WORKING PAPERS

N° TSE-572

April 2015

"Health Care Insurance Payment Policy when the Physician and Patient May Collude"

Yaping Wu, David Bardey and Sanxi Li

Toulouse School of Economics

Health Care Insurance Payment Policy when the Physician and Patient May Collude

Yaping Wu* David Bardey† Sanxi Li‡

April 25, 2015

Abstract

This paper analyzes the three-party contracting problem among the payer, the patient and the physician when the patient and the physician may collude to exploit mutually beneficial opportunities. Under the hypothesis that side transfer is ruled out, we analyze the mechanism design problem when the physician and the patient submit the claim to the payer through a reporting game. To induce truth telling by the two agents, the weak collusion-proof insurance payment mechanism is such that it is sufficient that one of them tells the truth. Moreover, we identify trade-offs of a different nature faced by the payer according to whether incentives are placed on the patient or the physician. We also derive the optimal insurance scheme for the patient and the optimal payment for the physician. Moreover, we show that if the payer is able to ask the two parties to report the diagnosis sequentially, the advantage of the veto power of the second agent allows the payer to achieve the first-best outcome.

JEL Code: I18, D82.

Keywords: collusion, falsification, health care insurance, physician payment.

*Southwestern University of Finance and Economics, ChengDu, China. Email: wuyp@swufe.edu.cn

†University of Los Andes, Bogotá, Colombia and Visiting Fellow at Toulouse School of Economics. Email: d.bardey@uniandes.edu.co - Phone: (571) 3324495.

‡Renmin University of China. Email: sanxi@ruc.edu.cn

1 Introduction

The strands of literature that deal with optimal health insurance on the one hand, and the optimal physician reimbursement rule on the other, have proven to be prolific. However, in recent decades medical health care has gradually shifted its emphasis from the disease to the patient. This has resulted in a more egalitarian relationship in which doctor and patient participate in a more balanced way in terms of their relative contributions, as well as the content of their interactions.[1] In particular, a single focus on the patient-payer or physician-payer relationship cannot account for complex contracts among the three parties. The health economics literature reveals that since the physican-patient interaction effectively makes communication between the two parties difficult to observe, the physician and the patient may try to coordinate and manipulate their report(s) to the payer (Alger and Ma [2003], Ma and McGuire [1997], Vaithianathan [2002]).

Both fraud and the abuse of health care programs cost the taxpayer billions of dollars. In 2009, the Centers for Medicare and Medicaid Services (CMS) estimated that overall 7.8 percent of the Medicare fee-for-service claims it paid out ($24.1 billion) did not meet program requirements. Roughly speaking, these claims should not have been paid.[2] Fraud in the Medicare program takes such forms as, but is not limited to, the falsification of Certificates of Medical Necessity (*e.g.*, misrepresenting the diagnosis for the patient to be able to justify the services or equipment furnished), claims involving collusion between a provider and a beneficiary, or between a supplier and a provider, resulting in unwarranted or higher costs or charges to the Medicare program. According to the United States Department of Justice, in many cases court documents allege that patient recruiters, Medicare beneficiaries and other conspirators supply beneficiary information to physicians so that physicians are then able to submit fraudulent billing to Medicare for services that were medically unnecessary or never provided. Such activities drive up health care costs, siphon taxpayer resources, and jeopardize the strength of the Medicare program.[3]

In this article, we explore a multi-agent adverse selection issue in which i) patients differ in severity; ii) the patient and the physician who share the same private information relating to this severity have a potential relationship whereby they can exploit mutual beneficial opportunities from the report transmitted to the payer/insurer and, iii) the physician may undertake an action (effort) that is non-contractable but which is also observable to the patient. Since physicians must be relied

[1] A.M. Van Dulmen (2002).

[2] U.S. Department of Health and Human Services.

[3] Medical fraud has also been documented for many transitional economies and regions, for instance, Bulgaria (Balabanova and McKee, 2002), Uganda (Huntb, 2010), and Taiwan (Chiua, Smithb, Morlockc and Wissowd, 2007).

upon to diagnose and prescribe clinical services, they exert an effort (*e.g.*, time or costly psychological and manual work) that represents an input for the production of health, which influences the patient's valuation of services and therefore affects the patient's decision regarding treatment services.

In the same vein as Ma and McGuire (1997), side transfer is ruled out and we assume that the physician does not lie to the patient and suggests a report of the diagnosis that is not necessarily the true diagnosis. Briefly, if the patient agrees then the suggested diagnosis is reported to the payer; if he disagrees, the true diagnosis is reported. We thus confirm the Ma and McGuire's insight that truthful reports in a claim must be individually rational. Our results show that mutually beneficial opportunities exist between the patient and the physician when joint deviation from truth telling is individually rational (even if side transfer between the two parties is ruled out). Since joint deviation is possible if and only if it is in both parties' self interests, then the collusion-proof insurance payment mechanism is such that it is sufficient that one of them tells the truth. Hence the payer has to decide whether to place incentives on the patient or to place incentives on the physician.

Our paper reveals how the trade-off faced by the payer differs according to whether the payer places incentives on the patient or the physician. When the payer provides incentives to the physician (scenario 1 hereafter), the trade-off is between the efficiency and the informational rent. However, if the payer provides incentives to the patient (scenario 2), his trade-off is between efficiency and risk sharing. Since the payer only cares about the patient's benefit, he does not care about the informational rent of the patient. In order to induce truth telling by the patient, the payer has to reduce the high-severity patient's consumption to make his menu less attractive for the low-severity type. Since this implies imperfect risk sharing for the patient, the high-severity patient is exposed to a higher risk. On the other hand, since the payer cares about the benefit of the patient, an upward distortion on the level of the treatment quantity increases the insurance of the high-severity patient.[4]

Our numerical analysis aims to compare the two *scenarii*. It shows that within the arrangement of interior solutions, when the benefit measurement for the high-severity is relatively small, placing incentives on the physician gives a higher expected utility than when placing incentives on the patient. On the contrary, when the parameter of high severity is relatively large, scenario 2 dominates. Interestingly, in the arrangement where both schemes generate corner solutions, both scenarios yield

[4]By making the insurance policy independent of the physician's effort, the patient can play the role of supervisor in relation to the physician and can be asked to report on the effort of the physician to the payer (Tirole, 1986).

the same expected utility for the patient. However, the numerical analysis also reveals that placing incentives on the physician generates solutions such that either the patient of low severity or that of high severity has zero allocation for treatment and effort. For an insurer who cares about equity and access to health care, this scheme should not be chosen.

Regarding implementation aspects, whatever the scenario at play, the optimal regulation implies that the insurance policy includes an insurance premium and a reimbursement rate based on treatment quantity. On the physician's side, the payment policy can be implemented by a fixed budget per patient, *i.e.* capitation. We derive the optimal marginal tax price on health insurance for different patients and the amount of capitation received by the physician when faced with different types of patient. Under scenario 2, the high-severity patient is subsidized and an implicit redistribution occurs from healthy and low-severity patients to high-severity patients. The physician is fully reimbursed for his cost on treatment and the cost of time spent on the patient (*e.g.*, number of office visits). If the payer places incentives on the physician **(scenario 1)**, the patient gets a zero marginal tax price on health care whatever his severity, whereas the physician is reimbursed more than his total cost when his patient is of low severity.

Finally, we provide a more powerful sequential mechanism that induces truth reporting from the patient and the physician and allows the insurer to achieve the first-best allocation. More precisely, we provide an adaptation of the two-stage subgame perfect mechanism suggested by Moore and Repullo (1988) for the patient-physician interaction issue. Since we assume that the physician does not lie to the patient, the private information of the patient and the physician are perfectly correlated. When side transfer is ruled out a sequential mechanism allows the insurer to take advantage of the veto power of the second agent in order to prevent misreporting. As a result, a two-stage mechanism can uniquely implement the first-best outcome as a subgame-perfect equilibrium.

Literature review

Ma and McGuire (1997) initiate the study of interactions among insurers, physicians, and patients with a focus on the simultaneous derivation of optimal insurance contracts to the consumers and payment contracts to providers. Our model also looks at the optimal policy mix. We follow their article on the reporting game that models the interaction between the physician and the patient leading to the property that misreporting is possible if and only if it is in both agents' self-interest. Our paper differs from Ma and McGuire's paper in three main aspects. First, our paper provides more insight into ways of preventing collusion. Since Ma and McGuire impose restrictions on some parameters in order to ensure that their optimal regulation is collusion-proof, these restrictions work as sufficient conditions, but they may not be necessary. In contrast, we seek for the optimal

regulation that handles distortions in efficiency by optimally placing incentives on the physician and the patient in order to ensure a collusion-proof outcome. In particular, we point out that one-side incentives are sufficient for the insurer to prevent collusive behavior under the same reporting game. Second, we introduce some heterogeneity regarding patient severity, which yields a different reporting game. In particular, Ma and McGuire consider that treatment quantity is not contractible. Due to the development of guidelines between insurers and hospitals (or within hospitals) during the last two decades, we rule out this quantity contractibility aspect in order to focus on the misreporting of the Diagnosis Related Group (DRG). Third, these authors consider linear policies while we allow for non-linear policies (on the patient's insurance copayment), which shows the properties of marginal prices. In practice, copayment rules are often linear, but some have non-linear features such that percentage of coverage by the social security and mutual insurance differs when the intensity of the treatment (*e.g.*, different medications) varies.[5] Such policies are then often piecewise linear. In this sense, our paper allows for a study of the optimal policy in a setting where an optimal general (possibly nonlinear) copayment scheme is also available.

This paper also relates to the collusion literature in health economics and borrows from the multi-agent collusion literature in contract theory. Alger and Ma (2003) consider a model of insurance and collusion in which they find that deterrence against collusion is optimal only if the probability that the provider is collusive is large enough. Vaithianathan (2002) considers a situation in which supply-side cost sharing imposes financial risks on a risk-averse physician, and the superiority of supply-side cost sharing arrangements over demand-side cost sharing is no longer assured. She also argues that when physicians and patients are asymmetrically informed the potential gains from collusion are more liable to become dissipated in informational rent. Thus the recent trend towards improving patient information may increase the cost of supply-side schemes. The above papers consider collusion with side transfers.

Another piece of literature that relates to our paper deals with the collusion issue in the principal-agent framework. Tirole (1986) analyzes a three-tier hierarchy with a supervisor between the principal and the agent. The possibility of collusion affects the efficiency of the organization. Laffont and Martimort (1997) analyze a mechanism design problem in which the agents can communicate among themselves and collude under asymmetric information. They characterize the set of implementable collusion-proof contracts. Quesada (2004) addresses the question of collusion in mechanisms by assuming that one of the colluding parties has all the bargaining power at the collusion stage and offers a side contract to the other. In contrast, the collusion behavior in this present paper follows

[5]See Bardey *et al.* (2015).

the one-shot bargaining process of Ma and McGuire (1997), where the physician proposes a report to the patient who subsequently decides whether to agree or not.

The paper is organized as follows. Section 2 presents the set-up. Section 3 characterizes the first-best solution. Section 4 is devoted to a behavior analysis of the physician and the implementation of the first-best optimum. Section 5 analyzes the collusion-proof mechanism, the characterization of the second-best optimum, and the sequential mechanism. Section 6 concludes the paper.

2 The Model

The model contains two agents and one principal: the agents are the patient and the physician, and the principal is the payer. The patient may suffer from a disease that corresponds to one of the Major Diagnostic Categories (MDC). The physician provides services to the patient and undertakes an effort that affects the patient's utility.

The patient's expected utility $\mathbb{E}U$ depends on his consumption $X \in \mathcal{R}^+$, the treatment quantity $q \in \mathcal{R}^+$, the physician's effort $m \in \mathcal{R}^+$ and the severity α of the disease. The physician's effort can be measured by the time devoted to the patient or manual work.[6] The treatment quantity can be measured by the number of repeated visits to the physician's office. Within this MDC, we consider three levels of severity $\{\alpha_0, \alpha_1, \alpha_2\}$, with $\alpha_2 > \alpha_1 > \alpha_0 = 0$, which occur with probabilities $1 - p_1 - p_2$, p_1 and p_2, respectively.[7] The higher α measures the higher the benefit that the patient may obtain from a certain amount of treatment and physician effort, we refer to $\alpha_0 = 0$ as the situation in which the patient is healthy. The patient with severity α_i suffers a pain K_i, with $K_2 > K_1 > K_0 = 0$, while the physician's intervention can relieve this pain by the amount $\alpha_i v(q_i, m_i)$, but with $K_i \geq \alpha_i v(q_i, m_i)$, $\forall q_i, m_i$. In words, the parameters α_i capture the patient's valuation for receiving treatment and physician's effort. Patients with a higher severity enjoy a higher marginal treatment benefit. On the other hand, the utility is lower under the high severity due to the difference between K_1 and K_2.

We denote y the patient's endowment and we consider a non-linear insurance scheme $(f, T(.))$ where the patient's tax price on health insurance $T(.)$ can only be based on the treatment quantity q.

[6] In the development of the Harvard Resource-based relative value scale (RBRVS), partially used by Medicare in the United States and by nearly all health maintenance organizations (HMOs), physician work includes the physician's time, mental effort, technical skill, judgement, stress and an amortization of the physician's education.

[7] Diagnosis-related group (DRG) classifies disease cases into one of originally 467 groups. DRGs are further grouped into Major Diagnostic Categories (MDCs). Hence within one MDC, there may be one or several DRGs. Its intention is to identify the services that a physician or a hospital provides.

Let $C > 0$ denote the marginal cost of treatment quantity which can be interpreted as the monetary cost per visit. It captures the functioning of the physician office and the practice exercised by all physicians in the discipline. The patient's expected utility $\mathbb{E}U$ is:

$$\begin{aligned}\mathbb{E}U &= p_0 u(y-f) \\ &\quad + p_1[u(y-f-Cq_1-T(q_1)) - (K_1 - \alpha_1 v(q_1, m_1))] \\ &\quad + p_2[u(y-f-Cq_2-T(q_2)) - (K_2 - \alpha_2 v(q_2, m_2))], \end{aligned} \tag{1}$$

where the term $(-Cq_i - T(q_i))$ reflects the demand side cost sharing. The term $-C - T'(.)$ reflects the marginal cost sharing. With zero tax $T'(.) = 0$ the patient bears all the cost for each unit of treatment he consumes; with a subsidy $T'(.) < 0$ (respectively $T'(.) > 0$), the patient pays less (resp. more) than the cost for each unit of treatment he consumes.

The physician maximizes his expected utility $\mathbb{E}V$, which depends on the revenue R_i received from the payer, which is a fixed budget received for each patient treated (capitation), the monetary cost of treatment Cq_i and the disutility of his effort $c_i m_i$, for $i = 1, 2$:

$$\mathbb{E}V = p_1 V(\pi_1) + p_2 V(\pi_2) = p_1 V(R_1 - Cq_1 - c_1 m_1) + p_2 V(R_2 - Cq_2 - c_2 m_2), \tag{2}$$

where $c_i > 0$ is the marginal disutility of the physician's effort.[8] The marginal disutility of effort c_i refers to the marginal cost of the physician's time and manual work, which it is assumed depends on the severity of the disease. The marginal disutility of effort differs from one severity to another and also from one speciality to another. If m denotes the consultation length, the marginal disutility of one additional unit of time spent on the patient is the stress or the psychological cost borne by the physician.[9] Accordingly, we restrict our attention to the separability of the effort and the treatment in the cost structure. Moreover, it is quite intuitive to assume that the marginal cost of effort is positively correlated with the severity of the disease. To simplify, we assume that they are perfectly correlated: $c_1/c_2 = \alpha_1/\alpha_2$.

Following Ma and McGuire (1997) and Ma and Alger (2003), we consider a risk-neutral payer who it is assumed operates in a competitive market and sets the policy to maximize the patient's expected utility $\mathbb{E}U$ defined in equation (1). Similarly, this risk-neutral payer could also be a public

[8] For the purpose of simplicity, we consider the financial equivalence of the psychological cost $c_i m_i$. If we consider the psychological cost outside the utility function $V(.)$, it does not qualitatively change our results.

[9] In a study realized by CREDES dealing with professional expenses of liberal physicians in France, each office visit consists of a number of physician working points such as the duration of consultation, the stress, the technical ability and psychological effort. These visits are classified in hierarchy relative to each other according to the quantity of physician working points in each discipline (intra-speciality) and then between the disciplines (inter-speciality).

regulator who only cares about the benefit of the patient as long as we assume a shadow cost of public fund, *i.e.* all rents paid to providers are costly. In the first-best environment we assume that both the type of severity and the physician effort are observable and verifiable by the payer. In the second-best environment we consider a case in which neither the type of severity nor the physician effort are observable, while the treatment quantity is always observable. We are looking for the optimal second-best collusion-proof mechanism when collusive behavior between the physician and the patient is taken into account. Due to the *Hippocratic Oath*, we assume that the physician does not lie to the patient about his true state of nature.

The timing of the game is as follows. In the first stage the payer sets the payment policies. Nature then decides the severity of the patient's condition. If the patient is healthy the game ends, otherwise he consults the physician. In the second stage the physician exerts effort. In the third stage, after observing the physician's effort, the patient chooses the treatment quantity.[10] Finally, the reimbursement for the patient and the payment to the physician are implemented.

3 The first-best optimum

At the first-best benchmark both the type of severity and physician effort are observable and verifiable. The problem of the payer can be written:[11]

$$\max_{X_0,X_1,q_1,m_1,X_2,q_2,m_2} p_0u(X_0) + p_1[u(X_1) - (K_1 - \alpha_1 v(q_1, m_1))] + p_2[u(X_2) - (K_2 - \alpha_2 v(q_2, m_2))] \quad (3)$$

s.t

$$p_0 f + p_1(f + Cq_1 + T(q_1)) + p_2(f + Cq_2 + T(q_2)) = p_1R_1 + p_2R_2, \qquad \lambda_0, \quad (4)$$

$$\pi_i \geq 0, \quad \forall i. \quad (5)$$

The budget constraint (4) requires that the money paid to the physician does not exceed the money received from the patient. Since $X_0 = y - f$, $X_i = y - f - Cq_i - T(q_i)$ for $i = 1, 2$, and the payer

[10] We may interpret this timing as a reduced form of a more complex compliance game in which the physician chooses the effort and the quantity and the patient the degree of compliance. Under this interpretation, the degree of compliance yields to the real level of treatment consumed.

[11] We consider the *ex post* participation constraints which guarantee that the physician does not obtain a negative profit whatever the severity of the disease. If instead we consider an *ex ante* participation constraint, in the second-best analysis the first-best optimum can be achieved without any informational rent being paid to the physician if the payer places incentives on the physician. Since an *ex ante* participation constraint would induce some patient selection from providers, we restrict our attention to the *ex post* participation constraints.

does not leave any rent to the physician when everything is observable, the budget constraint can be rewritten such that:

$$p_0(y - X_0) + p_1(y - X_1) + p_2(y - X_2) = p_1(Cq_1 + c_1m_1) + p_2(Cq_2 + c_2m_2). \qquad \lambda_0 \qquad (6)$$

The first-order conditions can be found in the appendix. For the first-best optimum it is worth noticing that the patient is fully insured. He consumes identical consumption levels across all states of nature $X_1 = X_2 = X_0$. Moreover, for each type of severity the marginal rates of substitution between the three allocations X, q, and m are equal to the ratio of their marginal costs in each state:

$$MRS^i_{X,q} = C, \qquad MRS^i_{q,m} = \frac{v_{q_i}(q_i, m_i)}{v_{m_i}(q_i, m_i)} = \frac{C}{c_i} \quad \text{for } i = \{1, 2\}. \qquad (7)$$

As pointed out in Ma and McGuire (1997), the classification of substitutes and complements between effort and treatment quantity is crucial. The characterization of the first-best optimum depends on whether the effort is a complement to or a substitute for the treatment, or how it interacts when the patient chooses a treatment. Under the assumption that the marginal cost of effort is perfectly correlated with the severity of the disease $c_1/c_2 = \alpha_1/\alpha_2$, the candidates of the first-best solution are summarized in Table 1. The proof is given in the appendix.

Table 1: Characterization of the first-best solution

the sign of v_{qm}	the first-best solution
$v_{qm} = 0$	$q_1^* < q_2^*, m_1^* = m_2^*$
$v_{qm} > 0$	$q_1^* < q_2^*, m_1^* < m_2^*$ or $q_1^* > q_2^*, m_1^* > m_2^*$
$v_{qm} < 0$	$q_1^* > q_2^*, m_1^* < m_2^*$ or $q_1^* < q_2^*, m_1^* > m_2^*$

When the effort and the treatment are separable in the benefit function, *i.e.* $v_{qm} = 0$, the high-severity patient consumes a higher level of treatment quantity while both types obtain the same level of physician effort. This comes from the fact that the marginal cost of treatment is independent of the severity. A higher marginal benefit of treatment ($\alpha_2 > \alpha_1$) implies a higher level of treatment. However, the marginal cost of effort is strictly positively correlated with severity. If the marginal benefit of effort is higher ($\alpha_2 > \alpha_1$) then so is the marginal cost of effort ($c_2 > c_1$). Hence the high-severity patient consumes a higher level of treatment quantity but benefits from the same level of effort as the low-severity type.

When the effort and the treatment are no longer separable, the level of effort affects the marginal benefit of treatment. When the effort and the treatment are complementary, *i.e.* $v_{qm} > 0$, if the

cross second derivative v_{qm} is positive but sufficiently small compared to the own second derivative v_{qq}, we must have $q_1 < q_2$ and $m_1 < m_2$. If the cross second derivative v_{qm} is positive but large enough to have a strong effect on the marginal benefit of treatment v_q, we obtain $q_1 > q_2$ and $m_1 > m_2$. On the contrary, when the effort and treatment are substitutive, *i.e.* $v_{qm} < 0$, which of the allocations is higher for the higher severity type becomes ambiguous. This depends on the curvature of the benefit function $v(.)$.

Whether the effort and the treatment quantity are complements or substitutes depends on the nature of the disease or on the stage of the physician's intervention (Ma and McGuire, 1997). In practice, both cases are possible according to the type of disease considered. For instance, acute care often exhibits substitution between physician effort and treatment quantity, while chronic care normally exhibits a complementary relationship between these two inputs. In the same spirit, acute appendicitis requires more physician effort than treatment quantity. If the diagnosis of the physician is accurate and if the preparation for the operation is adequate, an operation of 20 minutes is sufficient to cure the patient. In such a case, no more treatment quantity is needed. Hence physician effort and treatment quantity exhibit substitutability between one another. However, chronic diseases such as chronic bronchitis or a stroke require physicians to make permanent efforts to analyze the progression of the disease in patients. According to continuous analysis and supervision, the physician continously needs to prescribe the appropriate treatment for the patient. Through physician intervention the treatment may continuously be prescribed for the remaining lifespan of the patient. Hence in this case physician effort and treatment quantity exhibit complementarity.

4 Implementation of the first-best optimum

Knowing his state of nature and having an *ex post* observation of physician effort, the patient's program becomes:

$$\max_{q_i} u(y - f - Cq_i - T(q_i)) - (K_i - \alpha_i v(q_i, m_i)). \tag{8}$$

The first-order condition yields:

$$C + T'(q_i) = \alpha_i \frac{v_{q_i}}{u_{X_i}} \equiv MRS^i_{X,q}, \tag{9}$$

where $T'(q_i)$ is the marginal tax price on health insurance, and $C+T'(q_i)$ is the marginal demand side cost sharing. The term $MRS^i_{X,q}$ denotes the marginal rate of substitution between the consumption and the treatment evaluated at point (X, q). The patient selects the treatment quantity such that

the marginal rate of substitution between consumption and treatment is equal to the marginal demand-side cost sharing (the price of consumption being normalized to 1).

Combining equations (9) and (7), we obtain the following marginal demand-side cost sharing**:**

$$C + T'(q_i) = C, \quad \text{for} \quad i = \{1, 2\}. \tag{10}$$

The first-best treatment and consumption levels can be decentralized through a non-distortionary fee-for-service insurance scheme with zero tax on health care $T'(q_i) = 0$. Since in the first-best optimum effort is observable, the payer leaves no rent to the physician of each type:

$$R_i^* = Cq_i^* + c_i m_i^* \quad \text{for} \quad i = \{1, 2\}, \tag{11}$$

and imposes a hefty punishment in the case where the *ex post* verifiable effort is not the optimal one. Hence in the first-best outcome the patient pays according to the treatment he has obtained, whereas supply side cost sharing, *i.e.* the total cost is reimbursed to the physician in its entirely.

Moreover, the resource constraint requires that the insurance premium is such that:

$$f^* + p_1 T(q_1^*) + p_2 T(q_2^*) = p_1 c_1 m_1^* + p_2 c_2 m_2^*. \tag{12}$$

Since in the first-best outcome the patient pays for the treatment quantity, the insurance premium plus the patient's expected tax price on health insurance must pay for the physician's expected cost of effort.

The first-best solution and its decentralization have been derived under the assumption that the payer observes patient severity and the effort exerted by the physician. Under the more realistic scenario where information is asymmetric, the first-best insurance and payment schemes are generally not feasible. Moreover, the interaction between the patient and the physician makes collusive behavior possible. In the next section we consider the second-best environment where neither the severity nor the effort are observable by the payer. We analyze the collusion-proof mechanism, characterize the second-best optimum and derive the second-best optimal policy.

5 The second-best

In this section we analyze the second-best optimal mechanism for when neither the type nor the effort are observable by the payer. Since the physician and the patient share the same information regarding the patient severity and on the other, that the patient observes the physician's effort, the asymmetric information refers to the information structure in relation to the payer. We study

the centralized message game where the patient and physician have to submit the diagnosis to the payer through a certain process. Since the physican-patient interaction effectively makes the communication between the physician and the patient feasible and hard to observe, they may try to coordinate themselves in order to manipulate their report(s) to the payer.

5.1 Weak collusion-proof mechanism

5.1.1 The mechanism design

Similar to Ma and McGuire (1997), we first consider a modeling of such collusive behavior as a one-shot bargaining process when side transfer is not feasible between the physician and the patient. The timing of the payer's mechanism, as well as the collusion formation, is as follows:

0. The payer proposes the mechanism $\mathcal{M}$: $\{X(\hat{\alpha}), q(\hat{\alpha}), m(\hat{\alpha}), T(\hat{\alpha}), R(\hat{\alpha})\}$, where $\hat{\alpha}$ is the physician and the patient's joint report on their private information to the payer, specifying for each type of report the consumption to be taken, the treatment to be produced, the effort to be exerted, as well as the patient's tax price on health insurance and the payment to the physician.

1. Nature determines the value of the patient's severity. If he is healthy then the next steps in the reporting game do not occur, otherwise the game goes to the next step.

2. The physician first suggests to the patient a report of severity $\hat{\alpha}$ which is not necessarily equal to the true severity α. If the patient agrees, $\hat{\alpha}$ is reported to the payer. If he disagrees, the true severity is reported.

3. The allocations and the monetary transfers defined by the mechanism $\mathcal{M}$ are enforced.

4. The patient reports the physician's effort $\hat{m}$ to the payer. If $\hat{m} = m(\hat{\alpha})$, no punishment is required for the physician, otherwise, a large punishment is enforced on the physician.

We consider the weak collusion-proof mechanism where side transfer is ruled out. From the revelation principle, there is no loss of generality in restricting the set of contracts to a direct revelation mechanism. When side transfer is ruled out, this mechanism induces truth telling by the patient concerning physician effort. The patient has no incentive to misreport physician effort because his price of treatment does not depend on the reported physician effort $\hat{m}$.

At the collusion formation stage of the game either the physician or the patient can reveal the true type if they wish. The physician can always propose a reporting of the true severity; the patient can reveal the true severity by disagreeing with a nontruthful suggestion. Hence collusion (misreporting) is possible if and only if it is in the self-interests of both parties. In other words, if one of them does not have any incentive to misreport then collusion will be impossible. Thus the

weak collusion-proof mechanism is such that it is sufficient that one of them tells the truth. We summarize the second-best weak collusion-proof mechanism in the following lemma:

Lemma 1 *When joint deviation is possible through collusion, the weak collusion-proof incentive compatible insurance payment scheme requires truth telling from only one of the agents.*

First, consider the case in which the first-best solution is such that $q_1^* > q_2^*$ and $m_1^* > m_2^*$.[12] If the true state is of severity α_2 then the physician never suggests misreporting because by mimicking severity α_1 he obtains a negative profit since $c_2 > c_1$:

$$R_1^* - Cq_1^* - c_1 m_1^* = 0,$$
$$R_1^* - Cq_1^* - c_2 m_1^* < 0.$$

The patient characterized by severity α_2 does have an incentive to misreport since he gains a higher level of treatment and physician effort by paying the same price $C + T'(q_i^*) = C$, for $i = 1, 2$. However, when they disagree the true severity is reported even if the patient wants to misreport. If the true state is severity α_1 then the physician has an incentive to misreport because he is then able to obtain a positive profit:

$$R_2^* - Cq_2^* - c_2 m_2^* = 0,$$
$$R_2^* - Cq_2^* - c_1 m_2^* > 0.$$

However, the patient disagrees since he does not want to obtain a lower allocation but still has to pay the same price $C + T'(q_i^*) = C$, for $i = 1, 2$. As a result, the true severity is reported. Consequently, when the first-best solution is such that $q_1^* > q_2^*$ and $m_1^* > m_2^*$, collusion is impossible because of the conflict of interest between the two agents. In such a case, these is no need to design a collusion-proof mechanism since the true state is always reported. The first-best solution can be achieved even in the presence of asymmetric information. The marginal price remains undistorted for the patient, whatever his severity.

In the case where the first-best solution is such that $q_1^* < q_2^*$ and $m_1^* < m_2^*$, if the patient is characterized by severity α_2 then the physician never suggests misreporting and the patient never disagrees. But if the patient is characterized by α_1 the physician wants to misreport and the patient always agrees. Accordingly, joint deviation by collusion is possible. For a patient with severity α_1 the physician proposes a false report and it is in the patient's interest to agree. Consequently, in

[12] Such a case occurs when $v_{qm} \geq 0$.

such a case the insurance and the payment policy are based on this false report. Therefore, in order to ensure a truthful report the policy must be weakly collusion-proof incentive compatible.

Remark: *When physician's effort and treatment quantity are complement, whatever the first-best solution, the physician and the patient have incentives to collude only when the patient's severity is* α_1.

Finally, if the first-best solution is such that $q_1^* < q_2^*$ and $m_1^* > m_2^*$ or $q_1^* > q_2^*$ and $m_1^* < m_2^*$, although it is certain that the physician wants to misreport (respectively, never wants to misreport) if the patient is characterized by severity α_1 (resp. α_2), in both cases, the direction of the binding incentive constraints for the patient remains ambiguous.

5.1.2 The second-best optimum and its implementation

When the first-best solution is such that $q_1^* < q_2^*$ and $m_1^* < m_2^*$, collusion is possible when the patient is characterized by severity α_1. The payer's program is:

$$\max_{X_0,X_1,R_1,q_1,m_1,X_2,R_2,q_2,m_2} p_0u(X_0) + p_1[u(X_1) - (K_1 - \alpha_1 v(q_1, m_1))] + p_2[u(X_2) - (K_2 - \alpha_2 v(q_2, m_2))] \tag{13}$$

s.t

$$(RC), \quad p_0(y - X_0) + p_1(y - X_1) + p_2(y - X_2) = p_1R_1 + p_2R_2, \qquad \lambda_0, \tag{14}$$

$$(IC_1 \quad patient), \quad u(X_1) - (K_1 - \alpha_1 v(q_1, m_1)) \geq u(X_2) - (K_1 - \alpha_1 v(q_2, m_2)), \qquad \lambda_1, \tag{15}$$

or

$$(IC_1 \quad physician), \quad R_1 - Cq_1 - c_1m_1 \geq R_2 - Cq_2 - c_1m_2, \qquad \lambda_1, \tag{16}$$

$$(PC_2), \quad R_2 - Cq_2 - c_2m_2 \geq 0, \qquad \lambda_2, \tag{17}$$

$$(PC_1), \quad R_1 - Cq_1 - c_1m_1 \geq 0, \qquad \lambda_3. \tag{18}$$

In the following we characterize the second-best solution, in each case with one (the patient's or the physician's) relevant incentive constraint. When incentives are placed on the patient, the payer's program is (13) subject to constraints (14), (15), (17) and (18). Rearranging the first-order conditions (see appendix) we obtain:

$$MRS^1_{X,q} = \alpha_1 \frac{v_{q_1}}{u_{X_1}} = C, \qquad MRS^1_{q,m} = \frac{v_{q_1}(q_1, m_1)}{v_{m_1}(q_1, m_1)} = \frac{C}{c_1}, \tag{19}$$

$$MRS^2_{X,q} = \frac{[p_2 - \lambda_1]C}{p_2 - \lambda_1 \frac{\widetilde{MRS^1}}{MRS^2}} < C, \qquad MRS^2_{q,m} = \frac{v_{q_2}(q_2, m_2)}{v_{m_2}(q_2, m_2)} = \frac{C}{c_2}. \tag{20}$$

We obtain the usual "no distortion at the top" for the low-severity patient (α_1). Since $\alpha_2 > \alpha_1$, a patient with α_2 is willing to sacrifice greater consumption in order to obtain an additional unit of treatment in comparison to a patient with severity α_1. Hence, when the low-severity type mimics the high-severity type by taking his allocations, it follows that $\left(\widetilde{MRS^1}/MRS^2\right) < 1$. Consequently, the high-severity type's marginal rate of substitution between consumption and treatment is downward distorted. Compared to the first-best trade-off, for the same level of physician effort m_2, the α_2-patient type consumes a greater treatment quantity q_2 and a lower consumption X_2.

Since the payer only cares about the patient's benefit, he does not care about the informational rent of the patient. In order to induce truth telling by the patient, the payer has to reduce the high-severity patient's consumption X_2 to make his menu less attractive for the low-severity type. But this implies imperfect risk sharing for the patient. The high-severity patient is then exposed to a higher risk. Since the payer cares about the benefit of the patient, an upward distortion on the level of the treatment quantity q_2 increases the insurance of the high-severity patient. Actually, when the payer provides incentives to the patient, his trade-off is between efficiency and insurance. Starting from the first-best tradeoff, for a same level of physician effort m_2, a variation $dX_2 < 0$ along with a variation $dq_2 > 0$ is a way to relax the otherwise binding incentive compatibility constraint of the low-severity patient and to guarantee insurance for the high-severity patient at the same time.

Furthermore, the first-order conditions in the appendix also imply that $\lambda_2 > 0$ and $\lambda_3 > 0$, which suggests that we do not give any rent to both types of physician. They obtain zero profit in both states.

Combining equation (9) with equations (19) and (20) and the first-order conditions (45), (46), (47), (48), (49) and (50), (51), (52), (53) given in the appendix, we derive the optimal insurance payment policy when the payer places incentives on the patient:

$$C + T'(q_1^{SB}) = C, \tag{21}$$

$$C + T'(q_2^{SB}) < C, \tag{22}$$

$$R_1 = Cq_1^{SB} + c_1 m_1^{SB}, \tag{23}$$

$$R_2 = Cq_2^{SB} + c_2 m_2^{SB}. \tag{24}$$

There is no distortionary tax on the low-severity patient $T'(q_1) = 0$. But $C + T'(q_2) < C$ implies a subsidy for the treatment of the high-severity patient $T'(q_2) < 0$. Truth telling implies imperfect

risk sharing for the patients. In order to balance the trade-off between insurance and efficiency, the high-severity patient is subsidized in terms of health care in order to encourage the consumption of health care treatment. Since no informational rent for the physician implies no supply-side cost sharing and full reimbursement for the physician, an implicit redistribution occurs from healthy and low-severity patients to high-severity patients.

We now turn to the case in which the payer places incentives on the physician. The payer's program is now (13) subject to constraints (14), (16) and (17). Rearranging the first-order conditions given in the appendix we obtain

$$MRS^1_{X,q} = \alpha_1 \frac{v_{q_1}}{u_{X_1}} = C, \qquad MRS^1_{q,m} = \frac{v_{q_1}(q_1, m_1)}{v_{m_1}(q_1, m_1)} = \frac{C}{c_1}, \tag{25}$$

$$MRS^2_{X,q} = C, \qquad MRS^2_{q,m} = \frac{v_{q_2}(q_2, m_2)}{v_{m_2}(q_2, m_2)} = \frac{C}{c_2 \frac{\lambda_0 p_2 + \lambda_1 - \lambda_1 \frac{c_1}{c_2}}{\lambda_0 p_2}} < \frac{C}{c_2}. \tag{26}$$

In this mechanism both type of patients' marginal rate of substitution between treatment and consumption remain undistorted, but the high-severity type's marginal rate of substitution between treatment and effort is distorted. This distortion comes from the fact that the physician's incentive constraint, when his patient is α_1, is relevant and λ_1 is positive. In such a case, the patient is no left any rent, but a positive rent is given to the physician. Compared to the first-best trade-off, for the same level of consumption X_2, the high-severity type patient obtains a greater treatment quantity q_2 but benefits from a lower physician effort m_2.

When the payer places incentives on the physician, we have the usual trade-off between the efficiency and the informational rent. The informational rent of the physician when his patient is α_1 depends positively on the effort m_2 that he would have exerted when his patient would have been α_2. Starting from the first-best trade-off, for the same level of consumption X_2, a variation $dm_2 < 0$ and a variation $dq_2 > 0$ have no first-order effect on efficiency, but they decrease the profit of the mimicking physician and hence decrease the informational rent given to the mimicking physician. Consequently, for the same level of X_2, a downward distortion in m_2 along with an upward distortion in q_2 presents a way to relax the otherwise binding incentive compatibility constraint of the physician when his patient is α_1.

Combining equation (9) with equations (25) and (26) and first-order conditions (54), (55), (56), (57), (58) and (59), (60), (61), (62) given in the appendix, we derive the optimal insurance payment

policy when the payer places incentives on the physician:

$$C + T'(q_i^{SB}) = C, \text{for} \quad i = \{1, 2\}, \tag{27}$$

$$R_2 = Cq_2^{SB} + c_2 m_2^{SB}, \tag{28}$$

$$R_1 = (c_2 - c_1)m_2^{SB} + Cq_1^{SB} + c_1 m_1^{SB}. \tag{29}$$

Whatever the severity, the patient gets a zero marginal tax price on health insurance. Informational rent requires a positive profit for the physician when the patient is α_1.

We summarize the characterization of the second-best incentive mechanism in the following proposition:

Proposition 1 *i) The payer faces different trade-offs when he provides incentives to the physician and the patient. If the payer places incentives on the physician, the trade-off is as usual between the efficiency and the informational rent. However, if the payer places incentives on the patient, his trade-off is between efficiency and insurance.*
ii) If the payer places incentives on the patient, he leaves no rent to the physician whatever his patient's type but encourages health care consumption of the high-severity patient through a subsidy. If the payer places incentives on the physician, he does not distort the patients' prices but leaves a positive rent to the physician when his patient is characterized by low severity.

5.2 Numerical Analysis

In this section, a numerical analysis is conducted by taking utility functions and parameters[13] as follows

$$u(X_i) = -e^{-X_i} \quad \text{for} \quad i = 0, 1, 2,$$
$$v(q_i, m_i) = q_i^{0.5} m_i^{0.5} \quad \text{for} \quad i = 1, 2$$

Table 2: Basic Parameters

$p_0 = 0.6$	$C = 1$	$y = 50$
$\alpha_1 = 1$	$c_1 = 1$	
$K_1 = 0.1$	$K_2 = 0.2$	

At the first step, we vary the parameters α_2 (and c_2 which is assumed to be perfectly correlated with α_2) to show the change of allocations and expected utility respectively under the two schemes.

[13]Similar patterns are obtained with different values.

Then, keeping p_0 constant, we vary p_1 and p_2. At the second step, we compare the expected utilities under the two schemes on a same figure. Then, keeping p_0 constant, we vary p_1 and p_2. The results are shown in the three groups of figures after the appendix.

The first group of figures corresponds to the first scenario, *i.e.* when incentives are placed on the patient. We exhibit that the patient's expected utility decreases when α_2 becomes higher relative to α_1 until corner solutions appear for the α_2 patient ($q_2 = 0$ and $m_2 = 0$). When the α_2 patient's corner solutions appear, the expected utility exhibits a kink. Within the arrangement of interior solutions, as α_2 increases, the treatment quantity for type 2 increases, the effort level for type 2 decreases and, the treatment quantity for type 1 decreases.

The second group of figures that corresponds to the second scenario, *i.e.* when incentives are placed on the physician, shows a similar pattern. The patient's expected utility decreases when α_2 becomes higher relative to α_1 until corner solutions appear for the type-2 patient ($q_2 = 0$ and $m_2 = 0$). However, under this second scenario, when type-2 has interior solutions, type-1 has corner solutions ($q_1 = 0$ and $m_1 = 0$). That is, either the type-1 patient or the type-2 patient has zero allocation on treatment and effort. Roughly, for an insurer who cares about equity and the access to health care, this scheme should not be chosen.

We then compare the two scenarios. In the third group of figures, the red lines stand for the expected utility under the first scenario while the blue lines stand for the expected utility under the second scenario. We focus on the arrangement of interior solutions (before the first kink of the two lines appears). The expected utilities of the patient under the two scenarios both decrease (not strictly) since α_2 increases relative to α_1. For low values of α_2, the second scenario dominates the first scenario. As α_2 increases, at some point, the two lines cross and the first scenario dominates the second. This means that when α_2 is relatively small, placing incentives on the physician gives a higher expected utility than placing incentives on the patient. On the contrary, when α_2 is relatively large, it is more efficient to place incentives on the patient. Moreover, when the proportion of the high-severity patient increases, the point of intersection becomes closer to α_1 (which is set to be equal to 1). When the proportion of the high-severity patient becomes equal to or greater than the proportion of the low-severity patient, we only obtain corner solutions and the two schemes give the same expected utility (the value on the vertical axe is around that of a unique value).

5.3 Subgame perfect mechanism

If the first-best solution is such that $q_1^* < q_2^*$ and $m_1^* < m_2^*$, we see that when the patient is characterized by a severity α_1, both the patient and the physician want to report a severity α_2.

However, since it is assumed that the physician tells the truth to the patient, the patient's and the physician's private information about the type are perfectly correlated. In the vein of Moore and Repullo (1988), we show that if the payer can ask the patient and the physician to report the severity then the following stage mechanism allows the insurer to implement the first-best outcome as a subgame perfect equilibrium:

(a) if the physician announces that the patient is characterized by α_1;

(b) the patient "agrees", in which case the allocation $A_1 = [q_1^*, m_1^*, R_1^*, X_1^*]$ is implemented; or "challenges", that is, announces that the severity is α_2;

(c) when challenged, the allocation $A_z = [q_z, m_z, R_z, X_z]$ is implemented;

(a') if the physician announces the severity α_2;

(b') the patient "agrees", in which case the allocation $A_2 = [q_2^*, m_2^*, R_2^*, X_2^*]$ is implemented; or "challenges", that is, announces that the severity is α_1;

(c') when challenged, the allocation $A_x = [q_x, m_x, R_x, X_x]$ is implemented.

The allocations A_1, A_z, A_2, A_x satisfy the following incentive constraints:

$$A_1 \succeq^{pat_1} A_z: \quad u(x_1) + \alpha_1 v(q_1, m_1) \geq u(x_z) + \alpha_1 v(q_z, m_z), \tag{30}$$

$$A_x \succeq^{pat_1} A_2: \quad u(x_x) + \alpha_1 v(q_x, m_x) \geq u(x_2) + \alpha_1 v(q_2, m_2), \tag{31}$$

$$A_1 \succeq^{phy_1} A_x: \quad R_1 - Cq_1 - c_1 m_1 \geq R_x - Cq_x - c_1 m_x, \tag{32}$$

$$A_2 \succeq^{pat_2} A_x: \quad u(x_2) + \alpha_2 v(q_2, m_2) \geq u(x_x) + \alpha_2 v(q_x, m_x), \tag{33}$$

$$A_z \succeq^{pat_2} A_1: \quad u(x_z) + \alpha_2 v(q_z, m_z) \geq u(x_1) + \alpha_2 v(q_1, m_1), \tag{34}$$

$$A_2 \succeq^{phy_2} A_z: \quad R_2 - Cq_2 - c_2 m_2 \geq R_z - Cq_z - c_2 m_z. \tag{35}$$

In this mechanism the reporting message to the insurer's mechanism belongs to the set $\{(\alpha_1, \alpha_1), (\alpha_1, \alpha_2), (\alpha_2, \alpha_2), (\alpha_2, \alpha_1)\}$, where the first term in a message represents the physician's report and the second term represents the patient's report.

The above constraints (30), (31), (32), (33), (34) and (35) ensure the truth telling of both the physician and the patient. Together, equations (31) and (32) mean that when the patient is characterized by severity α_1, the physician wants A_2, but the patient prefers A_x to A_2, so will challenge to get A_x, but the physician dislikes A_x, hence the physician will tell the truth. Similarly, together equations (34) and (35) mean that if the patient is characterized by severity α_1, the physician wants A_1, since the patient prefers A_z to A_1, he will challenge to get A_z, but the physician dislikes A_z, hence the physician will not lie. Equations (30) and (33) ensure that when the physician tells the truth the patient prefers to agree rather than to challenge.

By choosing $X_z = X_1^*$, $q_z = q_1^*$, $m_z = m_1^*$, $X_x = X_2^*$, $q_x = q_2^*$, $m_x = m_2^*$, it is obvious that constraints (30), (31), (33) and (34) are all binding for the first-best solution. In order to satisfy the physician's two constraints, *i.e.* (32) and (35) are satisfied, the off-equilibrium reimbursement R_x and R_z can be set to zero, so that the off-equilibrium profit is negative and thus less than the first-best profit (zero). Accordingly, the physician will prefer to tell the truth in both states. Consequently, under the assumption that the physician does not lie to the patient, so that their private information is perfectly correlated, the first-best solution can be implemented as the unique equilibrium by the above mechanism. Our finding is summarized in the following proposition:

Proposition 2 *Asking both the physician and the patient to report sequentially is strictly better than asking only one of them to report. By asking the two agents to report the diagnosis sequentially, the advantage of the veto power of the second agent allows the payer to achieve the first-best outcome.*

However, in practice this form of stage mechanism is not used due to high administrative costs. Both agents would have to report the information and the payer analyze the two reports.[14] Furthermore, although this subgame perfect mechanism is highly powerful when side transfer is ruled out, it must be said that it is not robust to collusion with side transfer. Suppose that, first, the payer proposes this subgame perfect mechanism and then the physician negotiates with the patient that: I always report severity α_2, and you always agree, and I give you a positive side transfer. For the allocations of this mechanism the preferences of the physician and patient for the allocations are as follows:

$$A_2 \succeq^{phy_1} A_1 \succeq^{phy_1} A_x, \tag{36}$$

$$A_2 \succeq^{phy_2} A_z \succeq^{phy_2} A_x, \tag{37}$$

$$A_x \simeq^{pat_1} A_2, \tag{38}$$

$$A_2 \simeq^{pat_2} A_x. \tag{39}$$

If the severity is α_1, the physician is better off taking A_2 while the patient is indifferent between A_x and A_2. He knows that if he colludes with the physician and agrees he will get A_2; if he does

[14] A mechanism such that the physician is always asked to report but the patient is asked to report with a certain probability saves some administrative cost for the payer. But since it requires more incentive compatibility constraints than the second-best program when the payer asks only the physician to report, it cannot do better than the second-best mechanism. Moreover, we do not consider stochastic contract with stochastic allocation and payment because it is not optimal according to the type of benefit function $\alpha_i v(q_i, m_i)$. Furthermore, the enforcement of the stochastic mechanism is problematic in that it requires the randomization of the contracts to be verifiable by a court.

not collude, and hence disagrees, he will get A_x. He is indifferent between the two options, but if the physician gives him a positive side transfer he will be better off colluding with him and agreeing. Moreover, since the physician is better off taking A_2, he will be able to pay a positive side transfer while still being better off than without colluding. Hence collusion and side transfer make, for severity α_1, the physician and the patient both better off. Note that collusive behavior is formalized after the Nature has decided upon the severity. Following this, the physician can decide on the amount of side transfer after observing the realized type. Consequently, this subgame perfect mechanism is not strongly collusion proof when side transfer is feasible.

6 Conclusion

In this paper we explore the three-party contracting problem when the patient and physician are able to exploit mutually beneficial opportunities without being directly observed or regulated by the payer. This problem is relevant since the activities of falsification of diagnosis and collusion between a physician and a patient siphon taxpayer resources, drive up health care costs, and jeopardize the strength of health care financing programs. Following Ma and McGuire (1997), we analyze the mechanism design problem in a richer environment where the physician and the patient submit the claim to the payer through a reporting game. We also derive the optimal insurance payment prices for different types of patient and physician.

First, our results confirm that truthful reports in a claim must be individually rational. When there are mutually beneficial opportunities between the patient and the physician joint deviation from truth telling is individually rational, even if side transfer is ruled out by the two parties. Since joint deviation is possible if and only if it is in both of their self-interests, the weak collusion-proof insurance payment scheme requires truth telling from only one of the agents.

The payer's trade-offs are different when he chooses different manners of providing incentives to the patient and the physician. If the payer provides incentives to the physician, we face the usual trade-off between the efficiency and the informational rent. However, if the payer places incentives on the patient, his trade-off is between efficiency and insurance. Since the payer only cares about the patient's benefit, he does not care about the informational rent of the patient. In order to induce truth telling from the patient, the payer has to reduce the high-severity patient's consumption to make his menu less attractive for the low-severity type. But this implies imperfect risk sharing for the patient. The high-severity patient is then exposed to a higher risk. Since the payer cares about the benefit of the patient, an upward distortion in the level of treatment quantity increases the

insurance of the high-severity patient. We derive optimal marginal tax prices on health insurance for different patients. We show that the high-severity patient is subsidized. Following this, there occurs an implicit redistribution from healthy and low-severity patients to high-severity patients.

Moreover, we apply a two-stage subgame perfect mechanism for health economics, namely the mechanism suggested by Moore and Repullo (1988). We show that when side transfer is ruled out, if the payer can ask the two agents to report sequentially then a two-stage mechanism is able to uniquely implement the first-best outcome as a subgame-perfect equilibrium.

Several directions for extension comprise part of our research agenda. First, when the Hippocratic Oath is assumed away, it is likely that some patient would desire to obtain the true information of the severity by visiting a second doctor in order to cross check the diagnostic established by the first doctor. Emons (1997) proves that market equilibria inducing nonfraudulent behavior do indeed exist. Since the physician's market is usually assumed to follow a monopolistic competition market structure, it would be interesting to explore how the Emons' results hold in such a monopolistic competitive environment. Second, if the patient cannot observe the physician's effort but only a correlated signal, then the moral hazard problem must be taken into account. Third, useful insights could be derived by studying the optimal approach to delegation: decentralized contracts *versus* centralized contracts, as in Macho-Stadler and Perez-Castrillo (1993), Laffont and Martimort (1996), and Sandeep and Tomas (1998).

References

[1] Alger I. and C.-t.A. Ma, 2003, Moral hazard, insurance, and some collusion, Journal of Economic Behavior and Organization, 50, 225-247.

[2] Balabanova D. and M. McKee, 2002, Understanding informal payments for health care: example of Bulgaria, Health Policy, 62, 3, 243-273.

[3] Bardey D., Cremer H. and J-M Lozachmeur, 2015, The Design of Insurance Coverage for Medical Products under Imperfect Competition, IZA DP No. 8815.

[4] Chiu YC., KC. Smith, L. Morlock, and L. Wissow, 2007, Gifts, bribes and solicitions: Print media and the social construction of informal payments to doctors in Taiwan, Social Science and Medicine,64, 3, 521-530.

[5] DME MAC Jurisdiction C Supplier Manua, Fraud and Abuse. Chapter 14.

[6] Ellis R.P. and M.M. Miller, 2007, Provider payment methods and incentives, Health Systems Policy, Finance, and Organzation, 322-329.

[7] Emons W., 1997, Credence goods and fraudulent experts, Rand Journal of Economics, 1997, 28, 1, 107-119.

[8] Hsiao W.C., P. Braun, D.L. Dunn, E.R. Becker, D. Yntema, D.K. Verrilli, E. Stamenovic, and S.P. Chen, 1992, An overview of the development and refinement of the Resource-Based Relative Value Scale. The foundation for reform of U.S. physician payment, Medical care, NS1-12.

[9] Hensgen F., V. Paris, B. Pierrard, and A. Vergeau, 2000, Charges professionelles des médecins lib'eraux, Centre de recherche d'étude et de documentation en économie de la santé.

[10] Huntb J., 2010, Bribery in health care in Uganda, Journal of Health Economics, 29, 5, 699-707.

[11] Laffont J-J. and D. Martimort, 1997, Collusion Under Asymmetric Information, Econometrica, 65, 4, 875-911.

[12] Ma C.-t.A. and T.G. McGuire, 1997, Optimal health insurance and provider payment, The American Economic Review, 685-704.

[13] Moore, J. and R. Repullo, 1988, Subgame perfect implementation, Econometrica, 56, 5, 1191-1220.

[14] Morris L., 2010, Testimony of: Lewis Morris, Chief Counsel Office of Inspector General, U.S. Department of Health and Human Services.

[15] Office of Inspector General, Carrier Fraud Units, Nov. 1996.

[16] Quesada L., 2003, Modeling collusion as an informed principal problem, Game Theory and Information 0304002, EconWPA.

[17] Sandeep B. and S. Tomas, 1998, Decentialization and collusion, Journal of Economic Theory, 83, 196-232.

[18] Tirole J., 1986, Hierarchies and Bureaucracies, The Journal of Law, Economics, and Organization.

[19] Vaithianathan R., 2003, Supply-side cost sharing when patients and doctors collude, Journal of Health Economics, 22, 5, 763-780.

[20] Van Dulmen A.M., 2002, Different perspectives of doctor and patient in communication, International Congress Series, 1241, 243-248.

A Appendix

A.1 The first-order conditions

The first-order conditions of the first-best social optimum:

$$u_{X_0} = u_{X_1} = u_{X_2} = \lambda_0, \tag{40}$$
$$\alpha_1 v_{q_1}(q_1, m_1) = \lambda_0 C, \tag{41}$$
$$\alpha_2 v_{q_2}(q_2, m_2) = \lambda_0 C, \tag{42}$$
$$\alpha_1 v_{m_1}(q_1, m_1) = \lambda_0 c_1, \tag{43}$$
$$\alpha_2 v_{m_2}(q_2, m_2) = \lambda_0 c_2. \tag{44}$$

The first-order conditions of the second-best weak collusion-proof scheme where the payer places incentives on the patient:

$$X_0: \quad u_{X_0} = \lambda_0, \tag{45}$$
$$X_1: \quad (p_1 + \lambda_1) u_{X_1} = \lambda_0 p_1, \tag{46}$$
$$q_1: \quad \alpha_1 (p_1 + \lambda_1) v_{q_1} = \lambda_0 p_1 C, \tag{47}$$
$$m_1: \quad \alpha_1 (p_1 + \lambda_1) v_{m_1} = \lambda_0 p_1 c_1, \tag{48}$$
$$X_2: \quad p_2 u_{X_2} - \lambda_0 p_2 - \lambda_1 \widetilde{u_{X_2}} = 0, \tag{49}$$
$$q_2: \quad p_2 \alpha_2 v_{q_2} - \lambda_1 \alpha_1 \widetilde{v_{q_2}} - \lambda_3 C = 0, \tag{50}$$
$$m_2: \quad p_2 \alpha_2 v_{m_2} - \lambda_1 \alpha_1 \widetilde{v_{m_2}} - \lambda_3 c_2 = 0, \tag{51}$$
$$R_1: \quad \lambda_2 = \lambda_0 p_1, \tag{52}$$
$$R_2: \quad \lambda_3 = \lambda_0 p_2. \tag{53}$$

The first-order conditions of the second weak collusion-proof mechanism where the payer places

incentives on the physician:

$$X_0: \quad u_{X_0} = \lambda_0, \tag{54}$$

$$X_1: \quad u_{X_1} = \lambda_0, \tag{55}$$

$$q_1: \quad p_1\alpha_1 v_{q_1} = \lambda_1 C, \tag{56}$$

$$m_1: \quad p_1\alpha_1 v_{m_1} = \lambda_1 c_1, \tag{57}$$

$$X_2: \quad u_{X_2} = \lambda_0, \tag{58}$$

$$q_2: \quad p_2\alpha_2 v_{q_2} = \lambda_2 C - \lambda_1 C, \tag{59}$$

$$m_2: \quad p_2\alpha_2 v_{m_2} = \lambda_2 c_2 - \lambda_1 c_1, \tag{60}$$

$$R_1: \quad \lambda_1 = \lambda_0 p_1, \tag{61}$$

$$R_2: \quad \lambda_2 = \lambda_0 p_2 + \lambda_1. \tag{62}$$

A.2 The characterization of the first-best solution

In this section we prove the characterization of the first-best solution in Table 1. From the first-order conditions and the assumption that the marginal cost of effort is perfectly positively correlated with the severity of the disease $c_1/c_2 = \alpha_1/\alpha_2$, we obtain the following two equations:

$$v_{q_1}(q_1, m_1) > v_{q_2}(q_2, m_2), \tag{63}$$

$$v_{m_1}(q_1, m_1) = v_{m_2}(q_2, m_2). \tag{64}$$

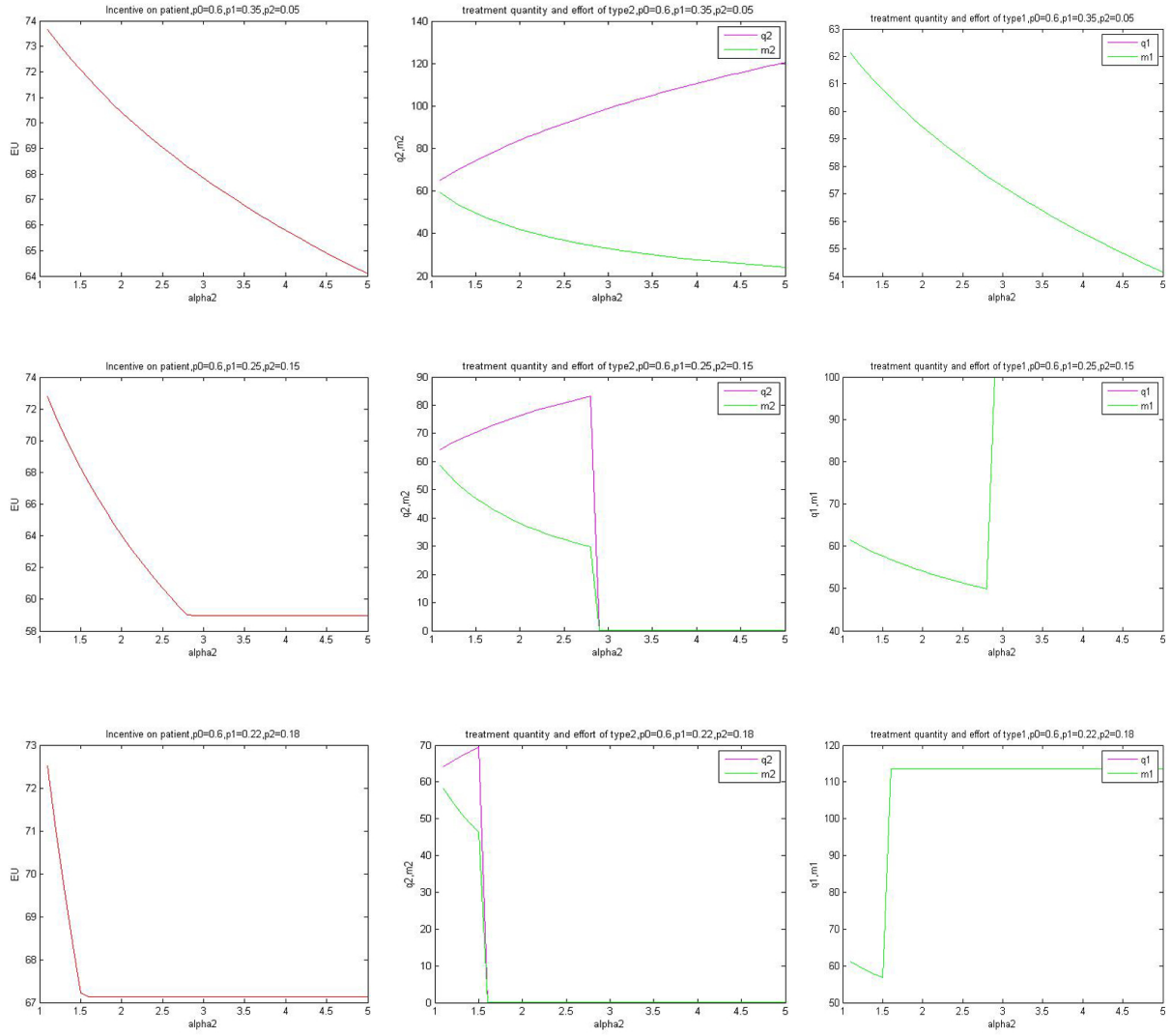
- If $m_1 = m_2$, equation (63) implies that $q_1 < q_2$, if $v_{qm} \neq 0$ it follows that equation (64) is impossible, hence only when $v_{qm} = 0$ we have the solution $m_1 = m_2$ and $q_1 < q_2$.
- If $m_1 < m_2$,
  - when $v_{qm} > 0$, (64) implies that $q_1 < q_2$, then, (63) is checked to be possible;
  - when $v_{qm} = 0$, (64) is impossible;
  - when $v_{qm} < 0$, (64) implies that $q_1 > q_2$, then, (63) is checked to be possible.
- If $m_1 > m_2$,

– when $v_{qm} > 0$, (64) implies that $q_1 > q_2$, then, (63) is checked to be possible;

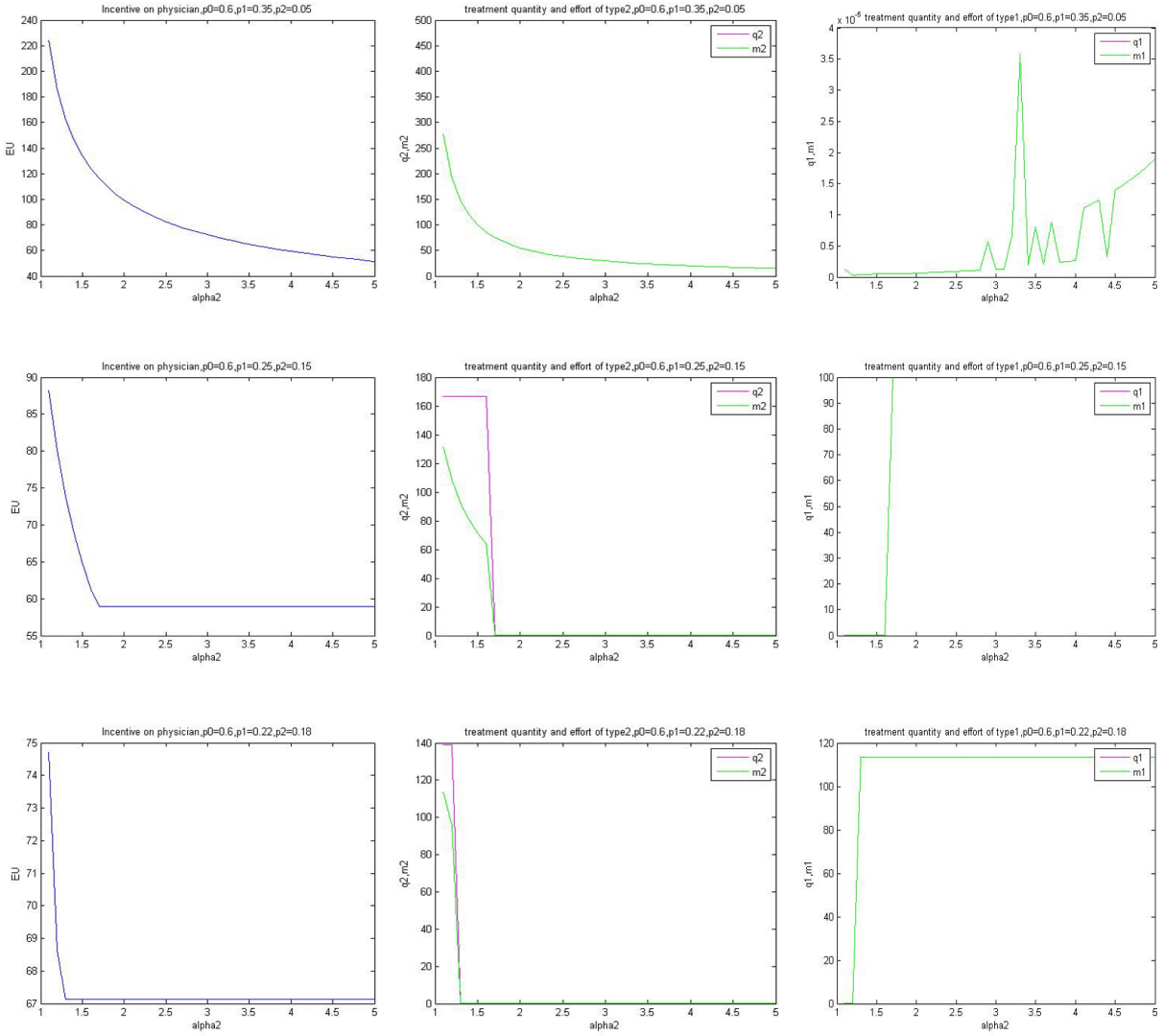
– when $v_{qm} = 0$, (64) is impossible;

– when $v_{qm} < 0$, (64) implies that $q_1 < q_2$, then, (63) is checked to be possible.

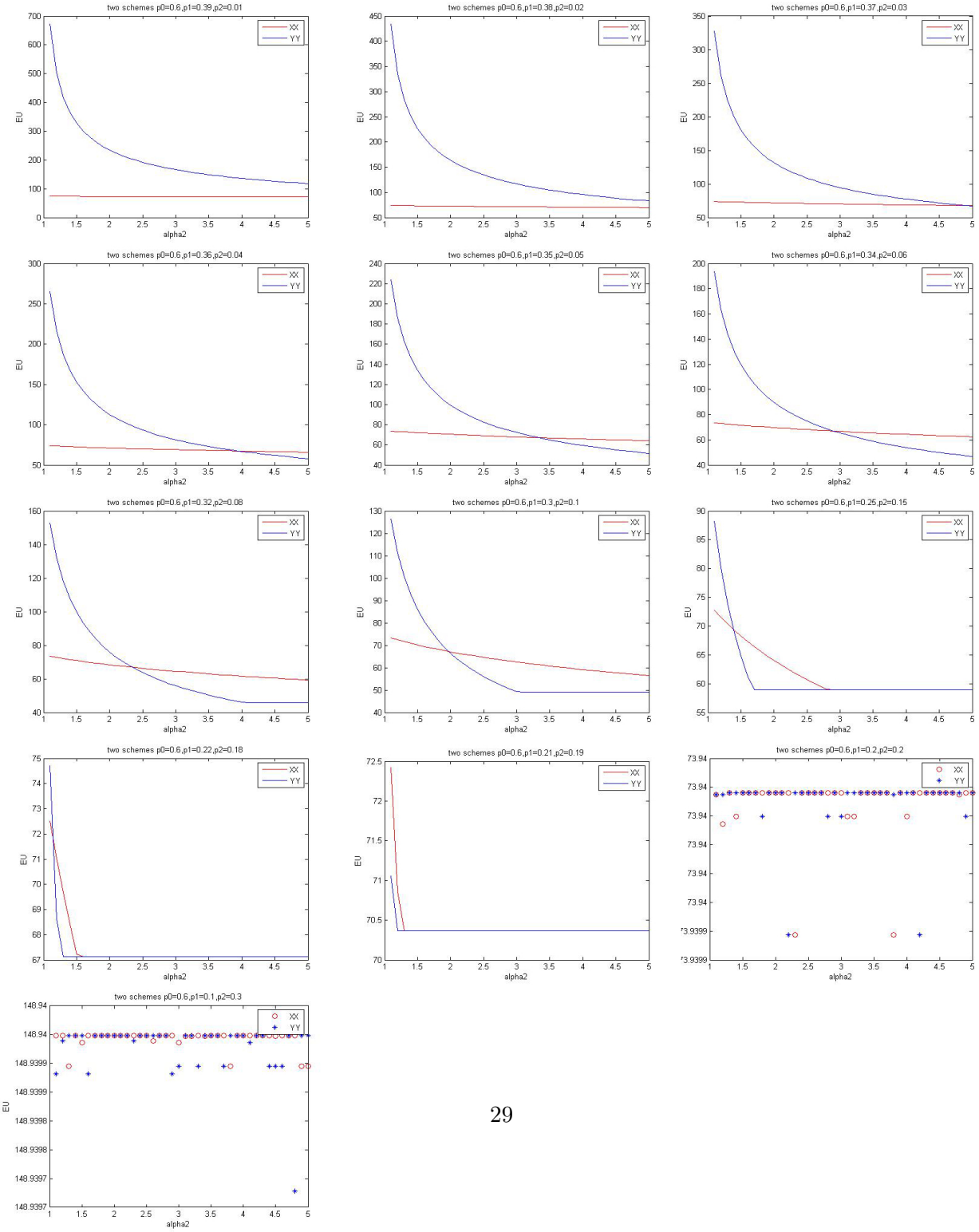
The first scheme: Placing incentives on the patient



The second scheme: placing incentives on the physician



Two schemes


two schemes p0=0.6,p1=0.39,p2=0.01
two schemes p0=0.6,p1=0.38,p2=0.02
two schemes p0=0.6,p1=0.37,p2=0.03
two schemes p0=0.6,p1=0.36,p2=0.04
two schemes p0=0.6,p1=0.35,p2=0.05
two schemes p0=0.6,p1=0.34,p2=0.06
two schemes p0=0.6,p1=0.32,p2=0.08
two schemes p0=0.6,p1=0.3,p2=0.1
two schemes p0=0.6,p1=0.25,p2=0.15
two schemes p0=0.6,p1=0.22,p2=0.18
two schemes p0=0.6,p1=0.21,p2=0.19
two schemes p0=0.6,p1=0.2,p2=0.2
two schemes p0=0.6,p1=0.1,p2=0.3
XX
YY
EU
alpha2