The importance of this paper is based on two facts:

1. Changes in Health Care markets include complex contracts between insurer-physician and insurer-patient. I.e ‘Capitation Contracts’ or ‘Contracts based on some measures of quality of treatment’.

2. These complex arrangements are not explained by previous literature. The need of a new theory arises.

   • Kenneth Arrow (1963) observed that an efficient (FIRST-BEST) health insurance policy would specify payment to the patient contingent on his state of health. Then, the patient would make his own decision to purchase health care. This payment scheme protects the patient from the financial risk ex ante and retains incentives for the patient to consume health care efficiently ex-post. **However** insurance policies contingent on health status are nonexistent because health status is too costly to verify. As a result, the market for health-status-based insurance is missing and: TREATMENT ITSELF IS NOT CONTRACTIBLE.

   • Zeckhauser (1970) described a model of interaction between a physician and her patient (as this paper does) but with the assumption that
     o the level of effort provided from the physician in the production of health care is contractible (In reality, the market for contracts based on level of effort is missing, effort is NOT OBSERVABLE for insurer, thus not contractible),
     o Physician and patient report truthfully, so they are paid according to the true quantity of treatment. (McGuire and Pauly argues that this truthful report can be attained IF AND ONLY IF the insurer imposes some restrictions on physician and patient’s payment parameters (patient shares some costs of the treatment and physician gets a partial reimbursement of the cost of treatment).).
SECTION I: THE MODEL

Extensive form of the game

1. Insurer
   - Chooses elements of insurance and payment system

2. Nature
   - Individual is sick
   - Individual is healthy

3. Physician
   - Chooses level of effort $e$

4. Patient
   - Chooses quantity of treatment $\tau$

5. Patient and Physician
   - Play a REPORTING SUBGAME
   - Financial terms of contract will be settled.

Reporting sub-game

- Physician
  - $\tau^R \neq \tau$
  - $\tau^R = \tau$

- Patient
  - Agrees
    - $\tau^R \neq \tau$
  - Disagrees
    - $\tau^R = \tau$
INTERPRETATION of the Demand response:

We can assume that a demand response of quantity of treatment at any level of physician’s effort, exists. In other words, by choosing the level of effort, the physician can induce demand of treatment from the patient.

Examples:

- Long term relationship between some physicians and their patients.
- Physicians with ‘acute specialties’ may still change their effort, alter information available to potential patients and create demand response.

ELEMENTS OF THE GAME:

1. **PATIENT’S UTILITY**

Assumption: The patient is risk averse in income. Thus, the utility function is strictly concave.

\[
EU = p \cdot U(w - \alpha - \beta \tau^R - s + F(\tau, \varepsilon)) + (1 - p) \cdot U(w - \alpha)
\]

\(EU\): Expected utility.

\(p\): Probability of getting sick

\(w\): Initial income.

\(\alpha\): Insurance premium. \((\alpha \geq 0)\)

\(\beta\): patient’s copayment per unit of reported quantity of treatment. \((0 \leq \beta \leq c\) where \(c\) is the marginal cost of treatment)

\(\tau^R\): quantity of treatment reported.

\(s\): Monetary equivalent of health shock when sick.

\(F(\tau, \varepsilon)\): Monetary equivalent of benefits to treatment.

2. **PHYSICIAN’S PREFERENCE**

The physician is risk neutral, has a utility function separable in money and effort, and a reservation utility normalized at zero.

\[
V = p\left[\rho + (\delta + c)\tau^R - ct - G(\varepsilon)\right]
\]

\(V\): Indirect utility function. Linear in income, quasilinear in the “price of effort”.
\( p \): Probability of getting sick

\( \rho \): Fixed fee per patient. “Prospective payment”.

\( \delta + c \): Reimbursement for each unit of reported treatment (\( \tau^R \)) where \( c \) is the marginal cost of treatment.

If \( \delta + c = 0 \), “Fully prospective system”.

If \( \delta < 0 \), “Payment system with supply side cost sharing”.

\( \tau \): real quantity of treatment.

\( G(\varepsilon) \): Physician incurs in some disutility in supplying effort. Effort is costly. \( G \) is monotonic increasing and convex in level of effort.

3. **INSURER’S PROBLEM**

The insurer maximizes the patient’s expected utility subject to a balanced budget.

This implies that the premium paid by the patient must equal the expected value of the insurer’s payment to physician.

\[
\alpha = p \times [\rho + (\delta + c - \beta)\tau^R]
\]  

(3)

This last equation can also be seen as:

\[
\alpha + \beta\tau^R = p \times [\rho + (\delta + c)\tau^R]
\]

Expected payment from patient. Expected reimbursement to physician.

**CONSTRAINTS IMPOSED ON PAYMENT PARAMETERS** (Table of incentives)

<table>
<thead>
<tr>
<th>PATIENT’S COPAYMENT</th>
<th>( \beta \geq 0 )</th>
<th>( \beta &lt; 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHYSICIAN’S REIMBURSEMENT</td>
<td>( \delta + c \geq 0 )</td>
<td>( \delta + c &lt; 0 )</td>
</tr>
<tr>
<td>Not Underreport, Underreport</td>
<td>( \tau^R = \tau )</td>
<td>Underreport, Underreport</td>
</tr>
<tr>
<td>Underreport, Not underreport</td>
<td>( \tau^R = 0 &lt; \tau )</td>
<td>Underreport, Not underreport</td>
</tr>
<tr>
<td>Not Underreport, Not underreport</td>
<td>( \tau^R = \tau )</td>
<td>Underreport, Not underreport</td>
</tr>
</tbody>
</table>
Section II: Patient Choice of Quantity and Physician Choice of Effort

PATIENT’S CHOICE

After the physician decides the level of effort $\varepsilon$ to put into a treatment in stage 3, the patient needs to choose $\tau$ to maximize the utility in stage 4. The patient’s utility function can be written as following:

$$U(w - \alpha - \beta \tau - s + F(\tau, \varepsilon)) \quad (1A)$$

Where $w$ is the wage of the patient; $\alpha$ is the insurance premium; $\beta \tau$ is the gross payment that the patient needs to pay for the treatment; $s$ is the monetary equivalent of the health shock, which can be also considered as the opportunity cost when the patient is ill; $F(\tau, \varepsilon)$ is the monetary benefit gained from a treatment, it depends on quantity of the and effort of the treatment.

To maximize patient’s utility (1A), take the first derivative with respect to $\tau$, the necessary and sufficient first-order condition is obtained as follow:

$$U'(w - \alpha - \beta \tau - s + F(\tau, \varepsilon))(-\beta + F'_{\tau}(\tau, \varepsilon)) = 0$$

$$U' \neq 0$$

$$\beta = F'_{\tau}(\tau, \varepsilon) \quad (4)$$

Equation (4) is the patient’s reaction function according to the physician’s decision. The left hand side $\beta$ is the marginal copayment, while the right hand side $F'_{\tau}(\tau, \varepsilon)$ is the marginal benefit gained from each treatment. When setting the marginal cost equals the marginal benefit, the patient utility reaches the maximum.

Take the total derivative of equation (4), yields:

$$0 = F'_{\tau\tau} \cdot d\varepsilon + F'_{\tau\tau} \cdot d\tau$$

$$\frac{d\tau}{d\varepsilon} = -\frac{F'_{\tau\tau}}{F'_{\tau\tau}} \quad (5)$$
F(τ, ε) is strictly concave, which implies the marginal benefit gained from the treatment is decreasing. So the sign of \( F_{ee} \) is always negative. The sign of the reaction function is the same as the sign of \( F_{ee} \).

If effort and treatment quantity are substitutes, then \( F_{ee} < 0, \frac{d\tau}{d\epsilon} < 0 \). When \( \tau \) increases, \( \epsilon \) decreases. To induce the patient to demand a higher \( \tau \), a lower \( \epsilon \) will be necessary. Alternatively, if effort and treatment are complements, meaning that \( F_{ee} > 0, \frac{d\tau}{d\epsilon} > 0 \). When \( \tau \) increases, the marginal benefit of effort \( \epsilon \) increases as well, then a higher \( \tau \) can only be induced by a higher \( \epsilon \).

**PHYSICIAN’S CHOICE**

Now the paper starts to consider the physician’s decision. Anticipating the patient’s treatment quantity reaction in stage 4, the physician chooses \( \epsilon \) to maximize the revenue less costs.

Consider in a subgame-perfect equilibrium, the physician’s choice of \( \epsilon \) and the patient’s subsequent choice of \( \tau \) are given the solution of the following Program A: for \( 0 \leq \beta \leq c \) and \( \beta \geq -c \), choose \( \epsilon \) and \( \tau \) to maximize

\[
\rho + \delta \tau - G(\epsilon) \tag{6}
\]

Subject to

\[
\beta = F_{\tau}(\tau, \epsilon)
\]

Equation (6) is the normalized physician’s utility function when the patient is sick. The physician is individual’s rational and need to maximize equation (6).

The physician’s prospective payment \( \rho \) is fixed, even in the worst case that the reimbursement of the physician is 0 (\( \delta + c = 0 \)), the physician still get the prospective payment, so \( \rho \) does not affect physician’s supply of effort. Only \( \beta \) and \( \delta \) appear as parameters in Program A.

Define \( \Omega \) as the solution set of Program A, and all the pairs \( (\tau, \epsilon) \) in \( \Omega \) can arise as subgame-perfect equilibrium:

\[
\Omega = \{ (\tau, \epsilon) : \text{there exist } (\beta, \delta), \text{with } 0 \leq \beta \leq c \text{ and } \beta \geq -c, \text{ for which } (\tau, \epsilon) \text{ solves Program A given } (\beta, \delta) \}\}

Take the first-order derivative of the objective function (6) with respect to \( \epsilon \):
According to equation (5), (A2) yields:

\[ \delta \frac{d\tau}{d\epsilon} - G'(\epsilon) \quad (A2) \]

To keep (7) negative, if \( \tau \) and \( \epsilon \) are substitutes (\( F_{\tau\epsilon} < 0 \)) then \( \delta \geq 0 \), or if \( \tau \) and \( \epsilon \) are complements (\( F_{\tau\epsilon} > 0 \)), then \( \delta \leq 0 \).

**PROPOSITION 1**

Consider a treatment quantity-physician effort pair \((\tau, \epsilon)\) belongs to the implementable set \( \Omega \); suppose at \((\tau, \epsilon)\), \( \epsilon \) and \( \tau \) are substitutes, then \( \epsilon \) is above its minimum if and only in the physician payment system includes supply-side cost sharing (\( \delta < 0 \)). Alternatively, suppose at \((\tau, \epsilon)\), \( \epsilon \) and \( \tau \) are complements, then \( \epsilon \) is above its minimum if and only if the physician payment system does not include the supply-side cost sharing (\( \delta > 0 \)).

Proposition 1 implies: Under the case of substitute, \( F_{\tau\epsilon} < 0 \), if \( \delta > 0 \), the physician gets paid for each treatment, so the physician’s utility increases only if \( \tau \) increases. When \( F_{\tau\epsilon} < 0 \), increase in \( \tau \) can only result from a decrease in \( \epsilon \). In order to maximize utility, the physician will therefore reduce effort \( \epsilon \) as much as she can if \( \delta > 0 \) and \( \epsilon \) may drop below 0. However, the minimum requirement for the effort is \( \epsilon \geq 0 \). It follows that in case of substitute, \( \delta \) must be negative, which means the physician needs to bear some of the costs of the treatment. Thus, \( \delta < 0 \).

Under the case of complement, \( F_{\tau\epsilon} > 0 \), if \( \delta < 0 \), the physician loses money in each treatment, the physician’s utility increases only if \( \tau \) decreases. When \( F_{\tau\epsilon} > 0 \), decrease in \( \tau \) can only result from a decrease in \( \epsilon \). Therefore, the physician will reduce effort \( \epsilon \) as much as she can if \( \delta < 0 \). However, in order to keep \( \epsilon \geq 0 \). It follows that \( \delta > 0 \), in other words, the physician needs to be reimbursed more than the costs of the requirement.

The \((\tau, \epsilon)\) must be carefully chosen that in fact the physician and the patient have the incentives to select the effort and treatment in equilibrium.

According to proposition 1, if effort and quantity are complements, \( \delta \geq 0 \), then the constraint for the truthful reporting equilibrium \( \delta \geq -c \) is not relevant for the pair to be implementable. The patient has the
incentive to under-report the truthful quantity of the treatment. On the other hand, the physician has
the incentive to report more quantities of the treatment. So equilibrium will be reached, the truthful
quantity is reported in the case of complement.

If effort and quantity are substitutes, the margin $\delta$ must be negative. The patient still wants to report
less quantity of the treatment; however, the physician has the incentive to under report as well. If no
other constraint exists here, the physician and the patient might not reach the equilibrium. The
condition $\delta<0$ for the subgame-perfect equilibrium puts a limit on the penalty the payment system can
impose. The situation that $\delta<0$ will never be the equilibrium.

Assume that the first-order conditions for Program A are necessary and sufficient for its solution.

From equation (7), at the optimal point:

$$\frac{F_{t\varepsilon}(\tau, \varepsilon)}{F_{t\tau}(\tau, \varepsilon)} \delta + G'(\varepsilon) = 0$$

Multiply $F_{t\tau}$ then:

$$F_{t\tau} \delta + G'(\varepsilon)F_{t\tau} = 0$$

Take the partial derivative with respect to $\delta$:

$$\delta F_{t\varepsilon} \frac{\partial \tau}{\partial \delta} + \delta F_{t\varepsilon} \frac{\partial \varepsilon}{\partial \delta} + F_{t\varepsilon} + F_{t\varepsilon} G' \frac{\partial \tau}{\partial \delta} + F_{t\varepsilon} \frac{\partial \varepsilon}{\partial \delta} + G' F_{t\varepsilon} \frac{\partial \varepsilon}{\partial \delta} = 0 \quad (A3)$$

From equation (4), take derivative with respect to $\delta$:

$$F_{t\tau} \frac{\partial \tau}{\partial \delta} + F_{t\varepsilon} \frac{\partial \varepsilon}{\partial \delta} = 0 \quad (A4)$$

From the (A3) and (A4) above, treat $\frac{\partial \varepsilon}{\partial \delta}$ and $\frac{\partial \tau}{\partial \delta}$ as unknown variables and solve the equation, yields:

$$\frac{\partial \tau}{\partial \delta} = \frac{-F_{t\varepsilon}^2}{F_{t\tau}[5F_{t\varepsilon} + F_{t\varepsilon}G]} - \frac{F_{t\varepsilon}^2}{F_{t\tau} [5F_{t\varepsilon} + F_{t\varepsilon}G + G'] F_{t\tau}} \quad > 0 \quad (8)$$

$$\frac{\partial \varepsilon}{\partial \delta} = \frac{F_{t\varepsilon} F_{t\tau}}{F_{t\tau} [5F_{t\varepsilon} + F_{t\varepsilon}G]} - \frac{F_{t\varepsilon} F_{t\tau}}{F_{t\tau} [5F_{t\varepsilon} + F_{t\varepsilon}G + G'] F_{t\tau}} \quad < 0$$

where $H = F_{t\tau} [5F_{t\varepsilon} + F_{t\varepsilon}G] - F_{t\tau} [5F_{t\varepsilon} + F_{t\varepsilon}G + G'] F_{t\tau}$, is the bordered Hessian.
From equation (8), if the $\delta$ decreases, then $\varepsilon$ increases. Back to the case when quantity and effort are substitutes, when $\delta<-c$, the equilibrium will never reached in $\Omega$. It also implies that very high effort $\varepsilon$ may not belong to the implementable set because more negative $\delta$ (such as $\delta<-c$) requires higher $\varepsilon$. The physician does not have the incentive to put more effort in treatment.

![Figure 1. The Set of Implementable Effort and Treatment Levels](image)

Figure 1 illustrates the implementable set under substitutes. The patient’s reaction functions ($F_{\tau} = \beta = 0$ and $F_{\tau} = \beta = c$) are negatively sloped, which is implied by equation (5). The physician’s indifferent curves are also shown for three sets of value of $\rho$ and $\delta$. The tangencies of these two kinds of curves are the equilibrium ($\varepsilon', \tau'$). The shaded area is the $\Omega$ set, all the points inside satisfied $\delta \leq c$; the points on the boundary of the $\Omega$ set yield $\delta = -c$; the points to the east of the boundary yield $\delta < -c$, which implies in the case of substitute, larger effort is infeasible.

**Section III: Optimal Insurance Payment Systems**

**Overview**

Having solved for stage 5 (the reporting subgame) as well as 4 and 3 (the physician-patient interaction that determines effort and quantity of treatment) the authors subsequently determine the optimal insurance payment system which is determined in stage 1 of the game. Thus they continue with backward induction by taking into account the equilibrium outcomes of the prior stages of the game (stage 2 of the game is not solved as here nature determines whether the patient is sick or not).
Take Away
The authors prove that even though neither of the two inputs of health care production (effort and quantity of treatment) are contractible (these two constraints constitute a third best regime\(^1\)), a second best solution may still be obtained if effort and quantity of treatment are complements. Furthermore, whenever this is possible the implementation does not lead to extra cost, even when the truthful reporting requirement is fulfilled \((\delta \geq -c)\).

When the truth telling constraint binds \((\delta = -c)\), however, the second best solution cannot be obtained at all.

The authors start by computing the third best solution and proceed with the case in which a second best solution can be obtained under the third best regime.

Obtaining a Third Best Solution in a Third Best Regime
In general the optimal insurance payment system is obtained by choosing the insurance payment parameters that maximize the patient’s expected utility. A possible interpretation for maximizing the patient’s utility is that the paper is written from the perspective of a social planner who’s goal it is to minimize social cost and improve welfare, i.e. minimizing the insurance payments by the patients while obtaining as much health benefits as possible, thus maximizing their utility.

In a first step the two optimality conditions, that the premium for the patient is actually fair (corresponds to the break-even or insurer’s budget constraint (3)) and that the physician’s normalized expected utility when the patient is sick is at least her reservation utility (i.e. zero) (corresponds to set (2) equal to zero), are combined to

\[
p[\epsilon \tau + G(\epsilon)] = \alpha + p\beta \tau,
\]

which sets the total expected cost of treatment (the probability of being sick times the total costs for the units of treatment plus the cost of effort) equal to the total premium and expected total co-payment from the patient.

This is obtained by setting \(p = G(\epsilon) - \delta \tau\) which ensures that (3), the insurer’s budget constraint, holds and (2), the physician’s expected utility, equals her reservation utility (zero).

As \(p\) can always be set in a way to extract any rent from the physician (see equation (6), the physician’s utility function) this holds for any arbitrary \(\delta\).

---

\(^1\) First Best (Efficient) Solution: Obtained from a maximization problem without constraints on the optimality condition

Second Best Solution: In this case obtained from a maximization problem with one constraint on the optimality condition (e.g. noncontractibility of the quantity of treatment)

Third Best Solution: In this case obtained from a maximization problem with two constraints on the optimality condition (noncontractibility of the physician’s effort and the quantity of treatment demanded by the patient)
Hence, \( \delta \) can be used to obtain any effort quantity pair in the subgame in stage 3 and one can maximize (1) subject to (9) and \( (\varepsilon, \tau) \in \Omega \).

The parameters \( \varepsilon \) and \( \tau \) are already given by the equilibrium outcome of the subgame in stage 3 of the game and are described by the implementable set \( (\varepsilon, \tau) \in \Omega (\delta \text{ is only part of this definition, not of (9))} \).

Neither effort nor quantity of treatment is contractible in this third best regime and the maximizations of the patient’s utility s.t. (9) and \( (\varepsilon, \tau) \in \Omega \) leads to a third best solution.

**Proposition 2: Under Certain Conditions a Second Best Solution Can Be Obtained in a Third Best Regime**

In order to show that under certain conditions a second best solution can still be obtained, a benchmark regime is considered that allows for the contractibility of the effort level \( \varepsilon \). The quantity \( \tau \) remains noncontractible.

Thus it follows that the constraint \( (\varepsilon, \tau) \in \Omega \) is not needed any more to reach equilibrium, as the insurance company can now force the physician to exert a certain effort level by paying her contingent on this parameter.

Hence, under this new regime, (1) is maximized subject to (9) and (4), as in this relaxed version it is only necessary for the patient to set

\[
\beta = F_\tau (\tau, \varepsilon),
\]

i.e. the marginal cost of treatment must equal the marginal benefit, instead of taking into account \( \Omega \) ((4) is the first order derivative of the patients expected utility function).

The solution is clearly a second best equilibrium, as effort is contractible.

The authors make their point by comparing the obtained second best solution with the third best solution from above. This comparison reveals that if \( \delta^{SB} \), which is obtained under the second best regime, is larger than or equal to \(-c\), i.e. \( \delta^{SB} \geq -c \), such that the effort and quantity of treatment pair belongs to the implementable set omega of the third best regime, \( (\varepsilon^{SB}, \tau^{SB}) \in \Omega \), the noncontractibility of the physician’s effort has no consequences and the maximization can be done as if it was in the second best regime.

The only difference between the maximizations under the second and the third best regime is the existence of the constraint expressed by \( \Omega \) in the latter one, which implies (4), but not vice versa.

However, the second best solution cannot be obtained when the constraint on \( \delta \) binds, i.e. \( \delta = -c \), as then the second best solution lies outside \( \Omega \) which incorporates the constraint on \( \delta \). In this case the patient’s equilibrium expected utility will be strictly inferior to the second best case.

As the constraints for \( \delta \) are different for the case in which effort and units of treatment are complements and substitutes, this feature separates these two cases.
When $\varepsilon$ and $\tau$ are complements, the implementation of any positive effort level requires $\delta > 0$ which does not conflict with the constraint $\delta \geq -c$.

When $\varepsilon$ and $\tau$ are substitutes, however, a positive level of effort requires supply-side cost sharing, or setting $\delta < 0$.

As in some cases a second best solution requires $\delta$ to be strictly larger than $-c$, these two conditions might conflict and a second best solution cannot be obtained.

**Summary Section III**

Thus, in summary, with effort and quantity of treatment being complements, the second best solution can be obtained under the third best regime, even though the two inputs are not contractible. The reason for this is that the set of possible equilibria of the third best regime in this case contains the equilibrium of the second best regime.

**Section IV: A Third-Best Equilibrium and Ethical Behavior**

**Overview**

To make their point that ethical behavior might render the results and improve a patient’s utility, even though both effort and quantity of treatment are not contractible, the authors continue by analyzing the case in which the second best solution cannot be obtained. This is the case when $\tau$ and $\varepsilon$ are substitutes, thus the cross partial derivative of the health care production function is negative ($F_{\tau \varepsilon} < 0$) and the constraint on $\delta$ binds ($\delta = -c$) with equality.

If this was not the case, circumstances might arise under which the physicians were punished per unit of treatment ($\delta < -c$) and incentives were not sufficient to induce a positive level of effort.
Proposition 3: If the constraint on $\delta$ binds ($\delta = -c$) and $\varepsilon$ and $\tau$ are substitutes, the second best solution cannot be obtained

Assuming that the third best equilibrium $A$, the constraint on $\delta$ binds (i.e. $\delta = -c$), the second best solution can never be part of the third best regime if $\varepsilon$ and $\tau$ are substitutes as the combination of $\varepsilon$ and $\tau$ that leads to the second best solution $B$ lies outside the implementable set $\Omega$.

In order to render this limit the authors introduce an ethics constraint that makes physicians exert a level of effort that leads to a minimum level of health care benefit for the patient, i.e. $F(\varepsilon, \tau) \geq \bar{F}$, where $\bar{F}$ is a constant.

With this lower bound, the physician is required, whether out of altruistic behavior or due to an implicit consensus as exchange for professional autonomy, to exert an amount of effort that produces a certain health benefit $\bar{F}$.

This does not necessarily have to bind as with a sufficiently low copayment per unit of treatment $\beta$ (and a associated higher demand of $\tau$) and a certain level of $\delta$, the patient might already receive a benefit of at least $\bar{F}$.

As it is however of special interest to analyze the cases in which the ethical constraint binds, proposition 3 establishes the case under which the second best solution will never be obtained. As described above, this case can only occur when effort and quantity of treatment are substitutes, as then in equilibrium the constraint on $\delta$, $\delta \geq -c$, binds, i.e. $\delta = -c$.

It furthermore states that if only a third best solution, i.e. point $A$ in the figure, is obtainable, the patient’s utility will be strictly lower than the utility gained from the second best solution, depicted by point $B$.

One can evaluate this intuitively with a comparison of the parameters that make up the patient’s utility function (1), i.e. $w, \alpha, \beta, \tau, s, \varepsilon$.

As $w$ and $s$ are exogenously determined they are fixed and are not different under solution $A$ and $B$. Furthermore, as $A$ and $B$ lie on the same reaction curve, they have the same $\beta$. The quantity of treatment, $\tau$ is decreased at the same time (reducing risk to the patient), thus increasing the patient’s utility by reducing the total amount of copayments, while $\varepsilon$ is increased, as they are substitutes.

As $\beta$ additionally is on a lower isosocial cost curve, the cost of treatment is reduced.

In conclusion, all components of the utility function improve with a decrease in $\delta$. By this, however, the constraint $\delta \geq -c$ is violated and the second best solution $B$ can never be attained under the third best regime if $\varepsilon$ and $\tau$ are substitutes and the patient’s utility under the third best solution is strictly lower than with the second best solution.

The formal proof is given in the appendix of the paper (see pp. 702 f.). In order to clarify the steps, however, we show in the following how to derive (14):

\[
F_{\tau\tau} d\tau + F_{\tau\varepsilon} d\varepsilon > \frac{F_s}{F_{\varepsilon}} [F_{\varepsilon} d\tau + F_{\varepsilon} d\varepsilon].
\]
From the hypothesis (see p. 702)

\[- \frac{F_{\tau\tau}}{F_{\tau}} > - \frac{F_{\tau}}{F_{\tau}} \]

and thus

\[\frac{F_{\tau\tau}}{F_{\tau}} < \frac{F_{\tau}}{F_{\tau}} \]

Because \( F \) is strictly concave, so \( F_{\tau\tau} < 0 \), we get

\((*)\)

\[F_{\tau\tau} > \frac{F_{\tau}}{F_{\tau}} \]

The right hand side of (14) equals

\[\frac{F_{\tau\tau}}{F_{\tau}} [F_{\tau} d\tau + F_{\epsilon} d\epsilon] = F_{\tau\tau} d\tau + \frac{F_{\tau}}{F_{\tau}} F_{\tau\tau} d\epsilon\]

The left hand side of (14) equals

\[F_{\tau\tau} d\tau + F_{\tau\tau} d\epsilon\]

Plugging (*) into the left hand side reveals that the left hand side is bigger than the right hand side.

**Ethics Constraint & Proposition 4:**

**Possible Reasons for Physicians**

It is possible and practical to impose ethics constraint on health insurance market. (1) As the paper indicated, physicians may be “accustomed” to provide health care service upper certain level, which means that they want to keep the convention that exists for many years. (2) Modern development of medical technology simulates physicians with high effort. (3) If physicians keep the patients’ benefit upon certain level, they can gain excellent reputations, and even attract more patients so that profit of physicians is ensured.

**Interpretation of Ethics Constraint**

Obviously that the ethics constraint provide a lower bound for patient’s benefit, *i.e.,*

\[F(\tau, \epsilon) \geq \overline{F}\]

In which \( \overline{F} \) is the lower bound. By imposing this lower bound, we can intuitively predict that the patient can gain more expected utility therefore improve their situations. In fact, by providing this lower bound, there is also a lower bound for physicians’ effort as shown in Figure 3. If we can get equilibrium in ethics
constraint regime, the new equilibrium should increase physicians’ effort $\varepsilon$ compared with that in the third best regime.

However, by proposition 3 we know that the patients’ expected utility reaches maximum within the boundary of implemental set $\Omega$. To further improve patients’ expected utility, the only way we can do is to relax the constraint of implemental set $\Omega$, which means untruthful reports are allowed. Although both physicians and physicians have incentive to underreport the quantity of treatment $\tau$, it can be proved that ethics constraint bring more benefits for patients. As a result, patients earn higher expected utility.

**Proposition 4:**

Suppose $F_{\varepsilon\tau} < 0$, and that for all $\varepsilon$ and $\tau$

$$\frac{F_{\varepsilon\tau}(\tau, \varepsilon)}{F_{\tau}(\tau, \varepsilon)} > \frac{F_{\varepsilon}(\tau, \varepsilon)}{F_{\tau}(\tau, \varepsilon)}$$

Under the ethics constraint $F(\tau, \varepsilon) \geq F(\tau^*, \varepsilon^*) \equiv F^*$, the patient’s equilibrium expected utility must be higher than the equilibrium expected utility without the ethics constraint ($EU^*$), and $\beta$ must increase above $\beta^*$.

**Interpretation of Proposition 4**

In proposition 4, $-\frac{F_{\varepsilon\tau}(\tau, \varepsilon)}{F_{\tau}(\tau, \varepsilon)}$ is the slope of isobenefit line $F^*$ while $-\frac{F_{\varepsilon}(\tau, \varepsilon)}{F_{\tau}(\tau, \varepsilon)}$ is the slope of patient’s reaction line $F_{\tau} = \beta$ in Figure 3. Proposition 4 hypothesizes that isobenefit line is steeper than patients’ reaction line. Under this hypothesis, keeping $\beta$ unchanged $F^\triangle$ must yield greater benefit than $F^*$ for patients as shown in Figure 3. With ethic constraint the implemental set can be relaxed. As a result, the new equilibrium improves patients’ expected utility although copayment is higher than that in the “third best” (Proposition 3).
Section V: Extension

Physician-Patient reporting with collusion possibilities (Side-contract Regime)

There are two kinds of situations, no collusion exists or side-contracting is allowed. Previous sections discussed the former situation, while the next section starts to explore how side-contracting changes the analysis above, intuitively. Although both situations are extreme cases, they are helpful for simplifying the analysis.

“Side-contract”

Side-contract refers to information misrepresentation resulting from a group incentive. We also assume that side-contracting is costless so that with positive profit both physicians and patients tend to enter the side-contract regime and agree with untruthful reports. Since implementation of the insurance contract is based on what τ the patient reports, if neither physician nor patient can get further benefit from untruthful report, no collusion occurs.

\[ \delta + c \] represents physicians reimbursement, while \[ \beta \] is patient’s copayment as mentioned. If coalition is allowed, we have the following cases:

| \( \delta + c > \beta \) | Feasible Maximum Report |
| \( \delta + c = \beta \) | Truthful Report |
\[ \delta \beta + < \text{Feasible Minimum Report} \]

\( \delta + c < \beta \)

With \( \delta + c > \beta \), physician gains money per unit of treatment. The side-contract can be established to diversify profit between physician and patient, then both of them accept the overstate \( \tau^r \), vice versa. Since physician’s net reimbursement is prospective payment, proposition 1 applies, which indicates the sign of margin \( \delta \) for complement and substitute cases. The requirement for truthful reporting includes not only margin for both cases, but also \( \delta + c = \beta \). This new constraint for distort the implementation of costly effort in optimal risk sharing. Although no mathematics are shown in this paper, we can predict that the second best solution might be infeasible with truthful constraint for coalition between physician and patient.

**Physician Competition**

**Overview**

Assuming that still both \( \varepsilon \) and \( \tau \) are not verifiable and thus not contractible, the authors proceed with an analysis of physician competition with respect to the quantity of effort.

Given the insurance payment parameters and the physician’s effort, patients will seek help and decide on their quantity of treatment. Anticipating these reactions, physicians will choose their profit maximizing level of effort, which determines demand and the quantity of treatment of each of the patients.

**Formal Analysis**

The analysis in this part is extended to a population of consumers with a mass normalized to one. Each consumer has the same utility function but a different, uninsurable out-of-pocket expense for different physicians (in addition to the copayment \( \beta \)). \( \theta_i \) denotes the out-of-pocket cost for obtaining service from physician \( i \), and \( \Phi_i(x) \) denotes the proportion of consumers who must incur a cost of at most \( x \) when they use physician \( i \).

Assuming that a patient gets a reservation utility of \( \bar{U} \) if not using physician \( i \), the condition for the optimal choice of treatment for a given set of parameters (insurance payment parameters and the physician’s effort level), the optimal choice of treatment for the patient is given by (4) and ex-post (optimized w.r.t. \( \tau \)) increasing in the level of effort. Intuitively, the patients still set, according to (4), the marginal cost of treatment equal to the marginal benefit, but the optimal choice of treatment is the higher the higher the physician’s effort is.
The service of physician $i$ will be demanded by those patients with values of $\theta_i$ below the threshold level (denoted by $\bar{\theta}_i$) of

$$U(w - \alpha - \beta \tau - s + F(\tau, \epsilon) - \bar{\theta}_i) = \bar{U}.$$  

Solving (11) the authors express the threshold as an increasing function of $\epsilon$: $\bar{\theta}_i(\epsilon)$. It follows that the higher the physician sets the level of effort, the more consumers, when sick, will demand his service. This share $\Phi_i(\bar{\theta}_i(\epsilon))$ will be denoted by $D(\epsilon)$. Thus it represents the demand for physician $i$.

Hence, the physician’s profit can be written as

$$D(\epsilon)[\rho + \delta \tau - G(\epsilon)],$$

where $\tau$, the optimal quantity, is again given by (4).

Considering the first-order derivative of (12) w.r.t. $\epsilon$ one gets:

$$D'(\epsilon)[\rho + \delta \tau - G(\epsilon)] + D(\epsilon)[\delta \frac{d \tau}{d\epsilon} - G'(\epsilon)].$$

(13) reveals that competition provides an extra incentive for costly effort. When comparing it to (7) it becomes obvious that the marginal benefit from increasing costly effort in the regime without competition is lower than in the regime with competition. With competition, the physician by increasing her effort does not only increase the treatment demanded per patient (second term of (13)), but additionally increases her profit by adding the marginal return through increasing demand (first term of the equation), if she expects to earn a positive profit from treating each patient (i.e. $\rho + \delta \tau - G(\epsilon) > 0$).

**Competition with a fixed level of effort, $\epsilon^*$**

As in section IV with the ethics constraint, a fixed level of effort can be identified with an appropriately defined second best regime.

Assuming that effort and treatment are complements, the results from section III reveal that the fixed level of effort, $\epsilon^*$, can be obtained in equilibrium by accordingly setting the payment parameter $\delta$. Thus, competition does not improve the results if effort and treatment are complements.

When effort and treatment are substitutes, however, the results can be different if without competition $\epsilon^*$ lies outside the implementable set $\Omega$. Following the lines of argumentation in sections III and IV, competition and an appropriate choice of $\delta$ and $\rho$ can improve the outcome and implement $\epsilon^*$. 

The reason for the inability to implement $\varepsilon^*$ without competition is that the truth telling constraint sets the lower limit of $\delta$ at $-c$. A lower $\delta$, however, decreases the physician’s reward for visits and induces him to increase her level of effort.

Competition thus provides an additional incentive to provide effort. This effect can be exploited by insurers to implement $\varepsilon^*$ by increasing the value of $\rho$, the prospective payment per patient, which increases the value of each patient attracted (the first part of equation (13)), thus increasing the incentive to exert higher effort.

Assuming that in equilibrium $\delta = -c$, (as effort and quantity of treatment are substitutes), and substituting $\delta$ in (13) by $-c$ one can formally pick $\rho^*$ to set (13) equal to zero, such that $\varepsilon^*$ will be implemented.

As physicians will under such regime ex-post earn a significant amount of profit, the insurer might want to set a participation fee that extracts the rent and allows physicians to compete with their colleagues for patients.

Section VI: Discussion

Unlike previous literature the authors do not only target the ex-ante noncontractibility of treatment but introduce a second the market failure in terms of the noncontractibility of effort. Thus they deal with the complex interplay of physicians’ and patients’ demand response and later on add an reporting subgame which’s truthful requirement poses restrictions on the equilibrium outcomes.

Furthermore, by incorporating ethics constraints and competition among physicians into their model they suggest means to alleviate the restrictions imposed by the truthful reporting requirement.