+\varphi)c_i}{\alpha\theta_i}\right]. \quad (31)$$

after substitution by $S(\underline{\alpha}) = \lambda\underline{\alpha}$. Define $l^{'} \equiv \lambda\frac{c_1}{\theta_1}\left[\frac{\underline{\alpha}(1+\varphi)}{\alpha}\right]$ and $l^{''} = \lambda\frac{c_2}{\theta_2}\left[\frac{\underline{\alpha}(1+\varphi)}{\alpha}\right]$. From the proof of Proposition 1, the physician will not prescribe any treatment if $l \leq l^{'}$, will use treatment Protocol 1 if $l^{'} < l < l^{''}$ and treatment Protocol 3 for $l^{''} < l$.

Next, we show that the Lagrangian multiplier φ must equal $\frac{\alpha}{\underline{\alpha}} - 1$ for a type-α physician. When $\varphi = \frac{\alpha}{\underline{\alpha}} - 1$, the loss thresholds $l^{'}$ and $l^{''}$ are identical to the first-best levels, l^* and l^{**}, respectively, so the first best is optimal. It remains to show that $\varphi = \frac{\alpha}{\underline{\alpha}} - 1$, and we do that by contradiction.

Suppose that $\varphi < \frac{\alpha}{\underline{\alpha}} - 1$. Then the loss thresholds satisfy $l^{'} < l^*$ and $l^{''} < l^{**}$. The difference between the physician's expected profit from choosing thresholds $l^{'}$ and $l^{''}$ and that from choosing the first-best thresholds l^* and l^{**} is

$$\begin{aligned} & \left\{T(\underline{\alpha}) - S(\underline{\alpha})\left[\int_{\ell'}^{\ell''} c_1 \,\mathrm{d}F(\ell) + \int_{\ell''}^{\bar{\ell}} c_3 \,\mathrm{d}F(\ell)\right]\right\} - \left\{T(\underline{\alpha}) - S(\underline{\alpha})\left[\int_{\ell^*}^{\ell^{**}} c_1 \,\mathrm{d}F(\ell) + \int_{\ell^{**}}^{\bar{\ell}} c_3 \,\mathrm{d}F(\ell)\right]\right\} \\ = \quad & -S(\underline{\alpha})\left\{\int_{\ell''}^{\ell^{**}} (1-\theta_1)c_2 \,\mathrm{d}F(\ell) + \int_{\ell'}^{\ell^*} c_1 \,\mathrm{d}F(\ell)\right\} < 0. \end{aligned}$$

Given that under $(S(\underline{\alpha}), T(\underline{\alpha}))$ the expected profit from the first-best treatments (the second term, in curly brackets, on the first line) is 0, the physician's expected profit from choosing thresholds $l^{'}$ and $l^{''}$is negative. This violates the nonnegative expected profit constraint, and contradicts the assumption that $\varphi < \frac{\alpha}{\underline{\alpha}} - 1$. Hence, we conclude that $\varphi \geq \frac{\alpha}{\underline{\alpha}} - 1$.

Next, suppose that $\varphi > \frac{\alpha}{\underline{\alpha}} - 1$. Then the loss thresholds satisfy $l^{'} > l^*$ and $l^{''} > l^{**}$. The difference between the physician's expected profit from choosing thresholds $l^{'}$ and $l^{''}$ and that from choosing the first-best thresholds is

$$\begin{aligned} & \left\{T(\underline{\alpha}) - S(\underline{\alpha})\left[\int_{\ell'}^{\ell''} c_1 \,\mathrm{d}F(\ell) + \int_{\ell''}^{\bar{\ell}} c_3 \,\mathrm{d}F(\ell)\right]\right\} - \left\{T(\underline{\alpha}) - S(\underline{\alpha})\left[\int_{\ell^*}^{\ell^{**}} c_1 \,\mathrm{d}F(\ell) + \int_{\ell^{**}}^{\bar{\ell}} c_3 \,\mathrm{d}F(\ell)\right]\right\} \\ = \quad & S(\underline{\alpha})\left\{\int_{\ell^{**}}^{\ell''} (1-\theta_1)c_2 \,\mathrm{d}F(\ell) + \int_{\ell^*}^{\ell'} c_1 \,\mathrm{d}F(\ell)\right\} > 0. \end{aligned}$$

Again, given that under $(S(\underline{\alpha}), T(\underline{\alpha}))$ the expected profit from the first-best treatments is 0, the physician earns a strictly positive expected profit. Hence, the nonnegative expected profit constraint does not bind, and the multiplier φ must be zero. This contradicts the assumption that $\varphi > \frac{\alpha}{\underline{\alpha}} - 1 > 0$. Hence we conclude that $\varphi \leq \frac{\alpha}{\underline{\alpha}} - 1$. In sum, we have $\varphi = \frac{\alpha}{\underline{\alpha}} - 1$. ■

Proof of Lemma 4: Because $W(\alpha)$ is the upper bound of affine functions of α, it is convex (Rockafellar, 1972, Theorem 5.5), and therefore almost everywhere differentiable (Rockafellar, 1972, Theorem 25.5). Incentive compatibility implies $V(\alpha, \widehat{l}(\alpha), \widehat{\widehat{l}}(\alpha); \alpha) = W(\alpha)$. By the envelope theorem,

$$W'(\alpha) = \frac{\partial V}{\partial \alpha} = U(Y-P) - \mathrm{E}(l) + \int_{\widehat{l}(\alpha)}^{\widehat{\widehat{l}}(\alpha)} \theta_1 l \mathrm{d}F(l) + \int_{\widehat{\widehat{l}}(\alpha)}^{\bar{l}} \theta_3 l \mathrm{d}F(l),$$

with the partial derivative being evaluated at $\alpha' = \alpha$, $\ell' = \widehat{l}(\alpha)$, and $\ell'' = \widehat{\widehat{l}}(\alpha)$, and we obtain the expression in the Lemma.

Next, rewrite $W'(\alpha)$ as

$$U(Y-P)-\mathrm{E}(l)+\int_{\widehat{l}(\alpha)}^{\bar{\ell}}\theta_1 l\mathrm{d}F(l)+\int_{\widehat{\widehat{l}}(\alpha)}^{\bar{l}}(1-\theta_1)\theta_2 l\mathrm{d}F(l). \tag{32}$$

Because $\dfrac{\mathrm{d}\widehat{l}(\alpha)}{\mathrm{d}\alpha}=\dfrac{\mathrm{d}(\frac{S(\alpha)}{\alpha})}{\mathrm{d}\alpha}\dfrac{c_1}{\theta_1}$and $\dfrac{\mathrm{d}\widehat{\widehat{l}}(\alpha)}{\mathrm{d}\alpha}=\dfrac{\mathrm{d}(\frac{S(\alpha)}{\alpha})}{\mathrm{d}\alpha}\dfrac{c_2}{\theta_2}$, $\dfrac{\mathrm{d}\widehat{l}(\alpha)}{\mathrm{d}\alpha}$ and $\dfrac{\mathrm{d}\widehat{\widehat{l}}(\alpha)}{\mathrm{d}\alpha}$ share the same sign as that of $\dfrac{\mathrm{d}(\frac{S(\alpha)}{\alpha})}{\mathrm{d}\alpha}$. If $\widehat{l}(\alpha)$ and $\widehat{\widehat{l}}(\alpha)$ were increasing at some α, then from (32) $W'(\alpha)$ would be decreasing at α. This contradicts incentive compatibility. We conclude that $\widehat{l}(\alpha)$ and $\widehat{\widehat{l}}(\alpha)$ must be decreasing. ■

Proof of Lemma 5: We only need to show that any menu satisfying $S(\alpha)/\alpha$ decreasing and $\pi(\overline{\alpha})\geq 0$ implies incentive compatibility and nonnegative expected profit. We start with a given cost-share rule $S(\alpha)$ with $S(\alpha)/\alpha$ decreasing. From Lemma 4 and the equilibrium condition for Stage 4 in (12), we have the thresholds $\widehat{l}(\alpha)$ and $\widehat{\widehat{l}}(\alpha)$ being decreasing. We can construct $T(\alpha)$ so that $(S(\alpha),T(\alpha))$, $\alpha\in[\underline{\alpha},\overline{\alpha}]$, is incentive compatible. First, we set $\widehat{l}(\alpha)$ and $\widehat{\widehat{l}}(\alpha)$ by (12) for the continuation equilibrium in Stage 4. Second, we use (11) to construct a function $W'(\alpha)$. Setting a value for $W(\overline{\alpha})$, we integrate $W'(\alpha)$ to obtain $W(\alpha)$. Third, we set

$$T(\alpha)=W(\alpha)-\alpha W'(\alpha)+S(\alpha)\left[\int_{\widehat{l}(\alpha)}^{\widehat{\widehat{l}}(\alpha)}c_1\mathrm{d}F(l)+\int_{\widehat{\widehat{l}}(\alpha)}^{\bar{l}}c_3\mathrm{d}F(l)\right]. \tag{33}$$

It is straightforward to check that $S(\alpha)$ and the $T(\alpha)$ in (33) satisfy incentive compatibility. Finally, we can choose $W(\overline{\alpha})$ so that $\pi(\overline{\alpha})\geq 0$. ■

Proof of Proposition 3: From the Lagrangian function L, where μ and γ are the multipliers for constraint (15) and (16), respectively.

$$\begin{aligned}L \;=\;& U(Y-P)-E(l)+\int_{\widehat{l}(\alpha)}^{\widehat{\widehat{l}}(\alpha)}\theta_1 l dF(l)+\int_{\widehat{\widehat{l}}(\alpha)}^{\bar{l}}\theta_3 l dF(l)+\mu\left\{P-W(\overline{\alpha})-\int_{\widehat{l}(\alpha)}^{\widehat{\widehat{l}}(\alpha)}c_1 dF(l)-\int_{\widehat{\widehat{l}}(\alpha)}^{\bar{l}}c_3 dF(l)\right\}\\&+\mu\left(\frac{G(\alpha)}{g(\alpha)}+\alpha\right)\left(U(Y-P)-E(l)+\int_{\widehat{l}(\alpha)}^{\widehat{\widehat{l}}(\alpha)}\theta_1 l dF(l)+\int_{\widehat{\widehat{l}}(\alpha)}^{\bar{l}}\theta_3 l dF(l)\right)\\&+\gamma\left\{W(\overline{\alpha})-\overline{\alpha}\left[U(Y-P)-E(l)+\int_{\widehat{l}(\overline{\alpha})}^{\widehat{\widehat{l}}(\overline{\alpha})}\theta_1 l dF(l)+\int_{\widehat{\widehat{l}}(\overline{\alpha})}^{\bar{l}}\theta_3 l dF(l)\right]\right\}.\end{aligned}$$

We use pointwise optimization for $\widehat{l}(\alpha)$, $\widehat{\widehat{l}}(\alpha)$, and take the derivatives of the Lagrangian function with respect to them at α. To simplify, we drop constant terms in the derivatives $\dfrac{\partial L}{\partial\widehat{l}}$ and $\dfrac{\partial L}{\partial\widehat{\widehat{l}}}$. These (simplified) derivatives are in expressions (34) - (37). The derivatives of the Lagrangian function with respect to P and $W(\overline{\alpha})$ are in (38) and (39).

$$
\begin{aligned}
\frac{\partial L}{\partial \widehat{l}}_{\alpha<\overline{\alpha}} &= -\theta_1\widehat{l}+\mu c_1-\mu\left(\frac{G(\alpha)}{g(\alpha)}+\alpha\right)\theta_1\widehat{l} && (34)\\
\frac{\partial L}{\partial \widehat{l}}_{\alpha=\overline{\alpha}} &= -\theta_1\widehat{l}+\mu c_1-\mu\left(\frac{G(\overline{\alpha})}{g(\overline{\alpha})}+\overline{\alpha}\right)\theta_1\widehat{l}+\gamma\overline{\alpha}\theta_1\widehat{l} && (35)\\
\frac{\partial L}{\partial \widehat{\widehat{l}}}_{\alpha<\overline{\alpha}} &= -\theta_2\widehat{\widehat{l}}+\mu c_2-\mu\left(\frac{G(\alpha)}{g(\alpha)}+\alpha\right)\theta_2\widehat{\widehat{l}} && (36)\\
\frac{\partial L}{\partial \widehat{\widehat{l}}}_{\alpha=\overline{\alpha}} &= -\theta_2\widehat{\widehat{l}}+\mu c_2-\mu\left(\frac{G(\overline{\alpha})}{g(\overline{\alpha})}+\overline{\alpha}\right)\theta_2\widehat{\widehat{l}}+\gamma\overline{\alpha}\theta_2\widehat{\widehat{l}} && (37)\\
\frac{\partial L}{\partial P} &= -U'(Y-P)+\mu\left[1-U'(Y-P)\int_{\underline{\alpha}}^{\overline{\alpha}}\left(\frac{G(\alpha)}{g(\alpha)}+\alpha\right)\mathrm{d}G(\alpha)\right]+\gamma\overline{\alpha}U'(Y-P) && (38)\\
\frac{\partial L}{\partial W(\overline{\alpha})} &= -\mu+\gamma && (39)
\end{aligned}
$$

We obtain (17) and (18) in the Proposition by setting (34), (36) to zero. From (39), we have $\mu=\gamma$. We then substitute γ by μ in and (38), set it to zero, and then apply integration by parts to obtain (19).

The first-order conditions for $\widehat{l}$ and $\widehat{\widehat{l}}$ at $\alpha=\overline{\alpha}$ are

$$
\begin{aligned}
\widehat{l}(\overline{\alpha}) &= \mu\frac{c_1}{\theta_1}\left[1+\frac{G(\overline{\alpha})}{g(\overline{\alpha})}\mu\right]^{-1} && (40)\\
\widehat{\widehat{l}}(\overline{\alpha}) &= \mu\frac{c_2}{\theta_2}\left[1+\frac{G(\overline{\alpha})}{g(\overline{\alpha})}\mu\right]^{-1}. && (41)
\end{aligned}
$$

The limit of $\widehat{l}(\alpha)$ as α converges to $\overline{\alpha}$ from below is $\frac{c_1}{\theta_1}\left[\frac{1}{\mu}+\left(\frac{G(\overline{\alpha})}{g(\overline{\alpha})}+\overline{\alpha}\right)\right]^{-1}$. Clearly, $\lim_{\alpha\to\overline{\alpha}}\widehat{l}(\alpha)<\widehat{l}(\overline{\alpha})$. Because incentive compatibility requires $\widehat{l}(\alpha)$ to be decreasing, the monotonicity constraint must bind at $\widehat{l}(\overline{\alpha})$, so $\widehat{l}(\overline{\alpha})=\lim_{\alpha\to\overline{\alpha}}\widehat{l}(\alpha)$. By the same argument, we have $\widehat{\widehat{l}}(\overline{\alpha})=\lim_{\alpha\to\overline{\alpha}}\widehat{\widehat{l}}(\alpha)$.

By assumption, the hazard rate $\frac{G(\alpha)}{g(\alpha)}$ is increasing, so $\frac{G(\alpha)}{g(\alpha)}+\alpha$ is increasing. Hence, $\widehat{l}(\alpha)$ and $\widehat{\widehat{l}}(\alpha)$ are decreasing in α. Finally, from (17) and (18), the ratio of $\widehat{l}(\alpha)$ to $\widehat{\widehat{l}}(\alpha)$ is a constant, so the equilibrium condition in Stage 4, (12), is satisfied. ■

Proof of Corollary 1: Suppose $P\leq P^*$. Then

$$\frac{1}{U'(Y-P)}\geq\frac{1}{U'(Y-P^*)}=\frac{c_1}{\theta_1}\frac{1}{\ell^*},$$

where the equality follows from Proposition 1. From (19), we have

$$\frac{1}{\mu}=\frac{1}{U'(Y-P)}\geq\frac{1}{U'(Y-P^*)}=\frac{c_1}{\theta_1}\frac{1}{\ell^*},$$

so

$$\frac{1}{\mu} \geq \frac{c_1}{\theta_1}\frac{1}{\ell^*}. \tag{42}$$

By (17), we have

$$\begin{aligned}\frac{c_1}{\theta_1}\frac{1}{\widehat{\ell}(\alpha)} &= \frac{1}{\mu}\left[1+\left(\frac{G(\alpha)}{g(\alpha)}+\alpha\right)\mu\right] \\ &\geq \frac{c_1}{\theta_1}\frac{1}{\ell^*}\left[1+\left(\frac{G(\alpha)}{g(\alpha)}+\alpha\right)\mu\right] > \frac{c_1}{\theta_1}\frac{1}{\ell^*}, \qquad (43)\end{aligned}$$

where the weak inequality is due to (42), and the strict inequality follows from the term inside the square brackets of (43) being strictly positive. Therefore, $\widehat{\ell}(\alpha) < \ell^*$ for all α. Repeating the same argument, we have $\widehat{\widehat{\ell}}(\alpha) < \ell^{**}$ for all α. The consumer receives more treatments and the physician receives profits. This therefore implies that $P > P^*$, which is a contradiction. ■

References

[1] Akerlof, G. and Kranton, R., "Identity and the economics of organizations," Journal of Economic Perspectives, 19 (2005), 9–32.

[2] Arrow, K., "Uncertainty and the welfare economics of medical care," American Economic Review, 53 (1963), 941–73.

[3] Bénabou, R. and Tirole, J., "Intrinsic and extrinsic motivation," Review of Economic Studies, 70 (2003), 489–520.

[4] Besley, T. and Ghatak, M., "Competition and incentives with motivated agents," American Economic Review, 95 (2005), 616–636.

[5] Chalkley, M. and Malcomson, J., "Contracting for health services when patient demand does not reflect quality," Journal of Health Economics, 17 (1998), 1-19.

[6] Chernew, M., Encinosa, W. and Hirth, R., "Optimal health insurance: the case of observable, severe illness," Journal of Health Economics, 19 (2000), 585-609.

[7] Choné, P. and Ma, C., "Optimal health care contract under physician agency," forthcoming in Annales d'Économie et de Statistique (2010).

[8] Delfgaauw, J. and Dur, R., "Signaling and screening of workers' motivation," Journal of Economic Behavior & Organization, 62 (2007), 605-624.

[9] Delfgaauw, J. and Dur, R., "Incentives and workers' motivation in the public sector," Economic Journal, Royal Economic Society, 118 (2008), 171-191.

[10] Dranove, D. and Spier, K., "A Theory of Utilization Review," Contributions to Economic Analysis and Policy, 2 (2003), article 9.

[11] Dusheiko, M., Gravelle, H., Jacobs, R. and Smith, P., "The effect of budgets on gatekeeping doctor behavior: evidence from a natural experiment," Journal of Health Economics, 25 (2006), 449-478.

[12] Ellis, R. and McGuire, T., "Provider behavior under prospective reimbursement cost sharing and supply," Journal of Health Economics, 5 (1986), 129-151.

[13] Ellis, R. and McGuire, T., "Optimal payment systems for health services," Journal of Health Economics, 9 (1990), 375-396.

[14] Francois, P. "Public service motivation as an argument for government provision," Journal of Public Economics, 78 (2000), 275–99.

[15] Jack, W., "Purchasing health care services from providers with unknown altruism," Journal of Health Economics, 24 (2005), 73–93

[16] Léger, P., "Physician payment mechanisms," Chapter 6 in Financing Health Care: New Ideas for a Changing Society (Ed: Lu, M. and E. Jonsson) (Wiley-VCH Press), (2008), 149-176.

[17] Ma, C., "Incentius de cost i qualitat en l'assisténcia sanitária. Proveidors altruistes," in Guillem Lopez-Casasnovas, La Contractació de Serveis Sanitaris, Generalitat de Catalunya, (1998), 65-80 (English version "Cost and Quality Incentives in Health Care: Altruistic Providers" available at http://people.bu.edu/ma/Papers_Archive/Q-C_ALT.pdf).

[18] Ma, C. and Riordan, M., "Health insurance, moral hazard, and managed care," Journal of Economics & Management Strategy, 11 (2002), 81-107.

[19] McGuire, T., "Physician agency," Chapter 9 in Handbook of Health Economics, 1 (2000), 461-536.

[20] Murdock, K., "Intrinsic motivation and optimal incentive contracts," RAND Journal of Economics, 33 (2002), 650–71.

[21] Myerson, R., "Optimal coordination mechanisms in generalized princiapl-agent problems," Journal of Mathematical Economics, 10 (1982), 67-81.

[22] Newhouse, J., "Toward a theory of nonprofit institutions: an economic model of a hospital," American Economic Review, 60 (1970), 64-74.

[23] Prendergast, C., “The motivation and bias of bureaucrats,” American Economic Review, 97 (2007), 180-196.

[24] Prendergast, C., “Intrinsic motivation and incentives,” American Economic Review Papers and Proceedings, 98 (2008), 201-205.

[25] Pauly, M., “The Economics of moral hazard: comment,” American Economic Review, 35 (1968), 531-37.

[26] Rockafellar, R.T., Convex Analysis, Princeton University Press, second printing (1972).

[27] Rochaix, L., “Information asymmetry and search in the market for physician services,” Journal of Health Economics, 8 (1989), 53-84.

[28] Rogerson, W., “Choice of treatment intensity by a nonprofit hospital under prospective pricing,” Journal of Economics & Management Strategy, 3 (1994), 7-51.

[29] Wennberg, E, Brownlee, S., Fisher, E., Skinner, J. and Weinstein, J., “An agenda for change: improving quality and curbing health care spending: opportunities for the congress and the Obama Administration,” Dartmouth Atlas White Paper, December 2008.

[30] Zeckhauser, R., “Medical Insurance: A case study of the tradeoff between risk spreading and appropriate incentives,” Journal of Economic Theory, 2 (1970), 10–26.