Optimal Compensation Rule: Pay-For-Performance Versus Fee-For-Service*

Yaping Wu†

January 12, 2014

Abstract

This paper examines the optimal non-linear compensation rule of physicians under pay-for-performance, fee-for-service and capitation in the presence of both adverse selection and moral hazard on the supply side. We provide an argument for the criticism on the shortcomings of fee-for-service. More importantly, we also provide a rationale for the continued use of fee-for-service payment even though the serious problems with fee-for-service have been widely acknowledged. We found that when moral hazard is the only problem, fee-for-service can only lead to the substitution of treatment quantity to physician effort, which is inefficient. Consequently, fee-for-service payments should not be used in this case. However, when moral hazard is combined with the adverse selection issue, an efficient screening requires a continued use of fee-for-service for the lower productivity physicians and less pay-for-performance. The design of the use of fee-for-service effectively improves screening.

JEL Code: I18, D82

Keywords: Physician compensation, fee-for-service, pay-for-performance, capitation, multitask

*I am grateful to Helmuth Cremer, Jean-Marie Lozachmeur, Patrick Rey, Sanxi Li, Bruno Jullien, Christian Hellwig, Guido Friebel for their support, advice and comments.

†Toulouse School of Economics (GREMAQ). Email: yapingwu.tse@gmail.com

1 Introduction

The world spends a significant and increasing share of its resources on health care. In 2007, the world spent 9.7% of global income on health care. General government expenditure on health represents 15.4% in total government expenditure, consisting of 59.6% in total expenditure in health care (WHO Statistics 2010). One of the key reasons for the high level of health care spending is the predominance of the fee-for-service payment system, which rewards quantity over quality. Instances of overuse, wasteful use, and misuse of care have been widely documented[1]. The economic literature has identified a number of reasons for the rise in medical cost, among those are moral hazard and adverse selection problems (McGuire 2000). Many health innovators and reformers believe the current fee-for-service system is unsustainable. Instead, they are experimenting with new pay-for-performance models, which assert that buyers should hold providers of health care accountable for both costs and quality of care[2].

However, even though the serious problems with fee-for-service payment have been widely acknowledged, many payment reforms do not change fee-for-service payment in any fundamental way, but merely add new forms of pay-for-performance bonuses or penalties on top of it. For instance, although many medical home payment programs provide a small, flexible, non-visit-based payment to primary care physicians, in most cases, the vase majority of the physicians' revenue continues to come from visit-based fees[3]. A look at "International Profiles of Health Care Systems" documented by the Commonwealth Fund[4] is fairly instructive. They document the health care system in 2009 in Australia, Canada, Denmark, England, France, Germany, Italy, the Netherlands, New Zealand, Norway, Sweden, Switzerland, and the United States. Physicians in all the above countries are paid through a combination of methods but all the systems include a fee-for-service. In France, physicians are self-employed and paid on a fee-for-service basis. In US, physicians are paid through charges or discounted fees paid by most private health plans, capitation rate contracts with some private plans, and fees paid by public programs.

In this paper we explore the issue of "fee-for-service versus pay-for-performance". We provide an argument for the criticism on the shortcomings of fee-for-service. More importantly, we also provide a rationale for the continued use of fee-for-service payment. Using a model inspired by the procurement model due to Laffont and Tirole (1986,1993), we analyze the optimal compensation rule of physicians in the presence of both moral hazard and adverse selection on the supply side.

[1] Center for American progress (2012) and UnitedHealth Center for Health Reform and Modernization (2012)

[2] Physicians may receive bonuses based on performance targets, such as high immunization rates, or low surgical complication rates (Rosenthal et al. 2005)

[3] Center for health care quality and payment reform (2013)

[4] A private foundation that promotes a high performance health care system in the US and other industrialized countries

We consider a nonlinear compensation scheme where the policy instruments include a capitation, a fee-for-service and a pay-for-performance. We incorporate both the characteristics and the actions of the physician. The physician's professional ability positively affects the performance. Moreover he can also exert effort or select treatment quantity to improve the performance. The physician, privately knowing his professional ability, performs a single task with two alterable and substituable dimensions: the treatment quantity and the effort. One dimension, the treatment quantity is contractable, while the other dimension, the effort, is not contractable. The effort can only be indirectly contracted by using the payment based on the outcome, the recovery probability. The current pay-for-performance incentive programs reward physicians upon performance of quality and efficiency measures using the treatment outcome such as the success rate of eye operation, the surgical complication rates or the success rate of preventive intervention. In previous literature on the issue of health care quality, the quality is a black box modeled as a deterministic variable entering into the benefit function or the demand function; in this present paper, we open the black box by allowing the physician to invest in the performance, the recovery probability.

We show that fee-for-service induces substitution of either treatment quantity or number of patients to physician effort. When only moral hazard problem is considered, the optimal compensation policy does not pay for the services. By making the physician residual claimant the payer solves all the dimensions of moral hazard. Payments for the services even make things worse. As the treatment quantity and the effort both contribute to the outcome, if one dimension is paid, while the other is not and the outcome is paid, the physician will naturally tend to select higher level on the dimension which is paid since it also increases the outcome which is also paid. Incentive payment on the contractable dimension will induce substitution of this dimension to the other uncontractable dimension. Consequently, when pure moral hazard is considered, in order not to induce such substitution behavior, the payer should not pay upon any dimensions especially the contractable dimension. As is well known, this is a result similar to the findings in multitask literature (Holmstrom and Milgrom 1991).

However, when both adverse selection and moral hazard are considered, the previous result no longer holds. Surprisingly, the design of the use of fee-for-service effectively improves screening. We show that the highest ability physician obtains a higher reward on the performance and zero fee-for-service, while the lower ability physicians are paid upon the fee-for-service but obtain less rewards on the performance. The reason comes from the fact that the effort is not contractable and can only be indirectly contracted by using the observed outcome. As usual in adverse selection models informational rents of the better types can be mitigated by reducing the performance pay of the less efficient types. In order to give partial incentives of exerting effort, the payer has to reduce the pay-for-performance. However the reduced pay-

for-performance may unwillingly induce too much lower treatment quantity. Hence, to avoid an excessive degradation in quality, it is then desirable to also use fee-for-service payments.

Furthermore, our model provides predictions on the allocations of physicians according to their professional ability. Physicians with higher professional ability are allocated to more serious medical specialties; physicians with lower professional ability are allocated to less serious medical specialties. For instance, a specialist must have high professional ability, and a general practitioner should not treat more serious diseases. As a result, for more serious diseases, we exclude a large range of low ability physicians and retain only the high ability physicians; while for less serious diseases, we do not exclude too much low ability physicians since they can still contribute to the benefit of the patients by dealing with less serious diseases.

This paper is closely related to the physician compensation literature dealing with the quality issue. Ma (1994) studies the quality and cost-reducing-effort tradeoff. The optimal scheme is a pure capitation if dumping is not possible; a piecewise linear scheme is optimal if dumping is not forbidden. Allen and Gertler(1991) assume heterogeneity of patients, and endogenize quantity and quality at the same time. Chalkley and Malcomson (1998 a) accentuate the tradeoff between quality and quantity, and introduce multidimensional quality. They intend to resume the quality issue by making the payment not only depend on the number of patients actually treated but also the number of patients who demand services. They assume that demand responds to some extent to the quality offered by the hospital. The contract is in fact composed of three margins: a lump sum transfer, a fixed price per patient treated and a fixed amount per patient wanting treatment. Chalkley and Malcomson (1998 b) extend their previous paper to analyse contracts when patient demand does not reflect quality.

Moreover, our results reinforce but also stand in contrast to the classic result of Holmstrom and Milgrom (1991), who find that it would be better to pay a fixed wage without any incentive scheme than to base the agent's compensation on the limited dimensions that can be effectively measured. The crucial difference is that HM only consider moral hazard problem. In this paper, we show that although moral hazard calls for no incentive pay on the contractable dimension, adverse selection requires a continued use of this incentive pay.

The paper proceeds as follows. Section 2 sets up the model and derives the first best social optimum. Section 3 studies the pure moral hazard problem and its implementation. Section 4 analyzes the problem with both moral hazard and adverse selection, as well as the contract design. Section 5 provides discussions and robustness check. Section 6 concludes.

2 The Model

2.1 The set-up

Consider a certain restorable disease corresponding to a certain Diagnosis-related group[5]. We study the regulation of physicians in the health care industry who provide services to treat this disease. The identity "physician" can be hospitals or specific health care providers. The analysis focusses on the supply side incentives; we ignore the possible incentives on the demand side and assume homogenous patients on the waiting list.

Let the continuous variable $x \in \mathcal{R}^+$ denote the number of patients selected to be treated by a certain physician. We assume that the population that is affected by the disease is large enough so that there are always patients on the waiting list. Let $q \in \mathcal{R}^+$ denote the treatment quantity per patient selected by a certain physician. This variable may be measured by the number of repeated visits to the physician's office, the quantity of disposable medical appliances, or by the length of hospital stay required by the physician, or the extent of surgical versus drug interventions.

Central to our analysis is the assumption that the provision of treatment to the patients generates recovery only with some probability. The factors that have an impact on the recovery probability are the nature of the disease, the physician's effort, the treatment quantity as well as the physician's professional ability. Let $\alpha \in (0,1)$ denote the nature of the disease, where lower α measures more serious disease for which it is more difficult to recover, and higher α measures less serious disease for which it is more easier to recover. Let $e \in \mathcal{R}^+$ denote the physician's effort per patient. It can be measured, for instance, by the time spent on the patient or a costly mental and manual work. Let $\theta \in \mathcal{R}^+$ denote the professional ability of the physician, being distributed according to the cumulative distribution $F(.)$ with density function $f(.)$ on the interval $[\theta_0, \theta_1]$, with higher θ designating higher professional ability. We denote the recovery probability by $\mathcal{P} = \alpha P(e,q,\theta)$, and assume that

$$
\begin{aligned}
&P_\theta(e,q,\theta) > 0, \quad P_q(e,q,\theta) > 0, \quad P_e(e,q,\theta) > 0 \\
&P_{qq}(e,q,\theta) < 0, \quad P_{ee}(e,q,\theta) < 0 \\
&P_{q\theta}(e,q,\theta) > 0, \quad P_{e\theta}(e,q,\theta) > 0
\end{aligned}
$$

Hence, given the nature of the disease, higher professional ability increases the recovery probability. Larger quantity of treatment or higher level of physician effort increases the recovery probability at a decreasing

[5]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 intent is to identify the services that a physician or a hospital provides.

rate. Higher professional ability also increases both the marginal gain of recovery probability due to treatment quantity and to effort. This is the Spence-Mirrlees condition which simply says that a more efficient type is also more efficient at the margin.

To simplify, we assume that the recovery probability takes the following function[6]:

$$\mathcal{P} = \alpha P(e, q, \theta) = \alpha[1 - exp\big(-\theta f(e, q)\big)]$$

where $f(e, q)$ is the constant elasticity of substitution production function:

$$f(e, q) = (e^{\rho} + q^{\rho})^{\frac{1}{\rho}}$$

with the coefficient of substitution $\rho \in (-\infty, 1]$.

Since x is continuous, according to the law of large numbers, the proportion of patients that successfully recover from the treatment almost surely converges to the recovery probability $\mathcal{P}$.

The monetary cost incurred by the physician depends on the total treatment quantity xq selected by him. We denote the cost function as $C = C(xq)$ with the property that the function C is increasing and convex in its argument. The disutility of effort depends on the per patient level of effort exerted by the physician and the number of patents. Let $\varphi(e)$ denote the disutility of effort per patient and assume that φ is increasing and convex.

We denote the benefit of a patient from recovery by b which is a strictly positive constant, and normalize the benefit in case of non-recovery by zero. The benefit from recovery b can be measured by the reacquisition of potential economic or non-economic losses due to the disease in case of no intervention from the physician[7]. While in case of non-recovery, even with the intervention from the physician, the patient still suffers from these losses and does not reacquire these potential losses. Hence $(b-0)$ measures the loss of benefit that a patient expects to enjoy but unfortunately does not in case of non-recovery.

Therefore, the expected benefit of the patients treated by a type θ physician is

$$x\mathcal{P}b = x(\theta)\alpha P\big(e(\theta), q(\theta), \theta\big)b$$

In the ideal first best, we assume that both the type and the effort are observable. Then, we consider the pure moral hazard where the type is still observable but the effort is not observable. Then, in the

[6]Under general form of probability function, our main results on the first best and second best policy mix remain unaffected. The only thing that is affected is that the direction of the distortion on treatment quantity is ambiguous in the second best.

[7]Economic losses include financial losses such as lost wages (sometimes called lost earning capacity). These losses may be assessed for future losses due to the disease in case of no intervention. Non-economic losses are assessed for the patient itself: physical and psychological harm, such as loss of vision, loss of a limb or organ, the reduced enjoyment of life due to a disability, severe pain and emotional distress in case of no intervention.

second best, we consider the case where neither the type nor the effort is observable. Regardless of which case we consider we assume that the number of patients and treatment quantity can always be observed and verified by the payer. While the disutility of effort experienced by the physician cannot be verified. Moreover, a key assumption to our analysis is that the number of patients that successfully recover is also observable. Thus, because of the law of large numbers, the proportion of patients that successfully recover, the success rate, can be calculated. Therefore the recovery probability can be inferred from the observables. However, the observation of this probability does not allow the payer to perfectly disentangle the type of the physician and his level of effort.

We consider three compensation methods: fee-for-service, capitation and pay-for-performance. Under fee-for-service, the payment is based on the treatment quantity selected by the physician. Under capitation, the payment is based on the number of patients treated by the physician. Under pay-for-performance, the payment is based on the success rate of the physician's intervention. Although the payment scheme cannot be based on the effort which is generally not contractable, the success rate is observable. The payer can take advantage of this additional piece of information by using it as an additional screening variable. To simplify the reading, the θ inside the parenthesis for the allocations will be omitted in the rest of the paper, except when the second best incentive constraints are considered.

We consider a non-linear expected payment scheme $S = S(x, q, \mathcal{P})$. The fee-for-service, the capitation and the pay-for-performance are defined respectively as $S_q(x, q, \mathcal{P})$, $S_x(x, q, \mathcal{P})$, $S_{\mathcal{P}}(x, q, \mathcal{P})$.

The physician is a profit-maximizer who maximizes revenue less the total cost and less the disutility of effort. Hence a type θ physician's net profit under the payer's compensation methods is:

$$\pi(x, q, e; \theta, \alpha) = S\big(x, q, \alpha P(e, q, \theta)\big) - C(xq) - x\varphi(e)$$

The reservation profit for the physician is normalized to zero.

The payer is a private insurance company seeking to attract consumers and maximize its profits in a competitive environment. It only cares about the patients' surplus. Its preference is given by the difference between the patients' expected benefit and the cost of delegating the production of health to the physician. We denote its objective as SW which is thus defined by

$$SW = \int_{\theta_0}^{\theta_1} [x\mathcal{P}b - S\big(x, q, \mathcal{P}\big)]f(\theta)d\theta$$

The timing of the game is as follows. At the first stage, the payer sets the payment policies. Then, the type of the physician realizes. At the second step, the physician selects the number of patients, the treatment quantity as well as the effort level. Finally the payment policies are implemented.

2.2 The first best social optimum

In the first best benchmark, both the type of physician and the effort are observable. The payer chooses the optimal allocations $\{\big(x(\theta), q(\theta), e(\theta), \pi(\theta)\big)_{\theta\in[\theta_0,\theta_1]}\}$ to maximise its objective SW, subject to the physician's participation constraint: $\forall\theta$

$$\pi(\theta) \geq 0$$

Since $S\big(x, q, \alpha P(e, q, \theta)\big) = \pi + C(xq) + x\varphi(e)$, replacing it in the SW function, the objective function of the payer becomes

$$SW = \int_{\theta_0}^{\theta_1} [x\alpha P(e, q, \theta)b - \pi - C(xq) - x\varphi(e)]f(\theta)d\theta$$

Since all information is publicly observable, the payer will leave no information rent to the physician by choosing, $\forall\theta$, $\pi(\theta) = 0$. Assume an interior solution, the maximization yields the following first order conditions: $\forall\theta$

$$x: \quad \alpha P(e, q, \theta)b = qC'(xq) + \varphi(e) \tag{1}$$

$$q: \quad \alpha P_q(e, q, \theta)b = C'(xq) \tag{2}$$

$$e: \quad \alpha P_e(e, q, \theta)b = \varphi'(e) \tag{3}$$

At the first best social optimum, the marginal cost of patient is equal to the marginal benefit. The marginal cost of treatment is equal to the marginal benefit of treatment, and the marginal cost of effort is equal to the marginal benefit of effort. If second order cross derivative of recovery probability with respect to e and q are sufficiently small in absolute value, equation (3) implies that under the Spence-Mirrlees condition $P_{e\theta}(e, q, \theta) > 0$, a higher ability physician exerts a higher level of effort. From equation (2), under the Spence-Mirrlees condition $P_{q\theta}(e, q, \theta) > 0$, a higher ability physician is allocated with a higher level of total treatment quantity, because he/she is more efficient at the margin. But whether he/she is allocated with higher level of treatment or more patients is ambiguous.

3 The pure moral hazard

In this section, we study the pure moral hazard where we assume that the effort is not observable but the type is observable. Since the link between effort, types and the recovery probability is completely deterministic, it entails no randomness at all. Given a target value of the recovery probability $\mathcal{P}$, which is

a contractual variable available to the payer, given the physician's type and the nature of the disease, effort is completely determined by the condition $e = e(\mathcal{P}, q, \theta)$, where $e(.)$ is implicitly defined by the identity $\mathcal{P} = \alpha P(e(\mathcal{P}, q, \theta), q, \theta)$ for all θ in Θ and all $\mathcal{P}$. In fact, the physician has no freedom in choosing his effort level when he takes his decision. As a result, the first best allocations can be achieved. Before moving to the contract design in the pure moral hazard setting, let us first analyze the physician's behavior.

3.1 The physician's behavior

In this section, we study the decision choice of the physician under a given contract. Given the nature of the disease, the physician's ability and the treatment quantity, the effort and the recovery probability is deterministically linked by the probability function. As a result, after having chosen the treatment quantity, choosing the effort is equivalent to choosing a recovery probability. We apply the change of variable: $e = e(\mathcal{P}, q, \theta)$, where $e(.)$ is implicitly defined by the identity $\mathcal{P} = \alpha P(e(\mathcal{P}, q, \theta), q, \theta)$ for all θ in Θ and all $\mathcal{P}$. The physician, privately knowing his type θ, selects the number of patients, the treatment quantity per patient and the recovery probability. Assuming an interior solution and maximizing the expected profit of a physician of type θ:

$$\pi = S(x, q, \mathcal{P}) - C(xq) - x\varphi(e(\mathcal{P}, q, \theta))$$

We obtain the following first order conditions

$$S_x(x, q, \mathcal{P}) = qC'(xq) + \varphi(e) \tag{4}$$

$$S_q(x, q, \mathcal{P}) = xC'(xq) + x\varphi'(e(\mathcal{P}, q, \theta))e_q(\mathcal{P}, q, \theta) \tag{5}$$

$$S_{\mathcal{P}}(x, q, \mathcal{P}) = x\varphi'(e(\mathcal{P}, q, \theta))e_{\mathcal{P}}(\mathcal{P}, q, \theta) = x\frac{\varphi'(e(\mathcal{P}, q, \theta))}{\alpha P_e(e, q, \theta)} \tag{6}$$

In order to show the incentives of the physician, we rearrange the above equations. Equation (3) implies that

$$\alpha P_e(e, q, \theta)S_{\mathcal{P}}(x, q, \mathcal{P}) = x\varphi'(e) \tag{7}$$

Using implicit function theorem, equation (2) implies

$$S_q(x, q, \mathcal{P}) = xC'(xq) - x\varphi'(e)\frac{P_q(e, q, \theta)}{P_e(e, q, \theta)} \tag{8}$$

From equation (4)(5), it follows that

$$S_q(x, q, \mathcal{P}) + \alpha S_{\mathcal{P}}(x, q, \mathcal{P})P_q(e, q, \theta) = xC'(xq) \tag{9}$$

Therefore, from equation (4)(7)(9), we observe that fee-for-service induces substitution of either treatment quantity or patients to physician effort. Keeping the capitation and pay-for-performance constant, if we increase the fee-for-service, equation (9) implies that the physician will choose either higher treatment or more patients or both. At least one allocation, q or x, will be higher. If the physician chooses both higher treatment quantity and more patients, equation (4) implies that effort will be lower. If the physician chooses only higher treatment quantity, then following equation (4) effort will be lower too. If the physician chooses only more patients, then equation (7) implies that effort will be lower too. However, the physician may choose a much larger amount of treatment with fewer patients or a higher number of patients with less treatment. In these two cases, which allocations are induced by the fee-for-service is ambiguous. But according to equation (9) the marginal cost of treatment is higher, at least one allocation, x or q, is selected to be higher. Hence, the fee-for-service induces substitution of either treatment quantity or patients to physician effort.

Since the payment method cannot be based on the effort, the only method that can induce a profit-maximizer physician's effort is pay-for-performance. If the percentage of recovered patients is too costly to be observed or verified so that the payer cannot use the pay-for-performance, equation (7) implies that $e = 0$: there is no incentive to exert any effort[8].

3.2 Implementation in a pure moral hazard setting

In this section, we turn to the implementation of the first best allocations in a pure moral hazard setting.

Lemma 1 *When there is no adverse selection problem, by making the physician residual claimant:* $S^*(x, \mathcal{P}) = x\mathcal{P}b - M(\theta)$ *where* $M(\theta)$ *is a constant depending on the type, the payer solves all the dimensions of moral hazard problem.*

No matter how many dimensions the effort e has[9], the treatment q is contractable while e is not. By making the physician residual claimant, the physician's profit is perfectly in line with the payer's objective. Consequently, there is no need to contract on any contractable dimension q. We thus derive the optimal

[8] Equation (4)(9) implies that the number of patients will not be zero because the fee-for-service and the capitation is still positive.

[9] e can be generalized to a N-dimensional vector $\{e_1, e_2, ..., e_N\}$ and q can be generalized to a N-dimensional vector $\{q_1, q_2, ..., q_N\}$

policy with pure moral hazard:

$$S^*_{\mathcal{P}}(x, q, \mathcal{P}) = xb \tag{10}$$

$$S^*_q(x, q, \mathcal{P}) = 0 \tag{11}$$

$$S^*_x(x, q, \mathcal{P}) = \mathcal{P}b \tag{12}$$

This gives rise to the following proposition:

Proposition 1 *When there is no adverse selection problem,*
i) The optimal compensation policy does not pay for services;
ii) The pay-for-performance incentivizes the physician to fully internalize the patients' benefit.

There is no need to reward services. Rewards on service only induce substitution behavior which are inefficient and inflate costs. The physician performs a single task: the recovery probability, which has two alterable dimensions: the treatment quantity and the effort. Both dimensions contribute to the recovery probability and they are substituable to some degree measured by ρ. One dimension, the treatment quantity is contractable, while the other dimension, the effort, is not contractable. The effort can and can only be indirectly contracted by using the payment based on the outcome, the recovery probability. As two dimensions equally contribute to the outcome, if one dimension is paid, while the other is not and the outcome is paid, the physician will naturally tend to select higher level on the dimension which is paid since it also increases the outcome which is also paid. We thus get a result which is similar to the finding of Holmstrom and Milgrom (1991). Incentive payment on the contractable dimension will induce substitution of this dimension to the other uncontractable dimension. In order not to induce such substitution behavior, the payer should not pay upon any dimensions especially the contractable dimension in the pure moral hazard case. If only the outcome is paid, the physician will select the two dimensions by comparing their contributions to the benefit and their cost just as what the payer does. Hence in the presence of pure moral hazard, the optimal policy mix includes no fee-for-service at all. The reason why the capitation based on the contractable number of patients is still used is that the number of patients does not affect the patients' recovery probability.

Since in any case the payment scheme cannot be based on the effort, if the treatment outcome is not observable, or even if it is observable, it is too costly to make the payments based on the treatment outcome, then the payer can only use the number of patients and the treatment quantity as contractable variables. It follows that a profit-maximizer physician will have no incentive to exert any effort. Thus, in order to implement the first best allocations, either, the payer is fully dictatorial on all the allocations

and leave no freedom to the physician to choose any allocations, or, the payer has to be dictatorial on the effort level and decentralize the treatment quantity and the number of patients by the following scheme:

$$\begin{cases} e = e^* \quad \text{dictatorship} \\ S(x, q \mid e = e^*) = \alpha P(e^*, q, \theta) bx - M \end{cases}$$

Otherwise, other measures must be used to induce quality-improvement effort (See Chalkley and Malcomson 1998 a and b).

4 Moral hazard and adverse selection

4.1 The characterization

The first best and pure moral hazard solution and their decentralization have been derived under the assumption that the payer observes the type of the physician. When there is asymmetric information on the type, the optimal scheme with pure moral hazard is generally not feasible. In this section, we adopt an information structure that is inspired by the procurement literature due to Laffont and Tirole (1986, 1993), and characterize the second best social optimum when neither the type nor the effort is observable.

By the change of variables, with observables being x, q and $\mathcal{P}$, we consider the direct revelation mechanisms $\{S(\hat{\theta}), x(\hat{\theta}), q(\hat{\theta}), \mathcal{P}(\hat{\theta})\}_{\hat{\theta}\in[\theta_0,\theta_1]}$ which are truth telling. The payer's problem is written as follows:

$$\max_{x(\theta),q(\theta),\mathcal{P}(\theta),S(\theta)} \int_{\theta_0}^{\theta_1} [x(\theta)\mathcal{P}(\theta)b - S(\theta)]f(\theta)d\theta$$

$$s.t \begin{cases} (IC_\theta): & S(\theta) - C\big(x(\theta)q(\theta)\big) - x(\theta)\varphi\big(e(\mathcal{P}(\theta), q(\theta), \theta)\big) \\ & = \max_{\hat{\theta}\in[\theta_0,\theta_1]} \big\{S(\hat{\theta}) - C\big(x(\hat{\theta})q(\hat{\theta})\big) - x(\hat{\theta})\varphi\big(e(\mathcal{P}(\hat{\theta}), q(\hat{\theta}), \theta)\big)\big\} \\ (PC_\theta): & S(\theta) - C\big(x(\theta)q(\theta)\big) - x(\theta)\varphi\big(e(\mathcal{P}(\theta), q(\theta), \theta)\big) \geq 0 \\ & x(\theta) \geq 0, \quad q(\theta) \geq 0, \quad e(\theta) \geq 0 \quad 0 \leq \mathcal{P}(\theta) \leq 1 \quad \text{for all} \quad \theta \in [\theta_0, \theta_1] \end{cases}$$

Applying Envelop Theorem to the incentive contraints implies

$$\dot{\pi} = -x\varphi'\big(e(\mathcal{P}, q, \theta)\big)\frac{\partial e(\mathcal{P}, q, \theta)}{\partial \theta}$$

It follows that $\dot{\pi} \geq 0$[10]. (PC_θ) reduce to $\pi(\theta_0) \geq 0$. Hence at optimum, there is no information rent for the lowest type physician.

[10]This is because $\varphi_e > 0$, then $\frac{\partial e(\mathcal{P},q,\theta)}{\partial \theta}$ is the partial derivative of effort with respect to the ability given the recovery probability and the treatment quantity. By implicit function theorem, $\frac{\partial e(\mathcal{P},q,\theta)}{\partial \theta} = -\frac{P_\theta(e,q,\theta)}{P_e(e,q,\theta)}$ which is negative.

The Hamiltonian function writes as:

$$\mathcal{H}(\theta,\pi,x,q,\mathcal{P},\lambda)=\Big[x\mathcal{P}b-\pi-C(xq)-x\varphi(e(\mathcal{P},q,\theta))\Big]f(\theta)+\lambda(\theta)\Big[-x\varphi'(e(\mathcal{P},q,\theta))\frac{\partial e(\mathcal{P},q,\theta)}{\partial\theta}\Big]$$

where the state variable is π, the control variables are $x,q,\mathcal{P}$, the co-state variable is λ. The Hamilton-Jacobi system gives that

$$\begin{cases}\dot{\lambda}^{SB}(\theta)=-\frac{\partial\mathcal{H}}{\partial\pi}=f(\theta)\\ \lambda^{SB}(\theta_1)=0\end{cases}$$

which yields

$$\lambda^{SB}(\theta)=F(\theta)-1$$

For each type of physician, the effort, type, treatment quantity and the recovery probability is deterministically linked by the relationship $\mathcal{P}=\alpha P(e,q,\theta)$. Optimizing with respect to the $\mathcal{P},x,q$ amounts to optimizing with respect to e,x,q. Expressing the payer's objective function in terms of efforts instead of the recovery probability and inserting the optimal co-state variable the Hamiltonian becomes:

$$\begin{aligned}\mathcal{H}(\theta,\alpha,\pi,x,q,e,\lambda^{SB})=&\Big[x\alpha P(e,q,\theta)b-\pi-C(xq)-x\varphi(e)\Big]f(\theta)\\&-\big(1-F(\theta)\big)\Big[-x\varphi'\big(e(\alpha P(e,q,\theta),q,\theta)\big)\frac{\partial e(\alpha P(e,q,\theta),q,\theta)}{\partial\theta}\Big]\end{aligned}$$

The first order conditions are as follows:

$$x:\begin{cases}\alpha P(e,q,\theta)b\leq qC'(xq)+\varphi(e)-\frac{1-F(\theta)}{f(\theta)}\Delta_x & \text{if } x=0\\ \text{with equality} & \text{if } x>0\end{cases}\tag{13}$$

$$q:\begin{cases}x\alpha P_q(e,q,\theta)b\leq xC'(xq)-\frac{1-F(\theta)}{f(\theta)}\Delta_q & \text{if } q=0\\ \text{with equality} & \text{if } q>0\end{cases}\tag{14}$$

$$e:\begin{cases}x\alpha P_e(e,q,\theta)b\leq x\varphi'(e)-\frac{1-F(\theta)}{f(\theta)}\Delta_e & \text{if } e=0\\ \text{with equality} & \text{if } e>0\end{cases}\tag{15}$$

where

$$\Delta_x=\varphi'(e)\frac{\partial e(\alpha P(e,q,\theta),q,\theta)}{\partial\theta}$$

$$\Delta_q=x\varphi'(e)\frac{\partial^2 e(\alpha P(e,q,\theta),q,\theta)}{\partial\theta\partial q}$$

$$\Delta_e=\frac{\partial\Big[x\varphi'(e)\frac{\partial e(\alpha P(e,q,\theta),q,\theta)}{\partial\theta}\Big]}{\partial e}$$

As usual in adverse selection models, for $\theta = \theta_1$, $F(\theta_1) = 1$, there is no distortion for the highest type. It is proved in the appendix that $\Delta_x < 0$, $\Delta_q < 0$, $\Delta_e < 0$. Hence, at the second best optimum, for all types such that $\theta < \theta_1$, the number of patients are downward distorted. The marginal benefit of effort is greater than the marginal cost. These types of physician under-provide effort. The treatment quantity is also downward distorted for these types[11].

4.2 Exclusion

In this section, we study respectively the conditions under which no type is excluded from the optimal contract and the conditions under which a subset of types are excluded.

Lemma 2 *Given the nature of the disease, assume that the hazard rate* $\frac{1-F(\theta)}{f(\theta)}$ *is strictly decreasing in* θ.
i) If $\rho \to 0$, *and*

$$C'(0) = 0$$
$$\varphi'(0) = 0$$

exclusion is never desirable. Otherwise,

ii) If there exists $\tilde{\theta}$ *such that,*

$$C'(0) \geq \alpha P_q(0, 0, \tilde{\theta})b$$
$$\varphi'(0) \geq \alpha P_e(0, 0, \tilde{\theta})b$$

then all types $\theta_0 \leq \theta \leq \tilde{\theta}$ *are excluded.*

iii) If $\forall\theta$,

$$C'(0) < \alpha P_q(0, 0, \theta)b$$
$$\varphi'(0) < \alpha P_e(0, 0, \theta)b$$

then, if these exists $\tilde{\theta}$ *such that,*

$$\frac{f(\tilde{\theta})}{1 - F(\tilde{\theta})} < Min\left\{ \frac{\varphi'(0)}{\tilde{\theta}(2^{\frac{1-\rho}{\rho}}\alpha\tilde{\theta}b - C'(0))}, \frac{\varphi'(0)}{\tilde{\theta}(2^{\frac{1-\rho}{\rho}}\alpha\tilde{\theta}b - \varphi'(0))} \right\}$$

then all types $\theta_0 \leq \theta \leq \tilde{\theta}$ *are excluded.*

[11] Although the downward distortion of treatment is derived under CES function, whether the treatment is downward or upward distorted does not affect our main conclusion in the second best policy mix.

Proof. Appendix.

In result i), if $\rho \to 0$, the Inada conditions are satisfied: $\forall\theta$

$$P_q(0,0,\theta) = +\infty$$
$$P_e(0,0,\theta) = +\infty$$

Hence i) simply says that if, whatever the type, the marginal benefits are infinity and the marginal costs are zero at zero allocations, no type is excluded, even the least efficient type. If these two conditions are not satisfied, result ii) says that, if there exists a type such that the marginal costs at zero are higher than the marginal benefits at zero, then the Spence-Mirrlees condition guarantees that all types that are lower than this type are excluded from the optimal contract. If for all types, the marginal benefits at zero are higher than the marginal costs at zero, then result iii) says that types whose ability is close to zero and whose presence among the physicians are insignificant are excluded from the optimal contract. Note that the conditions depend only on the primitives.

Furthermore, result iii) predicts the allocation of physicians according to their professional ability. A higher α implies a lower critical value of exclusion $\tilde{\theta}$, while a lower α implies a higher critical value of exclusion. Therefore, we obtain the following proposition:

Proposition 2 *Physicians with higher professional ability are allocated to more serious medical specialties; physicians with lower professional ability are allocated to less serious medical specialties.*

For more serious diseases, the patients have less chance to recover, thus the criteria of exclusion is more stringent. We need much higher ability physicians and we exclude a large range of low types. For example, specialists must have high professional ability and general practitioners (GPs) should not treat serious diseases. For less serious diseases, the patients have more chance to recover, the criteria of exclusion is less stringent. We retain some low ability physicians within a certain range. Examples are the general practitioners. For the GPs, we do not exclude too much low types, because they can still contribute to the benefit of the patients by dealing with less serious diseases.

4.3 The second best compensation policy

In this section we derive the second best compensation scheme for the types who are not excluded from the contract. By comparing equations (1)(4)(6) with equations (13)(14)(15), we obtain the following prices:

$$S_x(x,q,\mathcal{P}) = \mathcal{P}b - \frac{1-F(\theta)}{f(\theta)}|\Delta_x| \tag{16}$$

$$S_q(x,q,\mathcal{P}) = \frac{1-F(\theta)}{f(\theta)}(-|\Delta_q| + |\Delta_e|\frac{P_q}{P_e}) \tag{17}$$

$$S_{\mathcal{P}}(x,q,\mathcal{P}) = xb - \frac{1-F(\theta)}{f(\theta)}\frac{|\Delta_e|}{\alpha P_e} \tag{18}$$

From equation (16), it follows that,

$$S_x^{SB}(\theta_1) = S_x^*(\theta_1)$$

$$S_x^{SB}(\theta) < S_x^*(\theta), \quad \forall \tilde{\theta} < \theta < \theta_1$$

The highest type physician obtains the first best capitation payment. For the other types, the capitation is less than the first best level.

From equation (18) it follows that,

$$S_{\mathcal{P}}^{SB}(\theta_1) = x^*(\theta_1)b = S_{\mathcal{P}}^*(\theta_1)$$

$$S_{\mathcal{P}}^{SB}(\theta) < x^{SB}b < x^*b = S_{\mathcal{P}}^*(\theta), \quad \forall \tilde{\theta} < \theta < \theta_1$$

The highest type physician obtains the first best pay-for-performance payments thus is given full incentive to exert effort in improving the recovery probability. He is residual claimant for his effort. For all types that are lower than the highest type $\theta < \theta_1$, they are rewarded less than the first best pay-for-performance, thus are just given partial incentive to exert effort. They get only a fraction of his marginal effort in improving recovery probability. They are partially residual claimant for their effort and thus under-provide effort.

Surprisingly, equation (17) implies that all types lower than the highest type get a positive fee-for-service, while the highest type still get zero fee-for-service as in the first best:

$$S_q^{SB}(\theta) > 0, \quad \forall \tilde{\theta} < \theta < \theta_1$$

$$S_q^{SB}(\theta_1) = 0$$

Proof. Appendix.

We obtain thus the following proposition:

Proposition 3 *In the presence of both adverse selection and moral hazard,*
i) We want to contract on all contractable q;
ii) The highest ability physician obtains a higher reward on the performance and zero fee-for-service, while the lower ability physicians are paid upon the fee-for-service but obtain less rewards on the performance.

Proposition 3 states that in the second best, the contractable dimension of the task, the treatment quantity, is paid for the lower ability types. This is the main, and maybe surprising difference with the pure moral hazard policy similar to the finding of Holmstrom and Milgrom (1991) when adverse selection is not considered. The reason is that the effort is not contractable and can only be indirectly contracted by using the observed outcome, the success rate. In the presence of adverse selection, the informational rents of the better types can be mitigated by reducing the performance pay of the less efficient types. In order to give partial incentives of exerting effort, the payer has to reduce the pay-for-performance. However the reduced pay-for-performance may unwillingly induce too much lower treatment quantity. Hence, to avoid an excessive degradation in quality, it is then desirable to also use fee-for-service payments[12].

Moreover, we observe that the larger the distortion on treatment, the lower the fee-for-service; the smaller the distortion on treatment, the higher the fee-for-service. It follows directly from equation (17) which states that the level of fee-for-service depends exactly on the magnitude of the downward distortion on treatment quantity. If the distortion is larger in the second best, it is better to give a lower fee-for-service; if the distortion is small, a higher fee-for-service is needed.

To sum up, although moral hazard calls for no incentive pay on the contractable dimension, adverse selection requires a continued use of this incentive pay. Due to the fact that effort can only be indirectly contracted by the pay-for-performance, the incentive payment on the contractable treatment quantity is used as an instrument to correct unwilling distortions.

5 Discussion and robustness check

5.1 Chronic disease

The above analysis on optimal compensation policy assumes that the patients recover with a certain probability. In the case of chronic disease, the patient may never be able to recover. In this case, another performance measure is needed, and our $\mathcal{P}$ function in the model can still be used for the analysis just with another interpretation. For example, a possible performance measure is how long in average the patient's

[12]Indeed, if under other forms of probability function the treatment quantities for the lower types are upward distorted in the second best, it even reinforces our result that the fee-for-service is positive for these types in order to encourage treatment quantity. Proof. Appendix

life has been extended under the physician's intervention[13]. If in the end, a performance measure is not available, other measures must be used to induce quality-improvement effort, for instance, payment based on the number of patients who demand services (See Chalkley and Malcomson 1998 a and b).

5.2 Time lag and noisy observation of performance

Our work can also be extended to the cases where the performance is a noisy observation. This could happen when heterogeneity of patients within one certain DRG is considered. Patients may differ in their severities even within one DRG, consequently they may recover at different moments. At the moment of survey, a patient may recover after the survey, therefore this data is missing. Or, it seems that he/she has recovered but actually he/she has not. Hence, the noisy observation of $\mathcal{P}$ depends on the patients' severities.

Within our framework, we propose a variation of our model. let γ denote the severity within one DRG, being distributed according to a cumulative function $G(.)$. We only observe $\mathcal{P}'$: $\mathcal{P}' = \mathcal{P} + \varepsilon(\gamma)$ where $\varepsilon(.)$ measures the noise from the heterogeneity of severities. Then, when only moral hazard is considered, the payer can offer a payment scheme such that $S'(x, \mathcal{P}') = x\mathcal{P}'b - x\mathbb{E}(\varepsilon(.))b - M$, where $\mathbb{E}(\varepsilon(.))$ is the expectation of the noise. Hence, noisy observation is not a problem for the pure moral hazard if $\mathbb{E}(\varepsilon(.))$ is known. There is still no need for fee-for-service. When both moral hazard and adverse selection are considered, noisy observation would not be a problem if other higher moments of $\varepsilon(.)$ can be estimated from medical statistics. The second best non-linear compensation scheme can be implemented through a menu of quadratic schemes which are corrected by the moments of the degree of the quadratic schemes. It would be an application of the polynomial approximation method proposed by Caillaud, Guesnerie and Rey (1992) to physician compensation. In practice, if we can estimate what is the distribution of patients' severities in the population within one certain DRG and try to estimate how the noise of performance observation depends on the patients' severities (linearly, polynomially, exponentially...) from medical statistics, then we can first find the menu of quadratic schemes to approximate our optimal non-linear scheme, then correct it by its moments. And finally, the pay-for-performance based on the noisy observation would implement our second best optimal allocations.

[13]For example, The disability-adjusted life year (DALY) is a measure of overall disease burden, expressed as the number of years lost due to ill-health, disability or early death.

6 Conclusion

This paper examines the optimal compensation rule of physicians under three payment methods: pay-for-performance, fee-for-service and capitation in the presence of both adverse selection and moral hazard on the supply side. We explore the issue of "fee-for-service versus pay-for-performance". We provide an argument for the criticism on the shortcomings of fee-for-service. More importantly, we also provide a rationale for the continued use of fee-for-service payment even though the serious problems with fee-for-service have been widely acknowledged. Using a model inspired by the procurement model due to Laffont and Tirole (1986,1993), we show that fee-for-service induces substitution of either treatment quantity or number of patients to physician effort. When moral hazard is the only problem, the optimal compensation policy includes a capitation and a pay-for-performance without fee-for-service. The pay-for-performance incentivizes the physician to fully internalize the patients' benefit. When moral hazard is combined with the adverse selection issue, in order to avoid an excessive degradation in quality, an efficient screening requires a continued use of fee-for-service for the lower productivity physicians and less pay-for-performance.

Furthermore, our model provides predictions on the allocations of physicians according to their professional ability. Physicians with higher professional ability are allocated to more serious medical specialties; physicians with lower professional ability are allocated to less serious medical specialties. For instance, a specialist must have high professional ability, and a general practitioner should not treat more serious diseases. As a result, for more serious diseases, we exclude a large range of low ability physicians and retain only the high ability physicians; while for less serious diseases, we do not exclude too much low ability physicians since they can still contribute to the benefit of the patients by dealing with less serious diseases.

Our results also contribute to the multitask principal-agent literature. We show that although moral hazard calls for no incentive pay on the contractable dimension due to the problem of substitution behavior, adverse selection requires a continued use of this incentive pay. The incentive pay on the contractable dimension is used as an instrument to correct unwilling distortions in the second best world.

All this being said, the present paper has proposed a framework to study the pay-for-performance incentive programs. A possible extension would be to study an alternative pay-for-performance system also used in practice which rewards physicians according to how well they perform relative to their peers on various quality or cost measures. To put it in a nutshell, the pay-for-performance remains an area promising possible dramatic advances and worthy of significant new research.

Reference

Allen and Gertler (1987) Regulation and the provision of quality to heterogenous consumers: the case of prospective pricing of medical services. National Bureau of Economic Research. Working Paper No.2269

Bardey David, Canta Chiara and Lozachmeur Jean-Marie, (2012), The regulation of health care providers payments when horizontal and vertical differentiation matter, Journal of Health Economics, Elsevier, vol. 31(5), pages 691-704.

Caillaud, Guesnerie and Rey (1992), Noisy observation in adverse selection models, The Review of Economic Studies, Vol. 59, No, 3, pp. 595-615

David Bardey, (2004). A paradoxical risk aversion effect on the consumers' demand for quality. Localisation et tarification, Recherches economiques de Louvain, De Boeck Universite, pages 109-115

Dranove, D., and M. Satterthwaite (1992), Monopolistic competition when price and quality are imperfectly observable, RAND Journal of Economics 23(4):518-534.

Edward P. Lazear (2000), Performance Pay and Productivity, The American Economic Review, Vol. 90, No. 5 (Dec., 2000), pp. 1346-1361

Ellis, R.P., and McGuire, T.G. (1986). Provider behavior under prospective reimbursement: cost sharing and supply. Journal of Health Economics 5, 129-152.

Farewell to fee-for-service? A real world strategy for health care payment reform, Working Paper 8 (December 2012), UnitedHealth Center for Health Reform and Modernization

Gal-Or, E. (1996), Optimal reimbursement rules and malpractice reform, mimeo (University of Pittsburgh).

Grilo I. and X.Wauthy (2000), Price Competition when Product Quality isUncertain, Recherches Economiques de Louvain, 66(4), pp.415-437

Helmuth Cremer, Jean-Marie Lozachmeur and Pierre Pestieau (2011), The Design of Long Term Care Insurance Contracts, 8 - 10 April 2011 CESifo Conference Centre, Munich

Hector Chade and Edward Schlee (2012), Optimal insurance with adverse selection, Theoretical Economics 7 (2012), 571C607

Holmstrom, Bengt and Milgrom, Paul, (1991), Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design, Journal of Law, Economics and Organization, Oxford University Press, vol. 7(0), pages 24-52, Special I

International Profiles of Health Care Systems, The Commonwealth Fund, June 2010

Jacob, McGuire and Newhouse (2007) Using performance measures to motivate 'report-averse' and 'report-loving' agents. Journal of Health Economics 26 (2007) 1170-1189

Krishna K. and T. Winston (2000), If at first you dont succed... Profits,Prices and Market Structure in a Model of Quality with Unknowable Consumer Heterogeneity, NBER Working Paper Serie

Jean-Jacques Laffont Jean Tirole, (1986), Using Cost Observation to Regulate Firms, Journal of Political Economy, Vol. 94, No. 3, Part 1 (Jun., 1986), pp. 614-641

Jean-Jacques Laffont and Jean Tirole (1993), A theory of incentives in procurement and regulation

Jean-Jacques Laffont and Martimort (2002)The theory of incentives: the principle-agent model.

Martin Chalkley, James M. Malcomson(1998 b),Contracting for health services when patient demand does not reflect quality. Journal of Health Economics 17 (1998) 1-19

Martin Chalkley, James M. Malcomson(1998 a), Contracting for health services with unmonitored quality. The Economic Journal, 108 (July),1093-1110

M., Chalkley and J. Malcomson (2000) Government Purchasing of Health Services. Handbook of Health Economics, Chapiter 15, pp. 847-890

Ma, C.-t.A.(1994) Health care payment systems: Cost and quality incentives, J. Econ. Manage. Strategy 3 (1), 93-112.

McGuire (2000), Physician agency, Handbook of Health Economics

Maura Calsyn and Emily Oshima Lee (September 2012), Alternatives to Fee-for-Service Payments in Health Care Moving from Volume to Value, Center for American progress.

Newhouse, J.P.(1970) "Towards a theory of nonprofit institutions: An economic model of a hospital." Am. Econ. Rev. 60 (1), 64-74.

Pauly, M. (1980), Doctors and their Workshops (University of Chicago Press, Chicago).

Paul V. Dutton, Health care in France and the United States: learning from each other, The Brookings Institution

Rosenthal MB, Frank RG, Li Z, and Epstein AM (2005) Early evidence aith pay-for-performance: from concept to practice. Journal of American Medical Association 294: 1788-1793.

Rosenthal MB,Landon BE, Normand S(L T, Frank RG, and Epstein AM (2006) Pay-for-performance in commercial HMOs. New England Journal of Medicine 355: 1895-1902.

R. P. Ellis and M. M. Miller (2007), Provider payment methods and incentives, Elsevier Inc

R. P. Ellis (1998) Creaming, skimping and dumping: provider competition on the intensive and extensive marhins. Journal of Health Economics 17 (1998) pp. 537-555

Rajiv L. Sharma, (1998) Health-care payment systems: cost and quality incentives–comment, Journal of Economics and Management Strategy 7 (1), 127-137

World Health Statistics (2010), World Health Organization

A Appendix

A.1 Downward distortion of the second best optimal allocations for lower types

In this section, we are going to prove that the second best allocations are downward distorted for the lower types. For this, we have to prove that

$$\Delta_x = \varphi'(e)\frac{\partial e(\alpha P(e,q,\theta),q,\theta)}{\partial \theta} < 0$$
$$\Delta_q = x\varphi'(e)\frac{\partial^2 e(\alpha P(e,q,\theta),q,\theta)}{\partial \theta \partial q} < 0$$
$$\Delta_e = \frac{\partial\left[x\varphi'(e)\frac{\partial e(\alpha P(e,q,\theta),q,\theta)}{\partial \theta}\right]}{\partial e} < 0$$

We first prove that $\Delta_x < 0$. $\frac{\partial e(\alpha P(e,q,\theta),q,\theta)}{\partial \theta}$ is the partial derivative of effort with respect to its own ability θ. By implicit function theorem:

$$\frac{\partial e(\mathcal{P},q,\theta)}{\partial \theta} = -\frac{P_\theta(e,q,\theta)}{P_e(e,q,\theta)} < 0$$

Moreover, $\varphi'(e) > 0$. It follows that $\Delta_x < 0$.

We next prove that $\Delta_e < 0$.

$$\Delta_e = x\varphi''(e)\frac{\partial e(\alpha P(e,q,\theta),q,\theta)}{\partial \theta} + x\varphi'(e)\frac{\partial^2 e(\alpha P(e,q,\theta),q,\theta)}{\partial \theta \partial e}$$

We have shown that

$$\frac{\partial e(\mathcal{P},q,\theta)}{\partial \theta} = -\frac{P_\theta(e,q,\theta)}{P_e(e,q,\theta)} < 0$$

Then,

$$\begin{aligned}
&\frac{\partial^2 e(\alpha P(e,q,\theta),q,\theta)}{\partial \theta \partial e} = \frac{\partial^2 e(\alpha P(e,q,\theta),q,\theta)}{\partial \theta \partial \mathcal{P}}\frac{\partial\big(\alpha P(e,q,\theta)\big)}{\partial e} \\
&= \frac{\partial \frac{\partial e(\mathcal{P},q,\theta)}{\partial \mathcal{P}}}{\partial \theta}\frac{\partial\big(\alpha P(e,q,\theta)\big)}{\partial e} \\
&= \alpha\frac{\partial\left(\frac{1}{\alpha \partial P(e,q,\theta)/\partial e}\right)}{\partial \theta}\frac{\partial P(e,q,\theta)}{\partial e} \\
&= \frac{\partial\left(\frac{1}{\partial P(e,q,\theta)/\partial e}\right)}{\partial \theta}\frac{\partial P(e,q,\theta)}{\partial e} < 0
\end{aligned}$$

by the single-crossing property.

Moreover, $x\varphi''(e) > 0$ and $x\varphi'(e) > 0$ by the convexity and the increasing property of the disutility function. Thus, it follows that $\Delta_e < 0$.

Finally, we prove that $\Delta_q < 0$.

$$\begin{aligned}
&\frac{\partial^2 e(\alpha P(e,q,\theta),q,\theta)}{\partial\theta\partial q} \\
&= \frac{\partial^2 e(\mathcal{P},q,\theta)}{\partial\theta\partial\mathcal{P}}\frac{\partial\big(\alpha P(e,q,\theta)\big)}{\partial q} + \frac{\partial^2 e(\mathcal{P},q,\theta)}{\partial\theta\partial q} \\
&= \alpha\frac{\partial\frac{1}{\alpha\partial P(e,q,\theta)/\partial e}}{\partial\theta}\frac{\partial P(e,q,\theta)}{\partial q} + \frac{\partial\frac{\partial e(\mathcal{P},q,\theta)}{\partial q}}{\partial\theta}
\end{aligned}$$

The first term is negative following the single-crossing property. By implicit function theorem, the second term becomes

$$\frac{\partial\frac{\partial e(\mathcal{P},q,\theta)}{\partial q}}{\partial\theta} = \alpha\frac{1}{P_e}(P_{q\theta} - \frac{P_q}{P_e}P_{e\theta})$$

For CES production function and the exponential function for the recovery probability,

$$\frac{\partial\frac{\partial e(\mathcal{P},q,\theta)}{\partial q}}{\partial\theta} = \alpha\frac{1}{P_e}(P_{q\theta} - \frac{P_q}{P_e}P_{e\theta}) = 0$$

Moreover, $x\varphi'(e) > 0$. It follows that $\Delta_q < 0$.

A.2 Proof of Lemma 2

We first prove Result i).

$$P_q = \theta exp(-\theta f(e,q))f_q(e,q)$$

$$P_e = \theta exp(-\theta f(e,q))f_e(e,q)$$

$$lim_{q\to 0^+,e\to 0^+} f_q(e,q) = lim_{q\to 0^+,e\to 0^+}(1+q^\rho e^{-\rho})^{\frac{1-\rho}{\rho}} = 2^{\frac{1-\rho}{\rho}}$$
$$lim_{q\to 0^+,e\to 0^+} f_e(e,q) = lim_{q\to 0^+,e\to 0^+}(1+e^\rho q^{-\rho})^{\frac{1-\rho}{\rho}} = 2^{\frac{1-\rho}{\rho}}$$

It follows that

$$lim_{q\to 0^+,e\to 0^+} P_q(e,q,\theta) = \theta lim_{q\to 0^+,e\to 0^+} f_q(q,e) = 2^{\frac{1-\rho}{\rho}}\theta$$
$$lim_{q\to 0^+,e\to 0^+} P_e(e,q,\theta) = \theta lim_{q\to 0^+,e\to 0^+} f_e(q,e) = 2^{\frac{1-\rho}{\rho}}\theta$$

Therefore, when $\rho \to 0$ the Inada conditions are satisfied: $\forall\theta$

$$P_q(0,0,\theta) = +\infty$$
$$P_e(0,0,\theta) = +\infty$$

Then, if

$$C'(0) = 0$$
$$\varphi'(0) = 0$$

result i) follows immediately from equation (14)(15).

If Inada conditions are not satisfied:

$$P_q(0,0,\theta) < +\infty$$
$$P_e(0,0,\theta) < +\infty$$

and

$$C'(0) > 0$$
$$\varphi'(0) > 0$$

Result ii) follows from equation (14)(15) with Spence-Mirrlees conditions : $P_q(0,0,\theta), P_e(0,0,\theta)$ are increasing with type. Thus, if there exists $\tilde{\theta}$ such that $\forall x > 0$, $C'(0) \geq \alpha P_q(0,0,\tilde{\theta})b$ and $\varphi'(0) \geq \alpha P_e(0,0,\tilde{\theta})b$, then all types that are lower than $\tilde{\theta}$ will satisfy these inequalities. Then equation (14)(15) imply that for these types, $q = 0$ and $e = 0$. Then, equation (13) implies that $x = 0$.

If $\forall\theta$, $\forall x(\theta) > 0$

$$C'(0) < \alpha P_q(0,0,\theta)b$$
$$\varphi'(0) < \alpha P_e(0,0,\theta)b$$

then, suppose that $x(\theta)$ is interior, since the social welfare function is concave in its arguments, we can obtain $q(\theta) = 0$, $e(\theta) = 0$ if and only if equation (14)(15) are satisfied with inequalities[14] at $q(\theta) = 0$,

[14]The reason why we have to consider both inequalities together is as follows: If among (14)(15), at the optimum, one is satisfied with equality and the other is inequality, which means that one of the allocation (q or e) is zero, the other is interior. If the optimal x is interior, then this type is not excluded which is not the case that we are interested in. If the optimal x is zero, then the assumed equality equation (either (14) or (15)) cannot be an equality. Contradiction, this is impossible. Hence, considering one inequality with the other being equality either is not interesting or is impossible. Exclusion must mean that $q = 0$ and $e = 0$ at the same time. Hence (14)(15) must be satisfied with inequality together to derive the condition of exclusion.

$e(\theta) = 0$. Rewritting these two inequalities evaluated at $q \to 0^+$, $e \to 0^+$:

$$\frac{f(\theta)}{1-F(\theta)} < \frac{-\Delta_q|_{q\to 0^+, e\to 0^+}}{x\alpha P_q(e,q,\theta)|_{q\to 0^+, e\to 0^+} b - xC'(0)} \tag{19}$$

$$\frac{f(\theta)}{1-F(\theta)} < \frac{-\Delta_e|_{q\to 0^+, e\to 0^+}}{x\alpha P_e(e,q,\theta)|_{q\to 0^+, e\to 0^+} b - x\varphi'(0)} \tag{20}$$

Remember that

$$\Delta_e = x\varphi''(e)\frac{\partial e(\alpha P(e,q,\theta),q,\theta)}{\partial \theta} + x\varphi'(e)\frac{\partial^2 e(\alpha P(e,q,\theta),q,\theta)}{\partial \theta \partial e}$$

where

$$\frac{\partial^2 e(\alpha P(e,q,\theta),q,\theta)}{\partial \theta \partial e} = \alpha \frac{\partial\left(\frac{1}{\alpha \partial P(e,q,\theta)/\partial e}\right)}{\partial \theta}\frac{\partial P(e,q,\theta)}{\partial e}$$

and

$$\begin{aligned}\Delta_q &= x\varphi'(e)\frac{\partial^2 e(\alpha P(e,q,\theta),q,\theta)}{\partial \theta \partial q} \\ &= x\varphi'(e)\big[\alpha \frac{\partial \frac{1}{\alpha \partial P(e,q,\theta)/\partial e}}{\partial \theta}\frac{\partial P(e,q,\theta)}{\partial q} + \frac{\partial \frac{\partial e(\mathcal{P},q,\theta)}{\partial q}}{\partial \theta}\big]\end{aligned}$$

Taking $P(e,q,\theta) = 1 - exp\big(-\theta f(e,q)\big)$, where $f(e,q) = (e^\rho + q^\rho)^{\frac{1}{\rho}}$ with $\rho \in (-\infty, 1]$,

$$P_q = \theta exp(-\theta f(e,q)) f_q(e,q)$$

$$P_e = \theta exp(-\theta f(e,q)) f_e(e,q)$$

$$P_{q\theta} = M f_q(e,q)$$

$$P_{e\theta} = M f_e(e,q)$$

where $M = [exp(-\theta f(e,q)) + \theta exp(-\theta f(e,q))(-f(e,q))] f_e(e,q)$. Then, we obtain

$$\frac{\partial \frac{\partial e(\mathcal{P},q,\theta)}{\partial q}}{\partial \theta} = \alpha \frac{1}{P_e}(P_{q\theta} - \frac{P_q}{P_e}P_{e\theta}) = \alpha \frac{1}{P_e}[M f_q(e,q) - M f_q(e,q)] = 0$$

Thus,

$$\begin{aligned}\Delta_q|_{q\to 0^+,e\to 0^+} =& x\alpha\varphi'(0)\frac{\partial \frac{1}{\alpha\partial P(e,q,\theta)/\partial e}}{\partial\theta}\frac{\partial P(e,q,\theta)}{\partial q}\\ =& x\alpha\varphi'(0)\frac{\partial \frac{1}{\alpha\theta exp(-\theta f(e,q))f_e(e,q)}}{\partial\theta}\theta exp(-\theta f(e,q))f_q(e,q)\\ =& -x\alpha\varphi'(0)\frac{1}{\alpha^2\theta^2 exp^2(-\theta f(e,q))f_e^2(e,q)}[\alpha M f_e(e,q)]\theta exp(-\theta f(e,q))f_q(e,q)\\ =& -x\varphi'(0)\frac{1}{\theta exp(-\theta f(e,q))f_e(e,q)}Mf_q(e,q)\\ =& -x\varphi'(0)\frac{[1-\theta f(e,q)]f_q(e,q)}{\theta f_e(e,q)}|_{q\to 0^+,e\to 0^+}\end{aligned}$$

With CES function,

$$\begin{aligned}&lim_{q\to 0^+,e\to 0^+}f(e,q)=0\\ &lim_{q\to 0^+,e\to 0^+}f_q(e,q)=lim_{q\to 0^+,e\to 0^+}(1+q^\rho e^{-\rho})^{\frac{1-\rho}{\rho}}=2^{\frac{1-\rho}{\rho}}\\ &lim_{q\to 0^+,e\to 0^+}f_e(e,q)=lim_{q\to 0^+,e\to 0^+}(1+e^\rho q^{-\rho})^{\frac{1-\rho}{\rho}}=2^{\frac{1-\rho}{\rho}}\end{aligned}$$

Hence,

$$lim_{q\to 0^+,e\to 0^+}\Delta_q=-\frac{1}{\theta}x\varphi'(0)$$

Then, the inverse function of $\mathcal{P}(e,q,\theta;\alpha)$ given α, q and θ is

$$e(\mathcal{P},q,\theta)=\left[\big[-\frac{ln(1-\mathcal{P}/\alpha)}{\theta}\big]^\rho-q^\rho\right]^{\frac{1}{\rho}}$$

then,

$$\begin{aligned}&\frac{\partial e(\mathcal{P},q,\theta)}{\partial\theta}|_{q\to 0^+,e\to 0^+}\\ &=\frac{1}{\rho}\left[\big[-\frac{ln(1-\mathcal{P}/\alpha)}{\theta}\big]^\rho-q^\rho\right]^{\frac{1}{\rho}-1}\rho\big[-\frac{ln(1-\mathcal{P}/\alpha)}{\theta}\big]^{\rho-1}[-ln(1-\mathcal{P}/\alpha)(-\frac{1}{\theta^2})]|_{q\to 0^+,e\to 0^+}\\ &=0\end{aligned}$$

Therefore,

$$\begin{aligned}
lim_{q\to 0^+,e\to 0^+}\Delta_e &= x\alpha\varphi'(0)\frac{\partial\frac{1}{\alpha\partial P(e,q,\theta)/\partial e}}{\partial\theta}\frac{\partial P(e,q,\theta)}{\partial e}\\
&= -x\alpha\varphi'(0)\frac{1}{\alpha^2\theta^2 exp^2(-\theta f(e,q))f_e^2(e,q)}[\alpha M f_e(e,q)]\theta exp(-\theta f(e,q))f_e(e,q)\\
&= -x\varphi'(0)\frac{[1-\theta f(e,q)]f_e(e,q)}{\theta f_e(e,q)}\\
&= -x\varphi'(0)\frac{1-\theta f(e,q)}{\theta}|_{q\to 0^+,e\to 0^+}\\
&= -\frac{1}{\theta}x\varphi'(0)
\end{aligned}$$

Thus

$$lim_{q\to 0^+,e\to 0^+}\Delta_q = lim_{q\to 0^+,e\to 0^+}\Delta_e = -\frac{1}{\theta}x\varphi'(0)$$

Moreover,

$$\begin{aligned}
lim_{q\to 0^+,e\to 0^+}P_q(e,q,\theta) &= \theta lim_{q\to 0^+,e\to 0^+}f_q(q,e) = 2^{\frac{1-\rho}{\rho}}\theta\\
lim_{q\to 0^+,e\to 0^+}P_e(e,q,\theta) &= \theta lim_{q\to 0^+,e\to 0^+}f_e(q,e) = 2^{\frac{1-\rho}{\rho}}\theta
\end{aligned}$$

Hence inequalities (19)(20) evaluated at zero become

$$\frac{f(\theta)}{1-F(\theta)} < \frac{x\varphi'(0)}{\theta(2^{\frac{1-\rho}{\rho}}\alpha\theta xb - xC'(0))} \tag{21}$$

$$\frac{f(\theta)}{1-F(\theta)} < \frac{x\varphi'(0)}{\theta(2^{\frac{1-\rho}{\rho}}\alpha\theta xb - x\varphi'(0))} \tag{22}$$

It follows that if

$$\frac{f(\theta)}{1-F(\theta)} < Min\left\{\frac{\varphi'(0)}{\theta(2^{\frac{1-\rho}{\rho}}\alpha\theta b - C'(0))}, \frac{\varphi'(0)}{\theta(2^{\frac{1-\rho}{\rho}}\alpha\theta b - \varphi'(0))}\right\} \tag{23}$$

is satisfied, inequalities (21)(22) are satisfied. Thus, the original equation (14)(15) are both satisfied with inequalities at $q = 0$ and $e = 0$. Thus, the optimal solutions are indeed $q = 0$ and $e = 0$ for the type θ. Hence, from equation (23), for θ which is close to zero and $f(\theta)$ is close to zero, this condition is surely satisfied because the left hand side is close to zero and the right hand side goes to infinity.

Then, with monotonicity assumption of the hazard rate $\frac{1-F(\theta)}{f(\theta)}$, the left hand side of equation (23) is increasing in θ and the right hand side is decreasing in θ, thus for all types that are lower than this θ, this

condition is also satisfied, which implies $q = 0$, $e = 0$ for these types too. When the optimal treatment and effort are zero, equation (13) with inequality implies that the optimal number of patients is zero. Which contradicts what we supposed at the beginning that x is interior. Consequently, with condition (23) being satisfied, it is impossible to have interior solution. It thus proved the iii) of Lemma 2.

A.3 Second best fee-for-service is positive for lower types

Equation (17) is:

$$S_q = \frac{1 - F(\theta)}{f(\theta)}(-|\Delta_q| + |\Delta_e|\frac{P_q}{P_e})$$

We are going to prove that $-|\Delta_q| + |\Delta_e|\frac{P_q}{P_e} = \Delta_q - \Delta_e\frac{P_q}{P_e} > 0$[15].

$$\Delta_q = x\varphi'(e)\big[\frac{\partial^2 e(\mathcal{P}, q, \theta)}{\partial\theta\partial\mathcal{P}}\frac{\partial\big(\alpha P(e, q, \theta)\big)}{\partial q} + \frac{\partial^2 e(\mathcal{P}, q, \theta)}{\partial\theta\partial q}\big]$$

where $\frac{\partial^2 e(\mathcal{P},q,\theta)}{\partial\theta\partial q} = 0$ under CES production function[16].

$$\Delta_e = x\varphi''(e)\frac{\partial e(\alpha P(e, q, \theta), q, \theta)}{\partial\theta} + x\varphi'(e)\frac{\partial^2 e(\mathcal{P}, q, \theta)}{\partial\theta\partial\mathcal{P}}\frac{\partial\big(\alpha P(e, q, \theta)\big)}{\partial e}$$

Thus,

$$\begin{aligned}
&\Delta_q - \Delta_e\frac{P_q}{P_e} \\
&= x\varphi'(e)\frac{\partial^2 e(\mathcal{P}, q, \theta)}{\partial\theta\partial\mathcal{P}}\frac{\partial\big(\alpha P(e, q, \theta)\big)}{\partial q} \\
&-\left(x\varphi''(e)\frac{\partial e(\alpha P(e, q, \theta), q, \theta)}{\partial\theta} + x\varphi'(e)\frac{\partial^2 e(\mathcal{P}, q, \theta)}{\partial\theta\partial\mathcal{P}}\frac{\partial\big(\alpha P(e, q, \theta)\big)}{\partial e}\right)\frac{\alpha P_q}{\alpha P_e} \\
&= -x\varphi''(e)\frac{\partial e(\alpha P(e, q, \theta), q, \theta)}{\partial\theta}\frac{P_q}{P_e} > 0
\end{aligned}$$

because $\frac{\partial e(\mathcal{P},q,\theta)}{\partial\theta} < 0$, $x\varphi''(e) > 0$, $P_q > 0$ and $P_e > 0$.

As a result, $S_q > 0, \forall\tilde{\theta} < \theta < \theta_1$.

[15] Notation: $P_e = \frac{\partial P(e,q,\theta)}{\partial e}$

[16] If under other functions, this term is not zero, then the Δ_q may or may not be negative. The treatment may be downward or upward distorted. If this term is negative, treatment is downward distorted and fee-for-service is positive as we prove in this section. If this term is positive and large enough, $\Delta_q > 0$, the treatment is upward distorted. The term $\Delta_q - \Delta_e\frac{P_q}{P_e}$ will be surely positive because Δ_e is negative under any form of probability function. Hence, fee-for-service is still positive.