INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2004-37

Asymmetric Information from Physician Agency : Optimal Payment and Healthcare Quantity

Ph. CHONÉ[1]
Ch.-T. A. MA[2]

[1] CREST-LEI and CNRS URA 2200, 28 rue des Saints-Pères, 75007 Paris, France, email : chone@ensae.fr

[2] Department of Economics, Boston University, 270 Bay State Road, Boston, Massachusetts 02215, USA, email : ma@bu.edu

Acknowledgement : We thank Richard Frank, Yuk-fai Fong, Billy Jack, Tom, McGuire, Joe New-house and Bernard Salanié for their suggestions. We are also grateful to many seminar participants for their comments. The second author acknowledges support from the Agency for Healthcare Research and Quality through grant PO1-HS-10803. Opinions here belong to the authors.

Asymmetric Information from Physician Agency: Optimal Payment and Healthcare Quantity

Philippe Choné[†] Ching-to Albert Ma[‡]

August, 2004

Abstract

We model asymmetric information arising from physician agency, and its effect on the design of payment and healthcare quantity. The physician-patient coalition aims to maximize a combination of physician profit and patient benefit. The degree of substitution between profit and patient benefit in the physician-patient coalition is the physician's private information, as is the patient's intrinsic valuation of treatment quantity. The equilibrium mechanism depends only on the physician-patient coalition parameter. Moreover, the equilibrium mechanism exhibits extensive pooling, with prescribed quantity and payment being insensitive to the agency characteristics or patient's actual benefit. The optimal mechanism is interpreted as managed care where strict approval protocols are placed on treatments.

JEL: D82, I1, I10, L15

Keywords: Physician Agency, Altruism, Optimal Payment, Healthcare Quantity, Managed Care

[†]CREST-LEI and CNRS URA 2200, 28 rue des Saint-Pères, 75007 Paris, France, E-mail: chone@ensae.fr;
[‡]Department of Economics, Boston University, 270 Bay State Road, Boston, Massachusetts 02215, USA, E-mail: ma@bu.edu

Acknowledgement: We thank Richard Frank, Yuk-fai Fong, Billy Jack, Tom, McGuire, Joe Newhouse, and Bernard Salanié for their suggestions. We are also grateful to many seminar participants for their comments. The second author acknowledges support from the Agency for Healthcare Research and Quality through grant PO1-HS-10803. Opinions here belong to the authors.

Asymmetric Information from Physician Agency: Optimal Payment and Healthcare Quantity

Philippe Choné et Ching-to Albert Ma

Abstract

We model asymmetric information arising from physician agency, and its effect on the design of payment and healthcare quantity. The physician-patient coalition aims to maximize a combination of physician profit and patient benefit. The degree of substitution between profit and patient benefit in the physician-patient coalition is the physician's private information, as is the patient's intrinsic valuation of treatment quantity. The equilibrium mechanism depends only on the physician-patient coalition parameter. Moreover, the equilibrium mechanism exhibits extensive pooling, with prescribed quantity and payment being insensitive to the agency characteristics or patient's actual benefit. The optimal mechanism is interpreted as managed care where strict approval protocols are placed on treatments.

Keywords: Physician Agency, Altruism, Optimal Payment, Healthcare Quantity, Managed Care

Résumé

Nous examinons l'effet des interactions entre patients et médecins sur les contrats optimaux passés entre ces derniers et les organismes payeurs (assurances santé par exemple). Pour tenir compte de la complexité et de la diversité de ces interactions, nous supposons que les caractéristiques précises de la relation d'agence entre le médecin et son patient, ainsi que le bénéfice des soins pour le patient, sont des informations privées du médecin. Le schéma optimal ne dépend que du paramètre de la relation d'agence. De plus, le pooling est robuste à l'optimum : la quantité de soins et le paiement ne dépendent ni des caractéristiques de l'agence ni de celles du patients. Nous interprétons ce phénomène comme des restrictions sur les quantités imposées au médecin par le payeur, restrictions qui correspondent à la pratique connue sous le nom de *managed care*.

Mots-clés : Interaction patient-médecin, altruisme, quantité de soins, paiement optimal, *managed care*

JEL: D82, I1, I10, L15

1 Introduction

The economics of the health market is concerned with the interaction between insurers, consumers and providers. In this trilateral relationship, the nexus between a patient and a physician is arguably the most fundamental and complicated. The term "Physician Agency" has been used in the literature to refer to a range of issues arising from the influence of physician on healthcare use (McGuire, 2000). Yet, researchers have not reached a consensus on the formal model of physician agency. The reason perhaps originates from our suspicion of a pure profit maximization paradigm to model physician agency. We tend to believe that physician-patient interactions are influenced by factors such as power, motivation, medical training and current practice, ethics, and altruism. An otherwise powerful and simple paradigm, profit-maximization is heavily contaminated by these other factors.

Once economists depart from a pure profit-maximization approach, it is unclear what is the most compelling alternative. There is a practical issue, too. Useful assumptions in theoretical models cannot afford to be too complicated. The literature, both theoretical and empirical, has somehow gyrated toward a pragmatic but natural assumption: physician-patient interaction leads to physician objectives including both physician profits and patient benefits. "We assume that the physician maximizes the sum of his income and patient benefit," is used frequently. In fact, the terms "perfect agency" and "imperfect agency" are often used to mean the extent in which patient's benefit counts towards physician preferences.[1]

Physicians are often the major, if not the only, decision maker for patients' medical treatments. Any hypothesis about their objectives affects the way payment and insurance mechanisms are to be designed. For example, recognizing that a physician may value patient benefit, an insurer may either impose capitation or reduce cost reimbursement rates. A simple assumption asserting a *fixed* combination of profit and altruistic motives misses the complexity of physician agency. In this paper, we model this complexity, and examine its implications on the design of payment and quantity.

We model the complexity of physician agency by asymmetric information: how the physician agency values profit and patient benefit is the physician's private information. The previous literature has adopted a complete-information assumption, but the parameter that measures the relative

[1] Here is a very small sample of papers using some form of this assumption: Chalkley and Malcomson (1998), Dranove and Spier (2003), Dusheiko et. al. (2004), Ellis and McGuire (1986, 1990), Ma (1998), Ma and Riordan (2002), Newhouse (1970), Rochaix (1989), Rogerson (1994), etc.

weight of income and patient benefit is critical for mechanism design. This is unsatisfactory. How does an insurer obtain this information? What is more, physicians with different degrees of concern for patient benefits have an incentive to manipulate this information.[2]

Our model of asymmetric information arising from physician agency aims at understanding a fundamental issue. It is, however, much more than a theoretical curiosity. We recognize that physician agency is diverse and complex. A design of healthcare payment and quantity use should respect this complexity. For example, managed care can be interpreted as an attempt to control physician agency. Various quantity controls such as utilization reviews, quantity limits and restrictions appear to be difficult to interpret in terms of conventional marginal benefit and cost consideration. We will show how readily our model endogenizes quantity restrictions when physician agency leads to asymmetric information.

Our model consists of a consumer whose health benefits may vary, perhaps according to her medical conditions or preferences. When she seeks treatment, this information is learned by the physician and becomes his private information. The physician-patient interaction is captured by preferences that weigh the physician's profit as well as the consumer's health benefit. The preferences from this physician agency are the physician's private information. The physician possesses two pieces of private information: the consumer's benefit from treatment, as well as how the physician-patient coalition values profit and consumer benefit. The consumer's benefit and the physician agency weigh on consumer benefit follow a joint density distribution.[3]

An insurer or managed care company designs a payment and quantity contract for the physician. We will assume that the consumer has full insurance. A general mechanism will be studied. The physician, who behaves according to the preferences in the physician agency, picks from a menu of quantity-payment pairs, indexed by the consumer's intrinsic benefit and the physician agency's weight on consumer benefit. According to the revelation principle, these schedules are equivalent to a direct mechanism, where the physician reports his private information, and where it is an equilibrium for him to do so honestly. Also, the physician must earn a nonnegative profit

[2] Our methodology is not unlike the one that changed the way regulation was studied in the 1980's. Regulating a monopolist would be rather easy when cost information was known: a lump-sum payment to cover the fixed cost and marginal cost pricing would achieve the efficient allocation. It was the recognition of incomplete information about the monopolist's cost by Baron and Myerson (1982) and technology by Laffont and Tirole (1986) that led to the new regulation economics in the past twenty years.

[3] Our model belongs to the class of multi-dimensional adverse selection problems; see Armstrong and Rochet (1999), and Rochet and Choné (1998).

from treating the patient. We study the incentive-compatible mechanism which maximizes the consumer's expected benefit less the payment to the physician.

Our first result asserts that the extraction of information about the consumer's intrinsic benefit information is impossible. The optimal schedule only depends on the physician agency weight on consumer benefit, not the consumer's true benefit. The insurer will have to infer the distribution of the consumer's benefit from the information of the physician agency. The design of payments and quantities is to provide incentives for the physician agency to reveal the weight truthfully.

The second key result is that the program for the optimal quantity-payment schedule actually translates to a choice of a pooling region. In the pooling regime, the quantity is insensitive with respect to the physician agency parameter. This is an unusual step, and is seldom found among solutions for optimal mechanisms.[4] The physician's profit level turns out to be decreasing in the agency weight, while the quantity must be increasing. Profits, however, can only be positively related to quantity. The tension caused by incentive compatibility between quantities and nonnegative profits leads to the choice of pooling.

The optimal mechanism must have pooling, and pooling can even be complete. So even information about the physician agency weights on patient benefits will never be completely extracted, and may not be extracted at all. The latter situation is likely when the physician agency weight on patient is very large compared to the intrinsic patient benefit. In other words, when the discrepancy between the physician agency's weight on patient benefit and the intrinsic valuation is sufficiently large, the insurer will *not* attempt to extract the intrinsic information through agency.

Our main result on the optimal mechanism can be interpreted as a form of quantity restriction. Managed care aims to control agency by limiting physicians' discretion over healthcare quantities, and we derive this result from our model. In earlier work, this is usually taken as an assumption; see for example Baumgardner (1991). Other attempts to consider managed care propose allocation rules, which are often left as exogenous (Frank, Glazer and McGuire (2000); Keeler, Carter and Newhouse (1998)).

We compare the optimal quantities with the first best. The expected quantities are the same across the two regimes, but on average there is a reduction in the range of quantities under asymmetric information. Compared to the first best, on average the optimal mechanism assigns more

[4] We assume no countervvailing incentives, and adopt the usual hazard rate conditions for monotonicity.

quantities to patients with low valuations, and the opposite is true for those with high valuations.

We next compare the optimal mechanism with the second best, where the physician agency weight is assumed to be known. Even when the physician agency preferences are known, incentives to misreport a patient's intrinsic valuation information persist. These incentives must be removed. In contrast to the third best (where information of both patient valuation and physician agency is unavailable to the insurer), the second-best mechanism can tie quantities to intrinsic patient information. Surprisingly, there is a strong symmetry between the second best and the third best where the agency preferences information is unknown to the insurer. In the second best, there is always pooling, and it may be complete. There is also compression of quantities in the same fashion. The minimum profit constraints are common across different regimes.

Arrow (1963) pointed out the market failure due to the missing information about health status. The subsequent literature has highlighted other sources of market failures such as risk selection (Glazer and McGuire (2000)), cost and quality effort (Ma (1994)), and creaming and dumping (Ellis (1998)), etc. In general the literature has concentrated on problems of "hidden information" and "hidden action" of the provider. Our model follows the same line of investigation: the consumer's health status is unknown. Our model of the asymmetric information arising from the physician agency is novel.

Our approach to model a physician-patient relationship can be regarded as a reduced form. While Ma and McGuire (1997) explicitly model collusion between a patient and a provider, there the interaction between the patient and the physician does not involve any uncertainty relevant to the design of insurance and payments. Dranove (1988) examines bilateral asymmetric information between the physician and the patient. While this interaction is studied explicitly in Dranove's model, the design of optimal payment is not considered. Our method uses a simple structure and can be extended in various ways. Appendix A contains some examples of structural models of physician agency. In these models, the insurer is unable to observe all relevant features, which then become the source of the asymmetric information.

Recently, there has been some interest in incentive theories when agents are partly motivated by monetary rewards and partly by work activities; see Besley and Ghatak (2003), and Dixit (2002), and the references there. These papers argue that public and private firms may adopt different goals than pure profit maximization to attract workers who are more motivated by their activities. Our model can be interpreted as one where physicians are altruistic towards their patients. So physicians preferences include both monetary rewards and their patients' benefits, which correspond to their activities.

Jack (2004) considers incentives for cost and quality choices by healthcare providers with unknown altruism. In his model, Jack assumes that the provider derives utility from supplying qualites, but this utility is unknown to the regulator. The provider's choices of quality and cost effect are unobservable either. So the model contains elements of hidden information and hidden action. He adopts a participation or reservation utility constraint, and dervies the optimal mechanism. The paper shows that a menu of cost-sharing schemes is optimal. We neither use a participation constraint nor consider hidden action in this paper.[5]

Section 2 presents the model. The following section contains the characterization of incentive compatible payment and quantity schedules. Section 4 derives the main results, compares the optimal mechanism with the first best, and provides some examples. We also derive the optimal mechanism when the physician agency parameter is public information while the intrinsic patient information remains unknown; the results and comparisons are in section 5. The last section contains some concluding remarks. Proofs of results are collected in an appendix.

2 The Model

We now describe a general model of physician agency and quantity-payment design. An insurance or managed care company establishes an insurance contract with a consumer and a payment contract with a physician. If the consumer becomes sick, she seeks medical treatment from the doctor. For simplicity, and as in most managed care plans, we assume that the insurance coverage is complete; the patient does not bear any monetary expense when she seeks medical care.

Due to specific illness conditions, a consumer's severities vary, and so do her benefit from treatment. Upon diagnosis the doctor learns the consumer's conditions or her benefit from treatment. This information becomes the physician's private information; it is unknown to the managed care company.

After interacting with the patient, the doctor prescribes a treatment quantity for her. The real-valued variable $q \geq 0$ denotes the healthcare quantity. We use a real-valued parameter $\alpha > 0$ to characterize patient severity or potential benefit; this parameter varies according to a distribution.

[5] In Appendix C, we solve our model using reservation utility instead of minimum profit constraints.

For a consumer with parameter α, her benefit from quantity q is $\alpha V(q)$, where V is a strictly increasing and strictly concave function. The function V is assumed to be common knowledge while the value of α is the physician's private information.[6]

The physician bears a cost $C(q)$ when he prescribes a quantity q. The function C is strictly increasing, and strictly convex. The cost function is common knowledge, and we assume that the cost of treatment is verifiable information. This also means that the treatment quantity is verifiable. If the doctor is paid an amount R after he provides quantity q to the patient, his profit is $R - C(q)$.

We will assume that the consumer is passive. It is infeasible for the patient to order treatment quantity or bargain with the insurer directly.[7] The patient must interact with the physician to obtain treatment. This interaction is the physician agency that we now describe. After their interaction, the joint action of the physician and the patient is based on the following coalition utility: $R - C(q) + \beta V(q)$, where $\beta > 0$ captures the coalition's weight on the patient's benefit. That is, the physician, representing the patient, aims to maximize $R - C(q) + \beta V(q)$.

The important assumption is that physician agency is complex: the parameter β is unknown to the insurance company, and the physician's private information. The parameter β may be related to the patient's benefit parameter α. For example, it can be that β equals α, or some other function of α. Here, we simply assume that (α, β) follows a joint distribution function, and that this is common knowledge. Nevertheless, we mostly will work with the marginal distribution of β, $G(\beta)$; we assume that the marginal distribution of β has a strictly positive and continuous density function g on the support $[\underline{\beta}, \overline{\beta}]$. An assumption on the expectation of α conditional on β will be made later. Again, both α and β are the physician's private information.

The coalition utility may be regarded as a form of altruism. The patient delegates her treatment decision to the altruistic physician who derives utility from treatment quantities. The degree of altruism is the physician's private information.[8] Alternatively, one may regard the coalition utility as a reduced form of physician-patient interaction. The patient and physician agree on a cooperative and implicit agreement, which is captured by the coalition utility. The physician agency may involve

[6] Alternatively, we can regard $\alpha V(q)$ as the valuation of a payer or regulator which provides health care to some insured population.

[7] The consumer may not understand fully the severity, and in this case, must delegate decisions to the physician.

[8] How a patient selects a physician depending on her belief on the degree of altruism is an important issue, but beyond the scope of the paper.

complex agreements and understanding unknown to the insurer. This complexity is modelled by the physician's private information on both the patient's valuation as well as the coalition's valuation on benefits. Although we use a reduced form interpretation, we have constructed a few structural examples that generate coalition utility functions. In Appendix A, we use a model of Nash bargaining, and a model of repeated interactions between a physician and a patient to substantiate our reduced forms. From now on, however, we will mostly use the altruism interpretation for ease of exposition.

An insurer or managed care company would like to provide the efficient level of treatment to a patient. Information about the patient's benefit from treatment, namely α, is important. If the patient's preference parameter α was public information, the efficient level of treatment would be one that maximized $\alpha V(q) - C(q)$. The first-best, efficient level of quantity, $q^\star(\alpha)$ satisfying $\alpha V'(q^\star) = C'(q^\star)$, is an increasing function of α.

It is, however, the valuation of patient benefit by the physician-patient coalition or the altruistic physician that leads to an incentive problem. Suppose for the moment that the value of β is set to 0; in this case, only monetary rewards for the physician matter. Because both costs and quantities are verifiable, the managed care company can promise to reimburse the full cost of treatment: $R = C(q)$. No matter which quantity the physician recommends, he nets a zero profit, and has no incentive to misrepresent the patient's benefit parameter α.

A cost reimbursement policy works poorly for the managed care company when the physician is altruistic. The concern for the patient's benefit translates into a desire for higher quantities of treatment. For any positive β, if exaggerating the patient's benefit parameter α leads to a higher quantity, and since V is strictly increasing in q, the physician will have an incentive to do so. The first-best quantities are not implementable. Clearly, the fact that the parameter β is the physician's private information makes the problem more serious.

The managed care company designs an optimal mechanism to maximize the patient's utility less the payment to the physician, respecting the physician's private information about α and β. We use a general mechanism. Since costs and quantities are verifiable, they can be explicitly specified by the mechanism. According to the Revelation Principle, the optimal mechanism must be one in which it is an equilibrium for the physician to reveal α and β truthfully.

A mechanism is defined by the following pair of functions: $(q(\alpha, \beta), R(\alpha, \beta))$, where $q(\alpha, \beta)$ is the quantity the physician provides and $R(\alpha, \beta)$ his payment when he reports α and β. We will let the mechanism be deterministic. That is, for each report, the determination of the quantity and payment does not involve a lottery. We believe that stochastic mechanisms are suboptimal. The convexity of C implies that the total expected cost is higher than the cost of the expected quantity if q is determined via a lottery. Likewise, the concavity of V implies that the expected utility is lower than the utility of the expected quantity. A stochastic mechanism either raises costs or reduces benefits, but has no impact on incentives since the physician's preferences exhibit no risk aversion.

We impose a minimum profit condition on the payment: $R(\alpha, \beta) \geq C(q(\alpha, \beta))$; physicians must not suffer a financial loss when they provide treatment. This is a natural requirement in a physician-patient bargaining situation because a physician has a refusal option. If the concern for patient benefit arises out of altruism, the physician should nevertheless insist on a minimum profit. Even an altruistic physician cannot afford to be reimbursed less than his cost consistently. A reservation utility actually puts no limit on the monetary transfers between the insurer and the physician, and this seems unreasonable.[9] We call $R - C(q)$ the physician's profit.

A mechanism is said to be *incentive compatible* if

$$(1) \qquad R(\alpha, \beta) - C(q(\alpha, \beta)) + \beta V(q(\alpha, \beta)) \geq R(\alpha', \beta') - C(q(\alpha', \beta')) + \beta V(q(\alpha', \beta')),$$

for all $\alpha, \alpha', \beta, \beta'$. Obviously, an incentive compatible mechanism induces truth telling. A mechanism is said to satisfy *minimum profit* if

$$(2) \qquad R(\alpha, \beta) - C(q(\alpha, \beta)) \geq 0,$$

for all α, β. We say that a mechanism is *admissible* if it is incentive compatible and satisfies minimum profit.

How is the optimal mechanism to be designed? There are two pieces of information unavailable to the insurance company: both α and β are the physicians's private information. One would expect that the physician will earn some information rent because of asymmetric information. We, however, allow for an altruistic consideration, which is uncommon in standard treatment. How then are the incentive and minimum profit constraints going to constrain the information extraction?

To see the intuition behind the derivation in the following sections, we appeal to a complete-information benchmark. Suppose for the time being that the values of both α and β are fixed and

[9] We can use a more general condition: $R(\alpha, \beta) - C(q(\alpha, \beta)) \geq L$, where L is a finite (and possibly negative) number.

public information. We will make one more modification to our setup. Suppose that the physician has a reservation utility $\bar{U}$, instead of a nonnegative profit constraint. Given this modification, the first-best quantity and payment are those R and q that maximize $\alpha V(q) - R$ subject to $R - C(q) + \beta V(q) \geq \bar{U}$. To solve this problem, simply use the constraint as an equality and substitute for R in the objective function to obtain:

$$\alpha V(q) + \beta V(q) - C(q) - \bar{U}. \tag{3}$$

Now q is chosen to maximize the above: $(\alpha + \beta)V'(q) = C'(q)$.

Why does the optimal quantity in the first best appear to be excessive? The original objective function is simply to maximize the consumer's benefit $(\alpha V(q))$ less any payment to the physician. The maximization of (3) with respect to q treats the total benefit as the *sum* of the consumer's benefit $\alpha V(q)$ and the physician's altruistic component $\beta V(q)$. The reason is this. The transfer R makes the utilities of the patient and the physician transferable. So the social benefit is the sum of the patient's and the physician's utilities; this results in the maximand (3). In other words, because the physician values the patient's benefit, he will be partly compensated by that benefit and partly by the transfer.[10]

Now suppose that the value of $\beta V(q)$ at this solution is bigger than $\bar{U}$. Since the reservation utility constraint binds, $R - C(q)$ becomes negative. Because the insurer is using patient benefit to reward the physician, and because $\beta V(q)$ is sufficiently big (possibly due to a high degree of altruism β), the transfer R can be reduced so much that the physician makes a loss.

Let us now return to our model and replace the reservation utility constraint by the nonnegative profit constraint: $R - C(q) \geq 0$. We must have $R = C(q)$, and after substituting this into the objective function, we obtain $\alpha V(q) - C(q)$, which is maximized at $\alpha V'(q) = C'(q)$. In contrast to the earlier case, the physician's altruism towards patient benefit does *not* affect the choice of q.

Already in the complete information benchmark, a binding minimum profit constraint implies that the optimal design becomes insensitive to the physician's altruism parameter. With incomplete information, the physician will earn information rent, and the minimum profit constraint does not bind always. The quantity can be made sensitive with respect to the physician's private information. Nevertheless, information rent is costly, and we will show that for a range of β, the minimum profit constraint must bind, and quantity becomes insensitive to β and α.

[10] The transfer R is chosen to satisfy the physician's reservation utility constraint.

3 Characterization of Admissible Mechanisms

The direct revelation mechanism consists of the quantity and payment functions that depend on both α and β, and it must be an equilibrium for the physician to report α and β truthfully. Nevertheless, the physician's preferences only depend on β, and it appears that the physician's private information about α cannot be extracted directly. This turns out to be valid, but some arguments are necessary.

For a given mechanism (q, R), we define a doctor's maximum or indirect utility by

$$U(\beta) \equiv \max_{\alpha', \beta'} \{R(\alpha', \beta') - C(q(\alpha', \beta')) + \beta V(q(\alpha', \beta'))\}, \tag{4}$$

which, obviously, is independent of α. Clearly, there cannot be any strict incentive for the physician to report any particular value of α.

Nevertheless, suppose that, for some β, the physician is indifferent between *all* quantity-payment pairs as α changes. That is, $R(\alpha, \beta) - C(q(\alpha, \beta)) + \beta V(q(\alpha, \beta)) = R(\alpha', \beta) - C(q(\alpha', \beta)) + \beta V(q(\alpha', \beta))$, for all α, α', and all these are equal to $U(\beta)$. Then it is an optimal response for him to report α truthfully. However, making quantities and payments contingent on α appears to rely on a delicate balance. This sort of knife-edge construction can become problematic when the incentives for truthful revelation of β are to be considered simultaneously. In other words, although, for a given β, it may be possible to construct R and q as functions of α to satisfy the indifference requirement, this must interfere with the incentive constraint for nearby values of β.

The proof of the following, as well as other results in the paper, can be found in Appendix B:

Lemma 1 *An incentive compatible mechanism $(q(\alpha, \beta), R(\alpha, \beta))$ must have both quantity q and payment R independent of α for almost every β.*

The basic idea for Lemma 1 is as follows. As we have already shown, the indirect utility function U is independent of α. Now because U is the maximum of affine linear functions of β, it must be convex in β, and hence almost everywhere differentiable. Incentive compatibility then implies that the function V must be constant in α; otherwise, the differentiability of U over a dense set of β will be violated. So for almost every value of β, q must not be a function of α; likewise for R.

Lemma 1 is a significant result. Extracting information on α directly is impossible. When the optimal mechanism is only based on β, it is potentially possible that two patients with identical characteristics will receive very different health care depending on the way each interacts with her physician. The optimal mechanism can only base healthcare quantities on the physician agency parameter β, and it must consider the conditional distribution of α given β, as we will see later.

We assume that the distribution of β admits a density. It follows that the set of β where the schedule could depend on α (which has zero Lebesgue measure according to Lemma 1) has no impact on the objective function. We therefore have

Corollary 1 *Without loss of generality, an incentive compatible mechanism can be written as* $(q(\beta), R(\beta))$.

The following Lemma characterizes incentive compatible mechanisms in terms of the quantity q and the indirect utility U.

Lemma 2 *A mechanism* $(R(\beta), q(\beta))$ *is incentive compatible if and only if the indirect utility* $U(\beta) = R(\beta) - C(q(\beta)) + \beta V(q(\beta))$ *is a convex function of* β *and satisfies*

$$\dot{U}(\beta) = V(q(\beta)),$$

for all $\beta \in [\underline{\beta}, \overline{\beta}]$.

This result is a straightforward consequence of the Envelope Theorem. The convexity of the indirect utility is equivalent to the monotonicity of the quantity: $\ddot{U}(\beta) = V'(q(\beta))\dot{q}(\beta) \geq 0$, so q must be nondecreasing. The function U, however, may not be monotone, its derivative depending on the sign of V. The benefit function V is an ordinal measure, its sign being irrelevant. We only insist on V being increasing and concave. The indirect utility must be continuous, since it is convex. However for any incentive compatible mechanism, the quantity q is not necessarily continuous. An upward jump in q corresponds to a kink in U.

Now we make use of the indirect utility function to simplify the set of minimum profit constraints. We do this by establishing a monotonicity result. The profit for a physician with parameter β is $\pi(\beta) = R(\beta) - C(q(\beta))$, and we use the definition of U to rewrite it as:

$$\pi(\beta) = U(\beta) - \beta V(q(\beta)) = U(\beta) - \beta\dot{U}(\beta). \tag{5}$$

There is a geometric representation of profit. Consider the graph of $U(\beta)$ on the (β, U) plane. The value of profit at any β is the intersection of the tangent of U at β with the U-axis.

Differentiating the right-hand side of (5), we show that the physician's profit is nonincreasing in β:

$$(6) \qquad \dot{\pi}(\beta) = \frac{\mathrm{d}}{\mathrm{d}\beta}[U(\beta) - \beta\dot{U}(\beta)] = -\beta\ddot{U}(\beta) = -\beta V'(q(\beta))\dot{q}(\beta) \leq 0.$$

The inequality follows from the convexity of U. In other words, incentive compatibility implies that the physician's profit must be nonincreasing in β; the more the physician cares about patient benefit, the lower his profit level. The monotonicity of physician profit gives the following:

Lemma 3 *For any incentive compatible mechanism, the minimum profit constraint is satisfied for all β if and only if it is satisfied for $\beta = \overline{\beta}$. Moreover, if the minimum profit constraint binds at $\tilde{\beta}$, then the minimum profit constraints for any $\beta > \tilde{\beta}$ must also bind; in other words, binding minimum profit constraints can only occur on an interval $[\hat{\beta}, \overline{\beta}]$. Finally, whenever minimum profit constraints bind, quantities must become constant with respect to β, resulting in pooling: $q(\beta) = \hat{q}$.*

Lemma 3 is the key to the analysis. As functions of β, quantities must be nondecreasing while profits nonincreasing in an incentive compatible mechanism. From (6), $\dot{\pi}(\beta) = 0$ if and only if $\dot{q} = 0$. Setting quantities constant for a range of β implies that the corresponding profits are zero. Pooling quantities for a range of β means binding minimum profit constraints, which save on information rent. The optimal mechanism must consider this pooling interval. In the next section, we prove that pooling must exist in an optimal mechanism; in fact, pooling may be so extensive that it is the only property of an optimal mechanism.

4 The Optimal Payment and Quantity

We let the objective function of the insurer be expected consumer benefit less the expected payment to the physician. This is consistent with competition in the insurance or managed care markets; a firm that fails to pick a mechanism to maximize consumer benefit will be driven out of the market. This is also a necessary condition for (constrained) efficiency. For a given mechanism (q, R), the insurer's objective function is

$$W = \iint [\alpha V(q(\beta)) - R(\beta)]\ h(\alpha, \beta)\ \mathrm{d}\alpha\ \mathrm{d}\beta,$$

where h is the joint density of α and β. Integrating out α, we write W as

$$(7) \qquad W = \int_{\underline{\beta}}^{\overline{\beta}} [\alpha_{\mathrm{m}}(\beta)V(q(\beta)) - R(\beta)]\ g(\beta)\ \mathrm{d}\beta,$$

where $\alpha_{\mathrm{m}}(\beta) \equiv E(\alpha|\beta) = \int \alpha \frac{h(\alpha,\beta)}{g(\beta)} \mathrm{d}\alpha$ is the conditional mean of α given β. In words, $\alpha_{\mathrm{m}}(\beta)$ is the insurer's assessment of a consumer's average valuation given the altruism or physician agency parameter β.

4.1 Deriving the optimal mechanism

An optimal mechanism is an admissible mechanism (q, R) that maximizes (7). We use the definition of U, Lemma 3, and integration by parts to eliminate R, and then find the optimal quantity schedule. From Lemma 3, we can write the integral (7) as the sum of two integrals, one over $[\underline{\beta}, \hat{\beta}]$ and the other $[\hat{\beta}, \overline{\beta}]$, where $\hat{\beta}$ is the lower limit of the pooling interval. Replacing the payment $R(\beta)$ by $U(\beta) + C(q(\beta)) - \beta V(q(\beta))$, we get

$$\int_{\underline{\beta}}^{\hat{\beta}} [\alpha_{\mathrm{m}}(\beta) V(q(\beta)) - R(\beta)]\ g(\beta)\ \mathrm{d}\beta = \int_{\underline{\beta}}^{\hat{\beta}} [(\alpha_{\mathrm{m}}(\beta) + \beta) V(q(\beta)) - C(q(\beta)) - U(\beta)] g(\beta)\ \mathrm{d}\beta.$$

Now we integrate the utility term by parts. Using $\dot{U} = V(q)$ (from Lemma 2) and $U(\hat{\beta}) = \pi(\hat{\beta}) + \hat{\beta} V(q(\hat{\beta})) = \hat{\beta} V(q(\hat{\beta}))$, we get

$$\begin{aligned} \int_{\underline{\beta}}^{\hat{\beta}} [\alpha_{\mathrm{m}}(\beta) V(q(\beta)) - R(\beta)]\ g(\beta)\ \mathrm{d}\beta &= \int_{\underline{\beta}}^{\hat{\beta}} \left\{ \left[\alpha_{\mathrm{m}}(\beta) + \beta + \frac{G(\beta)}{g(\beta)} \right] V(q(\beta)) - C(q(\beta)) \right\} g(\beta)\ \mathrm{d}\beta \\ &- G(\hat{\beta}) \hat{\beta} V(q(\hat{\beta})). \end{aligned}$$

For the integral on the pooling interval $[\hat{\beta}, \overline{\beta}]$, we let the quantity be a constant $\hat{q}$. Because profit is zero over this interval, the optimal payment is $C(\hat{q})$. The objective function is

$$\begin{aligned} W &= \int_{\underline{\beta}}^{\hat{\beta}} \left\{ \left[\alpha_{\mathrm{m}}(\beta) + \beta + \frac{G(\beta)}{g(\beta)} \right] V(q(\beta)) - C(q(\beta)) \right\} g(\beta) \mathrm{d}\beta \\ &- G(\hat{\beta}) \hat{\beta} V(\hat{q}) + \int_{\hat{\beta}}^{\overline{\beta}} [\alpha_{\mathrm{m}}(\beta) V(\hat{q}) - C(\hat{q})] g(\beta) \mathrm{d}\beta, \end{aligned} \tag{8}$$

where $\hat{\beta}$ denotes the lower limit of the pooling interval.[11] The problem of finding the optimal mechanism can now be reformulated as follows: choose a nondecreasing function $q(\beta)$, $\underline{\beta} \leq \beta \leq \hat{\beta}$, and numbers $\hat{q}$ and $\hat{\beta}$, with$\underline{\beta} \leq \hat{\beta} \leq \overline{\beta}$, to maximize (8).

[11] This expression presents a problem in terms of $\hat{\beta}, \hat{q}$ and $q(\beta), \beta \leq \hat{\beta}$. Another form of the problem is also tractable: the problem can be expressed in terms of the indirect utility alone. The objective function can be written as

$$W = \int_{\underline{\beta}}^{\overline{\beta}} [(\alpha_{\mathrm{m}}(\beta) + \beta) \dot{U} - C(V^{-1}(\dot{U})) - U] g(\beta) \mathrm{d}\beta,$$

which is linear in U and strictly concave in $\dot{U}$. The incentive (U convex) and the minimum profit ($U - \beta \dot{U} \geq 0$) constraints define a convex set. The existence and the uniqueness of the optimum follow from standard arguments. In this formulation, calculus of variation can be used to characterize the optimal mechanism. We present here the more intuitive method. The solution via variation is available from the authors.

We now introduce our main assumption, which we adopt for the rest of the paper:

$$\textit{Assumption A: } \alpha_{\mathrm{m}}(\beta) + \beta + \frac{G(\beta)}{g(\beta)} \textit{ is continuous and nondecreasing in } \beta.$$

Assumption A is true under the following two assumptions:

Assumption A1 : $\frac{G(\beta)}{g(\beta)}$ *is continuous and nondecreasing in* β.

Assumption A2: $\alpha_{\mathrm{m}}(\beta)$ *is continuous and nondecreasing in* β.

Assumption A1 is the familiar monotone hazard rate condition, and satisfied for many classes of distributions (uniform, normal, exponential, etc.) Assumption A2 says that the conditional expectation of α is increasing in the physician's altruism parameter β. This seems a natural assumption to make. If the physician's concern for the patient takes into account the patient's valuation, a higher patient average valuation is associated with a physician who exhibits a higher degree of altruism. Nevertheless, the monotonicity of $\alpha_{\mathrm{m}}(.)$ does not imply a nonnegative correlation between α and β; neither does a nonnegative correlation between α and β implies the monotonicity of $\alpha_{\mathrm{m}}(\cdot)$.[12] In any case, Assumption A is weaker than Assumptions A1 or A2.

The form of the objective function in (8) actually reveals the various aspects of the problem. The integral from β to $\hat{\beta}$ refers to the regime of positive profits for the physician. As we have noted at the end of Section 2, when the minimum profit constraint does not bind, the total social benefit $[\alpha + \beta]V(q)$ becomes relevant. Because the information of α cannot be extracted directly, it is replaced by the conditional expectation $\alpha_{\mathrm{m}}(\beta)$. The hazard rate G/g is the familiar adjustment for the rent due to asymmetric information: the "virtual" social benefit is $\alpha_{\mathrm{m}}(\beta) + \beta + \frac{G(\beta)}{g(\beta)}$. The term $\hat{\beta}V(q(\hat{\beta}))$ is a measure of the indirect utility at the beginning of the pooling region. For any $\beta > \hat{\beta}$, the quantity becomes fixed, and the physician's profit becomes zero. The indirect utility becomes βV. So the pooling quantity $\hat{q}$ determines the indirect utility level $\hat{\beta}V(\hat{q})$, which is the base indirect utility level for all those doctors with β smaller than $\hat{\beta}$, hence the factor $G(\hat{\beta})$. Finally, the choices of $\hat{q}$ and $\hat{\beta}$ completely determine the pooling regime, which is the last term in (8).

The optimization program is separable with respect to $\hat{q}$ and $q(\beta)$, β in $[\underline{\beta}, \hat{\beta}]$. So we apply pointwise optimization to obtain the first-order condition for $q(\beta)$, β in $[\underline{\beta}, \hat{\beta}]$:

$$\left[\alpha_{\mathrm{m}}(\beta) + \beta + \frac{G(\beta)}{g(\beta)}\right] V'(q(\beta)) = C'(q(\beta)). \tag{9}$$

[12] The correlation between α and β can be written $\int[\alpha_{\mathrm{m}}(\beta) - \alpha_{\mu}]\beta g(\beta)\mathrm{d}\beta$, where α_{μ} denotes the unconditional mean of α. The monotonicity of $\alpha_{\mathrm{m}}(.)$ does not imply that the correlation is nonnegative.

Under Assumption A, the quantity defined by (9) is nondecreasing (recall that C'/V' is increasing).

Next we show that the optimal quantity is continuous at $\hat{q}$. The intuition is this. If the optimal quantity jumps upward at $\hat{q}$, then the pooling interval can be reduced. The value of $\hat{\beta}$ can be increased while the (higher) quantity at $\hat{q}$ can be kept constant. In other words, if there is a jump, in quantity at $\hat{q}$, the minimum profit continues to hold while the pooling interval can be made smaller; less pooling means that more information about α is revealed.

Lemma 4 *Suppose that $\underline{\beta} < \hat{\beta} < \overline{\beta}$. The optimal quantity is continuous at $\hat{\beta}$. The quantity $\hat{q}$ in the pooling region $[\hat{\beta}, \overline{\beta}]$ must satisfy*

$$\left[\alpha_{\mathrm{m}}(\hat{\beta}) + \hat{\beta} + \frac{G(\hat{\beta})}{g(\hat{\beta})}\right] V'(\hat{q}) = C'(\hat{q}). \tag{10}$$

Next, we characterize the optimal choice for $\hat{q}$. Using the first-order condition with respect to $\hat{q}$, we have the following

Lemma 5 *The pooling interval satisfies the condition*

$$\int_{\hat{\beta}}^{\overline{\beta}} \left\{\alpha_{\mathrm{m}}(\beta) V'(\hat{q}) - C'(\hat{q})\right\} g(\beta) \mathrm{d}\beta = \hat{\beta} G(\hat{\beta}) V'(\hat{q}). \tag{11}$$

The left-hand side of (5) measures the usual (expected) marginal benefit and cost of the quantity $\hat{q}$. The term on the right-hand side measures the change in the base indirect utility level. Raising $\hat{q}$ has a negative effect on the objective function since it gives more profit to all physicians with β less than $\hat{\beta}$. Equation (11) implies that there is no solution with an empty pooling interval; that is, $\hat{\beta} = \overline{\beta}$ cannot satisfy (11). If $\hat{\beta}$ was set at $\overline{\beta}$, then reducing $\hat{\beta}$ must improve the objective function. This would only lead to a second-order loss in the efficiency of q since (9) was satisfied at $\overline{\beta}$, but this would result in a first-order gain since profits for all physicians would be reduced.

Finally, we characterize the pooling region $\hat{\beta}$. Equations (10) and (11) together determine $\hat{\beta}$ and $\hat{q}$ if in fact they yield an interior solution $\hat{\beta} > \underline{\beta}$. It is, however, possible that these two equations yield a solution of $\hat{\beta}$ below $\underline{\beta}$, in which case, the quantity will be constant, and equation (11) with $\hat{\beta}$ set at $\underline{\beta}$ will determine $\hat{q}$. With α_μ denoting the unconditional mean of α, we present the condition for an interior solution:

Lemma 6 *The pooling interval is in the interior of the support of β, $\hat{\beta} > \underline{\beta}$, if and only if*

$$\alpha_\mu > \alpha_{\mathrm{m}}(\underline{\beta}) + \underline{\beta}. \tag{12}$$

When can separation be optimal? We know that the best prospect is for β near $\underline{\beta}$. So now suppose that there is complete pooling, say, the quantity is fixed at $\hat{q}$ for all β. A complete-pooling quantity must be based on the unconditional mean α_μ. What can be gained by some separation at $\underline{\beta}$? Recall that in a separating region, the social benefit $[\alpha_{\mathrm{m}}(\beta) + \beta]V(q)$ is relevant due to strictly positive profits. When $\alpha_\mu > \alpha_{\mathrm{m}}(\underline{\beta}) + \underline{\beta}$, the complete pooling quantity $\hat{q}$ is too high for $\underline{\beta}$. Lowering the quantity from $\hat{q}$ for $\underline{\beta}$ and then subsequently increasing it for higher β (due to Assumption A) will reduce the inefficiency due to the excessive quantity $\hat{q}$. Although this will entail some profits for the physician, it is worthwhile since inequality (12) is strict.

The condition for some separation (12) will fail to hold if the support of β is much larger than that of α, or when the variation of $\alpha_{\mathrm{m}}(\beta)$ is small. When (12) is violated, extracting information of α via β must lead to high profits to the physician due to quantities always higher than under complete pooling. In this case, the optimal quantity is constant on the whole interval $[\underline{\beta}, \overline{\beta}]$ and given by

$$\alpha_\mu V'(\hat{q}) = C'(\hat{q}). \tag{13}$$

We summarize our results by the following:

Proposition 1 *Under Assumption A, the optimal mechanism is defined as follows:*

1. *If $\alpha_\mu \leq \alpha_{\mathrm{m}}(\underline{\beta}) + \underline{\beta}$, the optimal quantity for each value of β is given by equation (13). The physician earns zero profit.*
2. *If $\alpha_\mu > \alpha_{\mathrm{m}}(\underline{\beta}) + \underline{\beta}$, there exists $\hat{\beta}$, with $\underline{\beta} < \hat{\beta} < \overline{\beta}$, and the following are properties for the optimal quantities:*
   - *(a) For $\underline{\beta} \leq \beta \leq \hat{\beta}$, the optimal quantity $q(\beta)$ is strictly increasing and satisfies (9).*
   - *(b) For $\hat{\beta} \leq \beta \leq \overline{\beta}$, the optimal quantity is constant and equal to $\hat{q}$, where $\hat{q}$ and $\hat{\beta}$ satisfy equations (10) and (11).*
   - *(c) The physician earns strictly positive profit if and only if his value of β is less than $\hat{\beta}$.*

In Figures 1, 2 and 3, we display the typical shapes of the optimal quantity, profit, and indirect utility. The indirect utility is convex.[13] It is strictly convex up to $\hat{\beta}$, and then becomes linear.

[13] Recall that the indirect utility may not be monotone; the derivative of U is $V(q)$, but we have made no assumption on the sign of V because it is an ordinal measure.

Indeed, for $\beta \geq \hat{\beta}$, $U(\beta) = \beta V(\hat{q})$. Accordingly, the physician's profit, $\pi = U - \beta \dot{U}$, is strictly decreasing until $\hat{\beta}$, and then becomes zero.[14]

Pooling in the optimal schedule can be interpreted as quantity limits in managed care. In the pooling interval, the quantity is insensitive with respect to β (and therefore α). Moreover, the limit applies to higher values of β, which correspond to higher expected severities or benefits. What is more, the quantity restriction may be extensive; in that case, the managed care plan offers the same quantity that is based on the average severity of the entire population. Where the optimal quantity is increasing with β, it is based on the expectation of α conditional on β. We next investigate how the optimal quantity is related to the first best.

4.2 Comparing the optimal mechanism with the first best

The comparison between the first best and optimal quantity in Proposition 1 is not quite straightforward, because the first best only depends on α while the optimal quantity depends on β. For a comparison, we calculate the expected first-best quantities conditional on β. For each value of β, we consider the conditional distribution of α, and the corresponding first-best quantities in this distribution. For ease of exposition, we compare $C'(q)/V'(q)$ (which is increasing by assumption) rather than the quantity q itself across the asymmetric information and first-best regimes.

Since $C'(q^\star)/V'(q^\star) = \alpha$ at the first best, we have

$$\text{(14)} \qquad \mathrm{E}\left\{ \frac{C'(q^\star(\alpha))}{V'(q^\star(\alpha))} \middle| \beta \right\} \equiv \alpha_{\mathrm{m}}(\beta).$$

In the optimal mechanism, $C'(q)/V'(q)$ is function of β given by Proposition 1. Now the comparison between the first best and second best quantity functions is given by the following proposition, which is proved in Appendix B.

Proposition 2 *At the optimum, we have*

$$\text{(15)} \qquad \int_{\underline{\beta}}^{\overline{\beta}} \frac{C'(q(\beta))}{V'(q(\beta))} g(\beta) \mathrm{d}\beta = \alpha_\mu.$$

[14] Proposition 1 includes the first best as a special case. If there is no uncertainty concering α, $\alpha_\mu = \alpha_{\mathrm{m}}(\beta)$, all β. So the first part of Propositin 1 applies , and equation (13) becomes exactly the one that defines the first-best quantity $q^*(\alpha)$.

If $\alpha_{\mathrm{m}}(\beta)$ is nondecreasing (Assumption A2), there exists $\tilde{\beta}$ with $\hat{\beta} < \tilde{\beta} < \overline{\beta}$ such that

$$\begin{cases} \dfrac{C'(q(\beta))}{V'(q(\beta))} \geq \alpha_{\mathrm{m}}(\beta) & \textit{for } \beta \leq \tilde{\beta} \\[2ex] \dfrac{C'(q(\beta))}{V'(q(\beta))} = \dfrac{C'(\hat{q})}{V'(\hat{q})} \leq \alpha_{\mathrm{m}}(\beta) & \textit{for } \beta \geq \tilde{\beta}. \end{cases}$$

From Proposition 1, we compute the expected value of C'/V' with respect to β. Equation (15) says that the unconditional expectation of $C'(q)/V'(q)$ is exactly α_μ, the unconditional mean of α or the consumer's average valuation. If we assume that $\alpha_{\mathrm{m}}(\beta)$ is nondecreasing, then on average there is overprovision of quantities for low values of β, and underprovision for high values of β. In the separating region (if this region does exist), the optimal quantity is always distorted upwards due to information rent and the social surplus consideration—see equation (9). So because the unconditional expectation of C'/V' must be the same across the two regimes, there must be downward distortion in the pooling regime. Managed care, on average, is associated with a compression of service variations. Figure 4 illustrates Proposition 2. The graph shows two plots of $C'(q)/V'(q)$ against β; the solid line is for the first best, and the other for the optimal mechanism.

4.3 Some examples

A few examples illustrate the scope of our results in Proposition 1. In Example 1, α and β are independent. By contrast, in each of Examples 2 and 3, there is a deterministic relationship between and α and β. In Examples 4 and 5, α and β are both unbounded and correlated.[15]

Example 1: Independent α and β. Suppose that α and β are independent. Then $\alpha_{\mathrm{m}}(\underline{\beta}) = \alpha_\mu$. The first part of Proposition 1 applies: there is complete pooling. Learning about α from any report of β is impossible.

Example 2. Multiplicative α and β. Suppose that $\beta = \alpha\theta$ where $\theta > 0$ is fixed and known (and can be interpreted as a degree of altruism). We have: $\underline{\beta} = \theta\underline{\alpha}$ and $\alpha_{\mathrm{m}}(\underline{\beta}) = \underline{\beta}/\theta = \underline{\alpha}$. There is complete pooling if and only if

$$\alpha_\mu - \underline{\alpha} \leq \theta\underline{\alpha} \qquad \text{or} \qquad \theta \geq \alpha_\mu/\underline{\alpha} - 1.$$

[15] Although our results in Proposition 1 have been written for distributions with finite support, they remain applicable to these examples.

That is, complete pooling is optimal when altruism is high compared to a measure of the range of α. In the separating region (if there is one), the optimal quantity is given by $C'/V' = \beta(1+1/\theta)+G/g$.

The case $\theta = 1$ corresponds to one where the physician puts equal weights on profits and consumer benefits. Here, there is complete pooling if and only if $\alpha_\mu \leq 2\underline{\alpha}$.[16]

Example 3. Additive α and β. Suppose that $\beta = \alpha + \theta$, where again θ is fixed and known. Then $\underline{\beta} = \underline{\alpha} + \theta$ and $\alpha_{\mathrm{m}}(\underline{\beta}) = \underline{\beta} - \theta = \underline{\alpha}$. There is complete pooling if and only if

$$\alpha_\mu - \underline{\alpha} \leq \underline{\alpha} + \theta \qquad \text{or} \qquad \alpha_\mu - 2\underline{\alpha} \leq \theta.$$

In the separating regime region (if there is one), the optimal quantity is given by $C'/V' = 2\beta - \theta + G/g$.

Example 4. Lognormal Distributions. Suppose that $\beta = \alpha.\theta$, and that α and θ are log-normally distributed. That is, $\ln \alpha$ and $\ln \theta$ are normal. Let a_μ and t_μ be the expectations of $\ln \alpha$ and $\ln \theta$, respectively, and σ_a^2 and σ_t^2 their variances. Finally, let ρ denote the correlation between $\ln \alpha$ and $\ln \theta$. The distributions of α, θ and β have a common support $[0, +\infty[$. In other words $\underline{\alpha} = \underline{\beta} = 0$ and $\overline{\alpha} = \overline{\beta} = +\infty$. The expectation of α is $\exp(a_\mu + \sigma_a^2/2) \equiv \alpha_\mu$.

The random variable $\ln \beta = \ln \alpha + \ln \theta$ is normal, with expectation $b_\mu = a_\mu + t_\mu$ and variance $\sigma_b^2 = \sigma_a^2 + \sigma_t^2 + 2\rho\sigma_a\sigma_t$. The distribution of $\ln \alpha$ conditionally on $\ln \beta$ is normal, with expectation

$$l_\mu(\beta) = \mathrm{E}\ (\ln \alpha | \ln \beta) = a_\mu + \frac{\sigma_a^2 + \rho\sigma_a\sigma_t}{\sigma_b^2}(\ln \beta - b_\mu)$$

and some variance σ_l^2 that is strictly smaller than σ_a^2.

We can compute the conditional mean of α given β:

$$\alpha_{\mathrm{m}}(\beta) = \mathrm{E}\ (\alpha|\beta) = \exp\ [l_\mu(\beta) + \sigma_l^2/2].$$

Assumption A2 is satisfied if and only if $\sigma_a^2 + \rho\sigma_a\sigma_t \geq 0$. Therefore, Assumption A2 in this example is equivalent to $\ln \alpha$ and $\ln \beta$ being nonnegatively correlated. Assumption A1 is satisfied since the log-normal distribution has a monotone hazard rate[17].

[16] If $\theta = 0$, then $\beta = 0$ always. The optimal quantity is one that maximizes $\alpha V(q) - C(q)$; this can be implemented by setting $R(q) = C(q)$.

[17] Let g and G be the density and distribution functions of the log-normal distribution, and ϕ and Φ the density and distribution functions of the standard normal distribution. Then $G(x) = \Phi(\ln x)$ and $g(x) = \phi(\ln x)/x$; so $G(x)/g(x) = x\Phi(\ln x)/\phi(\ln x)$. Because Φ/ϕ is increasing, G/g is increasing.

Finally, we have $\underline{\beta} + \alpha_{\rm m}(\underline{\beta}) = \alpha_{\rm m}(0) = 0 < \alpha_\mu$. According to Proposition 1, pooling must not be complete. There exists $\hat{\beta} > 0$ such that the optimal quantity is constant for $\beta \geq \hat{\beta}$ and is given by $C'(q)/V'(q) = \alpha_{\rm m} + \beta + G/g$ for $\beta \leq \hat{\beta}$.

Example 5. Independent Exponential Distributions. Suppose that $\beta = \alpha + \theta$, and that α and θ are independently and exponentially distributed, each with density $\exp(-x)$, on $[0, +\infty)$. Then we have $\underline{\alpha} = \underline{\beta} = 0$ and $\overline{\alpha} = \overline{\beta} = +\infty$. An exponential distribution is a special case of a gamma distribution; more precisely, α and θ each is a gamma distribution with parameters 1 and 1. The sum of two independent gamma distributions with an identical second parameter is again a gamma distribution (DeGroot, 1986, pp288-290). So β follows a gamma distribution with parameters 2 and 1. Accordingly, the density of β is $g(\beta) = \beta \exp(-\beta)$ on $[0, +\infty)$ and the distribution function is $G(\beta) = 1 - (1+\beta)\exp(-\beta)$. The hazard rate is

$$\frac{G(\beta)}{g(\beta)} = \frac{\exp(\beta) - 1 - \beta}{\beta},$$

and increasing in β on $[0, +\infty)$. The distribution of α conditional on β is the uniform distribution on $[0, \beta]$. Therefore, we have $\alpha_{\rm m}(\beta) = \beta/2$, and Assumption A is satisfied. Because $\alpha_\mu = 1$ and $\underline{\beta} = \alpha_{\rm m}(\underline{\beta}) = 0$, Proposition 1 says that there exists $\hat{\beta} > 0$ such that the optimal quantity is constant for $\beta \geq \hat{\beta}$, and for $\beta \leq \hat{\beta}$ the quantity $q(\beta)$ is given by

$$\frac{C'(q)}{V'(q)} = \frac{3}{2}\,\beta + \frac{\exp(\beta) - 1 - \beta}{\beta}.$$

5 Second Best Physician Agency

In the previous section, information about α and β is only known to the physician; this may be regarded as a third best. If the value of α were known to the managed care company, the physician's concern for the patient's benefit was irrelevant and the first-best quantity that maximized $\alpha V(q) - C(q)$ could be implemented. In this section, we consider a second best, where the value of β is known to the managed care company, but the information on α remains the physician's private information.[18]

When $\beta > 0$ is known, Corollary 1 does not apply. So we must consider mechanisms in which the physician is asked to report α. Now we assume that the distribution of α admits a strictly

[18] The situation where the physician agency parameter is known can be regarded as the typical scenario considered by many papers in the literature. These papers, however, did not study the problems due to asymmetric information on the patient's valuation.

positive density f on the interval $[\underline{\alpha}, \overline{\alpha}]$, with $\underline{\alpha} \geq 0$. A mechanism, a pair of functions $(q(\alpha), R(\alpha))$, is said to be incentive compatible if for all α and α'

$$R(\alpha) - C(q(\alpha)) + \beta V(q(\alpha)) \geq R(\alpha') - C(q(\alpha')) + \beta V(q(\alpha')). \tag{16}$$

The physician's preferences do not depend on α. So (16) can be written as

$$R(\alpha) - C(q(\alpha)) + \beta V(q(\alpha)) = U, \tag{17}$$

for all α, and some constant U. Instead of working with $(q(\alpha), R(\alpha))$, we shall use the quantity function $q(.)$ and the level of utility U (a scalar) as instruments. Given a quantity schedule $q(\alpha)$ and a constant U, we can use (17) to recover the payment $R(\alpha)$. Besides the incentive constraints, a mechanism must ensure that the physician makes a nonnegative profit: $R(\alpha) - C(q(\alpha)) \geq 0$, for all α.

Although a mecahnism satisfying (17) removes all incentives for the physician to misreport α, there cannot be any strict incentive for the truthful revelation of this information. The physician's preferences do not directly depend on α. Here, we make the usual assumption that a physician truthfully reveals the information of α if there is no incentive to do otherwise. In effect, we select one equilibrium among a large set of equilibria induced by $(q(\alpha), U)$. These other equilibria are supported by other physician reporting strategies, for example, the physician always reporting the highest (or lowest) value of α whenever he is indifferent between reports.

Given the truthful revelation of the information of α, the objective function of the insurer is

$$W = \int_{\underline{\alpha}}^{\overline{\alpha}} \{\alpha V(q) - R(\alpha)\} f(\alpha) \mathrm{d}\alpha = \int_{\underline{\alpha}}^{\overline{\alpha}} \{(\alpha + \beta) V(q(\alpha)) - C(q(\alpha))\} f(\alpha) \mathrm{d}\alpha - U.$$

The optimal mechanism maximizes W subject to the minimum profit constraints: $\beta V(q(\alpha)) \leq U$ for all α.

The solution is easy to describe. Obviously, pointwise optimization can be applied where the minimum profit constraint does not bind. This yields a first-order condition: $(\alpha + \beta) V'(q) = C'(q)$. When the physician earns positive profits, the social benefit $(\alpha + \beta) V(q)$ should be considered, and so the first-order condition describes the appropriate marginal benefit and cost calculations. This also yields an optimal quantity schedule $q(\alpha)$ that is increasing in α. So for a given U, the minimum profit constraints will bind for all values of α above a certain threshold, say, $\hat{\alpha}$; once $\alpha > \hat{\alpha}$, the optimal quantity becomes constant.

The optimal choice of U is never too high, so that some of the minimum profit constraints must bind. Again, this can be explained by the envelope argument. If the threshold $\hat{\alpha}$ was originally at the upper support, then lowering U would reduce profits for all physicians, a first-order gain; this only would lead to a second-order loss since the marginal conditions originally were satisfied. In other words, there must be some pooling. If the value of β is very high, however, the minimum profit constraint may become binding even at the lower support $\underline{\alpha}$; that is, $\hat{\alpha} = \underline{\alpha}$. In this case, the optimal quantity becomes constant for all values of α, and given by $\alpha_\mu V'(q) = C'(q)$. Again, there may be complete pooling.

Proposition 3 *When the value of β is public information, the optimal mechanism is defined as follows.*

1. *If $\alpha_\mu - \underline{\alpha} \leq \beta$, the optimal quantity is constant and given by*

$$\alpha_\mu V'(q) = C'(q) \tag{18}$$

2. *If $\alpha_\mu - \underline{\alpha} > \beta$, then there exists $\hat{\alpha}$, with $\underline{\alpha} < \hat{\alpha} < \overline{\alpha}$ such that the optimal quantities have the following properties:*

   (a) *For $\underline{\alpha} < \alpha < \hat{\alpha}$, the optimal quantity is strictly increasing and satisfies*

$$(\alpha + \beta)V'(q) = C'(q). \tag{19}$$

   (b) *For $\hat{\alpha} \leq \alpha \leq \overline{\alpha}$, the optimal quantity is a constant $\bar{q}$, and given by*

$$\int_{\hat{\alpha}}^{\overline{\alpha}} \left\{\alpha V'(\bar{q}) - C'(\bar{q})\right\} f(\alpha)\mathrm{d}\alpha = \beta V'(\bar{q})F(\hat{\alpha}). \tag{20}$$

   The optimal quantity is continuous at $\hat{\alpha}$ so that the equation

$$(\hat{\alpha} + \beta)V'(\bar{q}) = C'(\bar{q}) \tag{21}$$

   together with (20) determine $\hat{\alpha}$ and $\bar{q}$.

   (c) *The physician earns strictly positive profit if and only if α is less than $\hat{\alpha}$.*

The symmetry in Propositions 1 and 3 is striking, although the incentive constraints in the second best and third best are quite different.[19] The characteristics of the optimal quantity and physician profits in the second best can easily be illustrated by Figures 1 and 2—the necessary

[19] There is no information rent term G/g for the physician agency in Proposition 3.

modification being a change in the label of the horizontal axis from β to α. The quantitative differences between the second best mechanism and that in section 4 can be quite large.

The symmetry does not end here. The comparison between the second best and the first best actually parallels that between the third best and the first best. From equations (19) and (20),

$$\int_{\underline{\alpha}}^{\overline{\alpha}} \frac{C'(q(\alpha))}{V'(q(\alpha))} f(\alpha) \mathrm{d}\alpha = \alpha_{\mu},$$

which is symmetric to equation (15) in Proposition 2. From this, we can easily compare the second best with the first best. Again, given the value of β, in the second best, there is always overprovision of quantities for lower values of α, and underprovision for high α. Figure 4 illustrates this comparison if the label of the horizontal axis is changed to α.

These surprising comparisons indicate that the design of optimal payment and quantity depends critically on the existence of physician agency. Asymmetric information concerning physician agency adds one more dimension to the problem, but the basic issue is the missing information about the consumer's valuation of healthcare quantities. Physician agency is a relationship through which an insurer must attempt to extract this missing information.

In the paper, we have maintained the assumption of minimum profits for the physician. In Appendix C, we provide a derivation of the optimal mechanism in the third best when the minimum profit constraints are replaced by reservation utility constraints. The results there actually show that reservation utility constraints are unappealing. Assumptions that should not have any economic bearing, such as the sign of an ordinal utility function, determine the characteristics of optimal mechanism. The appendix also draws some connection between the model here and the literature on countervailing incentives (Lewis and Sappington (1989), Jullien (2000)).

6 Conclusion

We hypothesize that physicians interact with patients in complex ways, and have proposed a model of asymmetric information for the complexity in physician agency. The way the physician agency weighs physician profit and patient benefit is unknown to the insurer. This, however, is only one piece of missing information. The insurer does not know the patient's valuation for healthcare either. We study the optimal mechanism when these two pieces of information are unknown to the insurer.

We view the design of an optimal mechanism as an attempt to base payment and quantity on a patient's valuation of healthcare benefit. The insurer recognizes that this valuation information is unavailable. What is more, an attempt to extract this information must face the difficult task of resolving the complexity of physician agency. It is only through the physician agency that an insurer can get to this information.

The optimal mechanism exhibits properties commonly found in managed care. For example, the insuer imposes a fixed quantity for an extensive range of patient characteristics; this is due to the information about patient valuation being too costly to extract. More important, any variation of quantities is related to physician agency characteristics. The optimal mechanism cannot tie quantities to intrinsic patient valuation directly. So two patients with identical healthcare problems may receive different healthcare quantities depending on the particular relationship each has with her physician.

The complexity of physician-patient interactions affects many aspects of the study of the health market. The usual program of inducing cost efficiency and service quality necessarily assumes some provider objective. Usually, the objective is assumed to be known. Obviously, when uncertainty of provider objectives is present, the optimal mechanism will have to consider tradeoffs differently. Moreover, problems such as dumping and skimping have to be reconsidered if the physician agency shows some preferences toward patient welfare or benefits, and if these preferences are private information.

A Appendix: Game Forms for Physician Agency

We provide three examples to illustrate physician agency. Recall that $\alpha V(q)$ denotes the patient's (or the regulator's perception) benefit from quantity q. On the other hand, the preferences for the physician agency are represented by $\beta V(q) + R(q) - C(q)$. Here we give explicit game forms to generate the physician agency preferences. In these examples, the physician is only interested in his own profit.

The first example is a generalized Nash bargaining solution. Let the physician and the patient bargain cooperatively; an agreement is a pair (q, t), where q denotes the treatment quantity, and t a transfer from the patient to the physician. Let $R(q)$ denote the physician's payment from the insurer when he provides quantity q. So under an agreement (q, t), the patient's utility is $\alpha V(q) - t$, and the physician's utility $R(q) + t - C(q)$. Let the disagreement point be defined by zero utility level for both the physician and patient.

Under generalized Nash bargaining, an agreement splits the surplus in the ratio γ and $1 - \gamma$, $0 < \gamma < 1$, for the physician and the patient.[20] Formally, the generalized Nash bargaining solution is a pair of q and t that solve:

$$\max_{q,t}[R(q) - C(q) + t]^{\gamma}[\alpha V(q) - t]^{(1-\gamma)}.$$

The first-order conditions with respect to t and q are:

$$\frac{\gamma}{R(q) - C(q) + t} - \frac{1-\gamma}{\alpha V(q) - t} = 0$$

$$\frac{\gamma[R'(q) - C'(q)]}{R(q) - C(q) + t} + \frac{(1-\gamma)\alpha V'(q)}{\alpha V(q) - t} = 0.$$

Combining them, we obtain the characterization of the coalition's choice of q:

$$\alpha V'(q) + R'(q) - C'(q) = 0.$$

Nash bargaining will guide the coalition to maximize $\alpha V(q) + R(q) - C(q)$, the total surplus, and divide the maximum surplus accordingly with a transfer. Nash bargaining generates a special case of the physician agency preferences if we let β in section 2 be α.

In the second example, we maintain the generalized Nash bargaining solution, but eliminate the transfer. The bargaining solution is given by the first-order condition:

$$\frac{\gamma[R'(q) - C'(q)]}{R(q) - C(q)} + \frac{(1-\gamma)\alpha V'(q)}{\alpha V(q)} = 0.$$

[20] The split of surplus may be a function of α, but here we will suppress that.

Rearranging terms, we have

$$\left\{\frac{1-\gamma}{\gamma}\ \frac{R(q)-C(q)}{V(q)}\right\}\ V'(q)+R'(q)-C'(q).$$

We can define a parameter β and a function $S(q)$ by the following:

$$\beta S'(q)\equiv\left\{\frac{1-\gamma}{\gamma}\ \frac{R(q)-C(q)}{V(q)}\right\}\ V'(q).$$

Here, Nash bargaining will lead to the coalition to maximize $\beta S(q)+R(q)-C(q)$. This expression is slightly different from the physician agency preferences in section 2. Our formulation there can be regarded as an approximation. In any case, we have worked out the conditions, derivations, and results when the physician agency is characterized by preferences $\beta S(q)+R(q)-C(q)$ while the consumer's intrinsic benefits from treatment are $\alpha V(q)$. The results are qualitatively similar to those in the propositions; details are available from us.

Finally, we discuss a repeated interaction or demand response example. Here, the physician picks a quantity for the patient; the patient may search for another provider if she feels that the quantity is unsatisfactory. A physician prefers the patient to continue with their relationship; continuation saves setup, information, and other fixed costs. Nevertheless, the patient's outside options are unknown to the physician. A higher likelihood of continuation results if the physician prescribes a higher quantity for the patient. Let a function M denote the discounted expected profits if the patient chooses to stay with the physician. The function M is assumed to be increasing in $\alpha V(q)$.

The physician's current profit is $R(q)-C(q)$. So his total discounted expected profit from possible future interactions with the patient is $M(\alpha V(q);\omega)+R(q)-C(q)$, where ω is a vector of random variables affecting the prospect of continuation, and may be correlated with α. The physician chooses q to maximize the total discounted expected profit. The first-order condition is:

$$\frac{\partial M}{\partial q}\alpha V'(q)+R'(q)-C'(q)=0.$$

Now we simply write $\frac{\partial M}{\partial q}(\alpha V(q);\omega)\alpha V'(q)$ as $\beta V'(q)$, and the physician agency preferences in section 2 emerge. Again, there is a slight difference. The function M depends on the quantity, which is affected by the insurer's payment and quantity mechanism. On the other hand, we have assumed that the distribution of β is exogenous. Our formulation in section 2 is an approximation.

B Appendix: Proof of Lemmas and Propositions

Proof of Lemma 1

Consider an incentive compatible mechanism $(q(\alpha,\beta), R(\alpha,\beta))$. Recall that the indirect utility is defined by

$$U(\beta) = \max_{\alpha',\beta'}\{R(\alpha',\beta') - C(q(\alpha',\beta')) + \beta V(q(\alpha',\beta'))\}.$$

Since U is the upper bound of affine functions of β, it is convex in β (Rockafellar, 1972, Theorem 5.5). Therefore U is differentiable for almost every β in $[\underline{\beta},\overline{\beta}]$ (Rockafellar, 1970, Theorem 25.5).

For $\beta < \beta'$ and for all α, α', the incentive constraints imply

$$V(q(\alpha,\beta)) \le \frac{U(\beta') - U(\beta)}{\beta' - \beta} \le V(q(\alpha',\beta')). \tag{22}$$

Therefore, for $\beta' < \beta < \beta''$ and all α'

$$\frac{U(\beta) - U(\beta')}{\beta - \beta'} \le \min_{\alpha'} V(q(\alpha',\beta)) \le \max_{\alpha'} V(q(\alpha',\beta)) \le \frac{U(\beta'') - U(\beta)}{\beta'' - \beta}.$$

As $\beta' \to \beta^-$ and $\beta'' \to \beta^+$, the left and right derivatives of U at β satisfy

$$\left(\frac{\mathrm{d}U}{\mathrm{d}\beta}\right)_- \le \min_{\alpha'} V(q(\alpha',\beta)) \le \max_{\alpha'} V(q(\alpha',\beta)) \le \left(\frac{\mathrm{d}U}{\mathrm{d}\beta}\right)_+. \tag{23}$$

If $\min_{\alpha'} V(q(\alpha',\beta)) < \max_{\alpha'} V(q(\alpha',\beta))$, U would *not* be differentiable at β. But U is differentiable for almost every β. So we must have $\min_{\alpha'} V(q(\alpha',\beta)) = \max_{\alpha'} V(q(\alpha',\beta))$, or $q(\alpha,\beta) = q(\alpha',\beta)$ for almost every β. In turn, this implies $R(\alpha,\beta) = R(\alpha',\beta)$ for almost every β (since U does not depend on α).

Proof of Corollary 1

On the set (of zero Lebesgue measure) where U is not differentiable, we change the schedule as follows. For any β in this set, we pick any α_0 and replace, for all α', $(q(\alpha',\beta), R(\alpha',\beta))$ by $(q(\alpha_0,\beta), R(\alpha_0,\beta))$. In other words, we select an arbitrary value in the subgradient of U at β, under the only constraint that this value is the *same* for all α. This selection does not change the value of the objective function, since it only affects a set of zero measure. The resulting schedule depends only on β, leads to the same value of the objective function as before, and is incentive compatible. It follows that without loss of generality we can consider only schedules depending only on β. For almost every β, the indirect utility function is differentiable at β and its derivative is $V(q(\beta))$ (see equation (23) or simply apply the Envelope Theorem).

Proof of Lemma 2

The first part of Lemma 2 follows from the proof of Lemma 1. We now prove the second part. So consider a pair (q, U), with U convex and $\dot{U} = V(q)$. Define the payment R by $R(\beta) = U(\beta) + C(q(\beta)) - \beta V(q(\beta))$. The incentive compatibility constraint is

$$R(\beta) - C(q(\beta)) + \beta V(q(\beta)) \geq R(\beta') - C(q(\beta')) + \beta V(q(\beta')).$$

After substituting R by $U(\beta) + C(q(\beta)) - \beta V(q(\beta))$, the incentive constraint becomes

$$U(\beta) \geq U(\beta') + V(q(\beta'))(\beta - \beta').$$

The inequality in the above is valid because $\dot{U}(\beta') = V(q(\beta'))$ and U is convex. So (q, R) is incentive compatible.

Proof of Lemma 3

Since the profit π is nonincreasing, $\pi(\beta) \geq 0$ for all β is equivalent to $\pi(\overline{\beta}) \geq 0$. Moreover, if there exists $\hat{\beta} < \overline{\beta}$ such that $\pi(\hat{\beta}) = 0$, then for all $\beta \geq \hat{\beta}$, we must have $\pi(\beta) = 0$. On that interval, the profit is identically 0, so its derivative $\dot{\pi}(\beta) = \beta V'(q(\beta))\dot{q}(\beta)$ must be zero as well; this implies that q is constant on that interval.

Proof of Lemma 4

Suppose that q jumps upward at $\hat{\beta}$; that is, $\hat{q} > q(\hat{\beta}_-)$, where $q(\hat{\beta}_-)$ is given by (9) at $\beta = \hat{\beta}$. Then we could slightly increase $\hat{\beta}$, while keeping $\hat{q}$ constant. This change would respect the monotonicity requirement. The impact on the objective function is given by

$$\left(\frac{\partial W}{\partial \hat{\beta}}\right)_{\hat{q} \text{ fixed}} = \left[\alpha_{\mathrm{m}}(\hat{\beta}) + \hat{\beta} + \frac{G(\hat{\beta})}{g(\hat{\beta})}\right] [V(q(\hat{\beta}_-)) - V(\hat{q})] - [C(q(\hat{\beta}_-)) - C(\hat{q})]. \tag{24}$$

Using $\alpha_{\mathrm{m}}(\hat{\beta}) + \hat{\beta} + G/g(\hat{\beta}) = C'(q(\hat{\beta}_-))/V'(q(\hat{\beta}_-))$ and the assumption that $\hat{q} > q(\hat{\beta}_-)$, we know that the above derivative is strictly positive (recall that the cost function C is strictly convex and the benefit V is strictly concave). So increasing $\hat{\beta}$ would increase the objective function. We conclude that q is continuous at $\hat{\beta}$.

Proof of Lemma 5

We consider a small variation in $\hat{q}$. In principle, we have to change $\hat{\beta}$ accordingly to respect the monotonicity requirement at $\hat{\beta}$. Nevertheless, due to the continuity of q (see equation (24) with

$\hat{q} = q(\hat{\beta})$), the direct impact of the induced variation in $\hat{\beta}$ on the objective function is zero:

$$\left(\frac{\partial W}{\partial \hat{\beta}}\right)_{\hat{q}\text{ fixed}} = 0.$$

So the total impact of the change is

$$\frac{\partial W}{\partial \hat{q}} = \int_{\hat{\beta}}^{\overline{\beta}} \left\{\alpha_{\mathrm{m}}(\beta)V'(\hat{q}) - C'(\hat{q})\right\} g(\beta)\mathrm{d}\beta - \hat{\beta}G(\hat{\beta})V'(\hat{q}),$$

which gives equation (11) and achieves the proof of the Lemma.

Proof of Lemma 6

Use equations (10) and (11) to eliminate $\hat{q}$. After simplifying and applying integration by parts, we obtain the following equation for $\hat{\beta}$

$$(25) \qquad \int_{\hat{\beta}}^{\overline{\beta}} \left\{ \left[\alpha_{\mathrm{m}}(\beta) + \beta + \frac{G(\beta)}{g(\beta)}\right] - \left[\alpha_{\mathrm{m}}(\hat{\beta}) + \hat{\beta} + \frac{G(\beta)}{g(\hat{\beta})}\right] \right\} g(\beta)\mathrm{d}\beta - \overline{\beta} = 0.$$

By Assumption A, the left-hand side of (25) is continuous and nonincreasing in $\hat{\beta}$; it is equal to $-\overline{\beta}$ at $\hat{\beta} = \overline{\beta}$. Also, it is equal to $\alpha_\mu - \alpha_{\mathrm{m}}(\underline{\beta}) - \underline{\beta}$ at $\hat{\beta} = \underline{\beta}$, where α_μ is the unconditional mean of α. So as $\hat{\beta}$ varies between $\underline{\beta}$ and $\overline{\beta}$, the left-hand side of (25) varies between $\alpha_\mu - \alpha_{\mathrm{m}}(\underline{\beta}) - \underline{\beta}$ and $-\overline{\beta}$.

If $\alpha_\mu - \alpha_{\mathrm{m}}(\underline{\beta}) - \underline{\beta} > 0$, there exists a unique solution $\hat{\beta}$ to (25) satisfying $\underline{\beta} < \hat{\beta} < \overline{\beta}$. Otherwise, if $\alpha_\mu - \alpha_{\mathrm{m}}(\underline{\beta}) - \underline{\beta} \leq 0$, there exists no value of $\hat{\beta}$ between $\underline{\beta}$ and $\overline{\beta}$ to fulfill (25), and we have a corner solution $\hat{\beta} = \underline{\beta}$.

Proof of Proposition 2

Equation (15) follows from (9) and (11):

$$\int_{\underline{\beta}}^{\overline{\beta}} \frac{C'(q(\beta))}{V'(q(\beta))} g(\beta)\mathrm{d}\beta = \int_{\underline{\beta}}^{\hat{\beta}} \left[\alpha_{\mathrm{m}}(\beta) + \beta + \frac{G(\beta)}{g(\beta)}\right] g(\beta)\mathrm{d}\beta + \int_{\hat{\beta}}^{\overline{\beta}} \alpha_{\mathrm{m}}(\beta) g(\beta)\mathrm{d}\beta - \hat{\beta}G(\hat{\beta}) = \alpha_\mu.$$

From equation (9), $C'(q)/V'(q) \geq \alpha_{\mathrm{m}}(\beta)$ for $\beta \leq \hat{\beta}$. Now the existence of $\tilde{\beta} > \hat{\beta}$ where $C'(\hat{q})/V'(\hat{q}) \leq \alpha_{\mathrm{m}}(\beta)$ on $(\tilde{\beta}, \overline{\beta})$ follows from equation (15).

Proof of Proposition 3

For a given level of utility U, we choose each $q(\alpha)$ to maximize the objective function subject to the minimum profit constraints. Pointwise maximization leads to equation (19), and we must check that it satisfies the minimum profit constraints. Let us define $\bar{q}(U)$ by $\beta V(\bar{q}(U)) = U$; for later use, note that

$$\bar{q}'(U) = \frac{1}{\beta V'(\bar{q}(U))}.$$

Then any $q(\alpha)$ satisfying (19) is an optimal quantity if and only if $q(\alpha) \leq \bar{q}(U)$. It is easy to verify that any $q(\alpha)$ satisfying (19) is increasing in α. Accordingly, we can define $\hat{\alpha}(U)$ such that for $\alpha \leq \hat{\alpha}(U)$, (19) holds, and the value of $\hat{\alpha}(U)$ is given by

$$\frac{C'(\bar{q}(U))}{V'(\bar{q}(U))} = \hat{\alpha}(U) + \beta. \tag{26}$$

Now the objective function can be written as a function of U alone:

$$W = \int_{\underline{\alpha}}^{\hat{\alpha}(U)} \{(\alpha+\beta)V(q(\alpha)) - C(q(\alpha))\} f(\alpha)\mathrm{d}\alpha + \int_{\hat{\alpha}(U)}^{\overline{\alpha}} \{(\alpha+\beta)V(\bar{q}(U)) - C(\bar{q}(U))\} f(\alpha)\mathrm{d}\alpha - U,$$

where $q(\alpha)$ satisfies (19) on $[\underline{\alpha}, \hat{\alpha}(U)]$. Differentiating with respect to U yields

$$\begin{aligned} W'(U) &= \bar{q}'(U) \int_{\hat{\alpha}(U)}^{\overline{\alpha}} \{(\alpha+\beta)V'(\bar{q}(U)) - C'(\bar{q}(U))\} f(\alpha)\mathrm{d}\alpha - 1 \\ &= \frac{1}{\beta} \int_{\hat{\alpha}(U)}^{\overline{\alpha}} \left\{(\alpha+\beta) - \frac{C'(\bar{q}(U))}{V'(\bar{q}(U))}\right\} f(\alpha)\mathrm{d}\alpha - 1, \end{aligned}$$

where the second equality follows from the definition of $\bar{q}'(U)$.

The functions $\bar{q}(U)$ and $\hat{\alpha}(U)$ are nondecreasing with respect to U. It follows that W' is nonincreasing and, therefore, W is a concave function of U. The optimal level of U is given by $W'(U) = 0$, if such a U exists. This yields equation (20).[21]

To prove the first part of the Proposition, recall that $\hat{\alpha}(U)$ is nondecreasing in U; the higher the value of U, the larger is the pooling interval. There is complete pooling if and only if $W'(U) \leq 0$ for a U where $\hat{\alpha}(U) = \underline{\alpha}$. For such a value of U, we have, from (26),

$$W'(U) = \frac{1}{\beta}\left[\alpha_\mu + \beta - \frac{C'(\bar{q}(U))}{V'(\bar{q}(U))}\right] - 1 = \frac{1}{\beta}[\alpha_\mu - \hat{\alpha}(U)] - 1 = \frac{1}{\beta}[\alpha_\mu - \underline{\alpha}] - 1.$$

So when $W'(U) \leq 0$, or $\alpha_\mu - \underline{\alpha} \leq \beta$ as in the first part of the Proposition, it is a corner solution. The physician earns zero profit, while the optimal quantity satisfies (18).

[21] If a physician can incur a loss $L < 0$, the minimum profit constraint is given by $\beta V(q(\alpha)) \leq U + L$. In the proof of Proposition 3, we must replace U by $U + L$. Since U is endogenous, the optimal value of α, which determines the pooling interval, is unchanged. We simply replace the payment R by $R - L$.

C Appendix: Minimum Profit versus Reservation Utility Constraints

To assess the role of minimum profit, we solve a version of the model with a reservation utility constraint for the physician (or the physician agency). Without loss of generality, we let the reservation utility be 0; a mechanism must guarantee a nonnegative indirect utility. That is, $U \equiv \beta V(q(\beta)) + R(\beta) - C(q(\beta)) \geq 0$. Results turn out to be rather different under this constraint. First, the sign of V does matter, while it does not when a minimum profit constraint is used instead (results in the propositions above depend only on the derivative of V, or the marginal utility). This is unsatisfactory. The function V is an ordinal measure of the patient's benefit, and its sign should not have a bearing on economic principles, but this is not true.[22] In other words, we argue here that a minimum profit constraint is more appealing.

Suppose first the function V is everywhere positive. Then the indirect utility is increasing. It follows that the reservation utility constraint binds at $\underline{\beta}$. Integrating by parts the utility term ($\int Ug = \int \dot{U}(1-G)$) yields

$$W = \int_{\underline{\beta}}^{\overline{\beta}} \left\{ \alpha_{\mathrm{m}}(\beta)\dot{U} - C(V^{-1}(\dot{U})) + \beta\dot{U} - \frac{1-G}{g}\,\dot{U} \right\} g(\beta)\mathrm{d}\beta.$$

Maximizing pointwise with respect to $\dot{U}$ leads to

$$\frac{C'(q)}{V'(q)} = \beta + \alpha_{\mathrm{m}}(\beta) - \frac{1-G(\beta)}{g(\beta)}. \tag{27}$$

The right-hand side of (27) is increasing in β under Assumption A2 and $(1-G)/g$ nonincreasing. So the quantity schedule satisfying (27) is incentive compatible, and therefore optimal. The optimal quantity schedule exhibits no pooling.

Suppose now that V is everywhere negative. Then U is decreasing and the reservation utility constraint binds at $\overline{\beta}$. Integrating by parts the utility term ($\int Ug = -\int \dot{U}G$) and maximizing pointwise yields

$$\frac{C'(q)}{V'(q)} = \beta + \alpha_{\mathrm{m}}(\beta) + \frac{G(\beta)}{g(\beta)}. \tag{28}$$

which is increasing under Assumption A. This is therefore the solution.

Finally, suppose that $V(q)$ is negative for small q and positive for large q. Let q_1 be defined by $V(q_1) = 0$. Since $\dot{U} = V(q)$, the indirect utility U first decreases and then increases. So suppose

[22] For example, results in this paper remain unchanged if we replace the function V by $V - 10{,}000$. This no longer holds true if a reservation utility constraint replaces our minimum profit constraint.

that U is decreasing on $[\underline{\beta}, \beta_1]$, constant on $[\beta_1, \beta_2]$, and increasing on $[\beta_2, \overline{\beta}]$, $\beta_1 \leq \beta_2$. Because the reservation utility constraint must bind, $U(\beta) = 0$ for all $\beta \in [\beta_1, \beta_2]$.

We now show that $\beta_1 < \beta_2$. Suppose to the contrary that $\beta_1 = \beta_2$. Integrating by parts on the two intervals and maximizing with respect to q leads to the following: the quantity is given by (28) on $[\overline{\beta}, \beta_1]$ and by (27) on $[\beta_1, \overline{\beta}]$. This leads to a downward discontinuity of U at β_1, violating the monotonicity of q. It follows that $\beta_1 < \beta_2$.

Since $U(\beta) = 0$ for all $\beta \in [\beta_1, \beta_2]$, we have $\dot{U} = V(q) = 0$ on that interval and $q(\beta) = q_1$. By the same computations (integration by parts and pointwise maximization), we conclude that the optimal quantity is given by

$$\frac{C'(q)}{V'(q)} = \begin{cases} \beta + \alpha_{\mathrm{m}}(\beta) + \frac{G(\beta)}{g(\beta)} & \text{if } \beta \leq \beta_1 \\ \dfrac{C'(q_1)}{V'(q_1)} & \text{if } \beta_1 \leq \beta \leq \beta_2 \\ \beta + \alpha_{\mathrm{m}}(\beta) - \frac{1-G(\beta)}{g(\beta)} & \text{if } \beta \geq \beta_2 \end{cases}$$

where β_1 and β_2 are given by

$$\frac{C'(q_1)}{V'(q_1)} = \beta_1 + \alpha_{\mathrm{m}}(\beta_1) + \frac{G(\beta_1)}{g(\beta_1)} = \beta_2 + \alpha_{\mathrm{m}}(\beta_2) - \frac{1 - G(\beta_2)}{g(\beta_2)}.$$

When the reservation utility constraint $U \geq 0$ replaces the minimum profit constraint $\pi \geq 0$, pooling may result; any pooling must occur in the strict interior of the support of β. Nevertheless, the reason for pooling is very different. Because of the change of the sign of V, there are countervailing incentives, as in Lewis and Sappington (1989). For small β, the physician has an incentive to under-report β while the opposite is true for high β. Generally, when a reservation utility constraint is imposed, the solution depends on the sign of V. By contrast, under the minimum profit constraint, the solution only depends on V'.

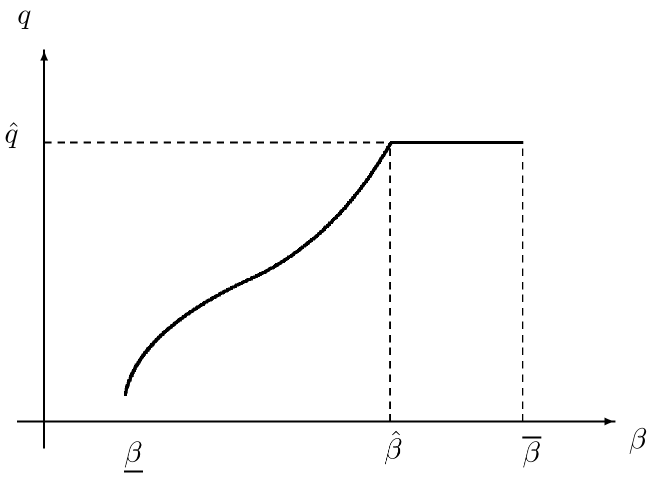

Figure 1: Optimal quantity

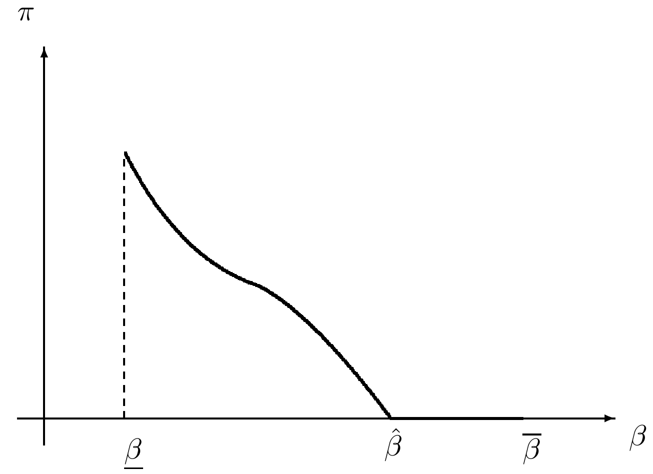

Figure 2: Physician profit in optimal mechanism

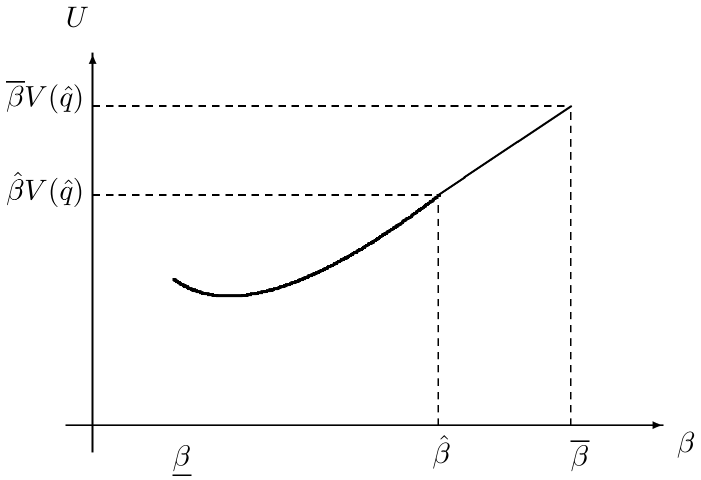

Figure 3: Indirect utility in optimal mechanism

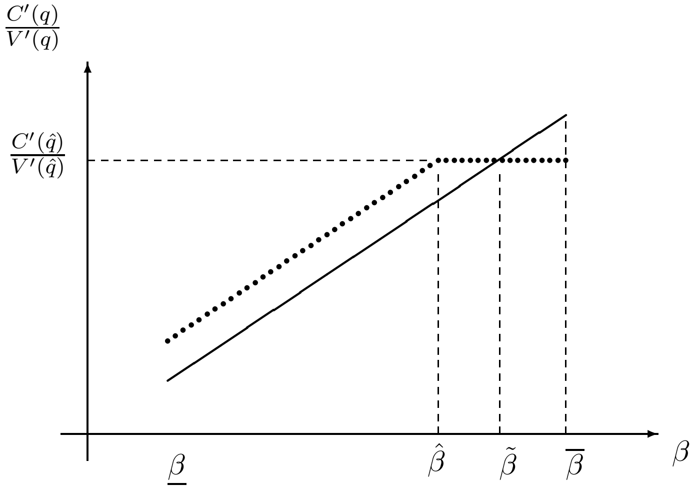

Figure 4: Comparison between first best (solid line) and third best (dotted line)

References

Armstrong, Mark and Rochet, Jean-Charles (1999). "Multidimensional Screening: A User's Guide," *European Economic Review*, 43:959-979.

Arrow, Kenneth J. (1963). "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, 53(5):941-969.

Baron, David P. and Myerson, Roger B. (1982). "Regulating a Monopolist with Unknown Costs," *Econometrica*, 50:911-930.

Baumgardner, James (1991). "The Interaction between Forms of Insurance Contract and Types of Technical Change in Medical Care," *Rand Journal of Economics*, 22(1):36-53.

Besley, Timothy and Ghatak, Maitreesh (2003). "Competition and Incentives with Motivated Agents," London School of Economics working paper.

Chalkley, Martin and Malcomson, James M. (1998). "Contracting for Health Services when Patient Demand Does Not Reflect Quality," *Journal of Health Economics*, 17(1):1-20.

DeGroot, Morris H. (1986). *Probability and Statistics, second edition*, Addison Wesley, Reading Mass.

Dixit, Avinash (2002). "Incentive Contracts for Faith-Based Organizations to Deliver Social Services," Princeton University working paper.

Dranove, David (1988). "Demand Inducement and the Physician/Patient Relationship," *Economic Inquiry*, 26:251-298.

_____ and Spier, Kathryn E. (2003). "A Theory of Utilization Review," *Contributions to Economic Analysis and Policy*, 2(1), article 9.

Dusheiko, Mark; Gravelle, Hugh; Jacobs, Rowena, and Smith, Peter (2004). "The Effect of Budgets on Gatekeeping Doctor Behavior: Evidence from a Natural Experiment," University of York working paper.

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

_____ and McGuire, Thomas G. (1986). "Provider Behavior Under Prospective Reimbursement," *Journal of Health Economics*, 5:129-151.

_____ (1990). "Optimal Payment Systems for Health Services" *Journal of Health Economics*, 9:375-396.

Frank, Richard G., Glazer, Jacob and McGuire, Thomas G. (2000). "Measuring Adverse Selection in Managed Health Care," *Journal of Health Economics*, 19:829-854.

Glazer, Jacob and McGuire, Thomas G. (2000). "Optimal Risk Adjustment in Markets with Adverse Selection: An Application to Managed Care," *American Economic Review*, 90(4):1055-1071.

Jack, William (2004). "Purchasing Health Care Services from Providers with Unknown Altruism," Georgetown University working paper.

Jullien, Bruno (2000). "Participation Constraints in Adverse Selection Models," *Journal of Economic Theory*, 93:1-47.

KEELER, E. B., CARTER, G., AND NEWHOUSE, J.P. (1998). "A Model of the Impact of Reimbursement Schemes on Health Plan Choice," *Journal of Health Economics*, 17(3):297-320.

LAFFONT, JEAN-JACQUES AND TIROLE, JEAN (1986). "Using Cost Observations to Regulation Firms," *Journal of Political Economy*, 94:614-641.

LEWIS, TRACY AND SAPPINGTON, DAVID (1989). "Countervailing Incentives in Agency Problems," *Journal of Economic Theory*, 49:294-313.

MA, CHING-TO ALBERT (1994). "Health Care Payment Systems: Cost and Quality Incentives," *Journal of Economics & Management Strategy*, 3(1):93-112.

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

——— AND MCGUIRE, THOMAS G. (1997). "Optimal Health Insurance and Provider Payment," *American Economic Review*, 87(4):685-704.

——— AND RIORDAN, MICHAEL H. (2002). "Health Insurance, Moral Hazard, and Managed Care," *Journal of Economics & Management Strategy*, 11(1):81-108.

MCGUIRE, THOMAS G. (2000). "Physician Agency," in A.J. Cuyler and J.P. Newhouse, *Handbook of Health Economics*, North-Holland, 467-536.

NEWHOUSE, JOSEPH P. (1970). "Toward a Theory of Nonprofit Institutions: An Economic Model of a Hospital," *American Economic Review*, 60:64-74.

ROCKAFELLAR, R T. (1972). *Convex Analysis*, Princeton Univ. Press, second printing.

ROCHAIX, LISE (1989). "Information Asymmetry and Search in the Market for Physician Services," *Journal of Health Economics*, 8:53-84.

ROCHET, JEAN-CHARLES AND CHONÉ, PHILIPPE (1998). "Ironing, Sweeping and Multidimensional Screening," *Econometrica*, 66(4):783-826.

ROGERSON, WILLIAM P. (1994). "Choice of Treatment Intensity by a Nonprofit Hospital under Prospective Pricing," *Journal of Economics & Management Strategy*, 3(1):7-51.