UNIVERSITÄT BERN

Faculty of Economics and Social Sciences

Department of Economics

Incentive Compatible Reimbursement Schemes for Physicians

Winand Emons

10-01

January 2010

DISCUSSION PAPERS

Schanzeneckstrasse 1
Postfach 8573
CH-3001 Bern, Switzerland
http://www.vwi.unibe.ch

Incentive Compatible Reimbursement Schemes for Physicians

Winand Emons*
University of Bern and CEPR
January 2010

Abstract

We consider physicians with fixed capacity levels. If a physician's capacity exceeds demand, she may have an incentive to overtreat, i.e., she may provide unnecessary treatments to use up idle capacity. By contrast, with excess demand she may undertreat, i.e., she may not provide necessary treatments since other activities are financially more attractive. We first show that simple fee-for-service reimbursement schemes do not provide proper incentives. If insurers use, however, fee-for-service schemes with quantity restrictions, they solve the fraudulent physician problem.

Keywords: credence goods, expert services, incentives, medical doctors, demand inducement, insurance.
Journal of Economic Literature Classification Numbers: D82, I11.

*Departement Volkswirtschaftslehre, Universität Bern, Schanzeneckstrasse 1, Postfach 8573, CH-3001 Bern, Switzerland, Phone: +41-31-6313922, winand.emons@vwi.unibe.ch, www.staff.unibe.ch/emons/. I thank Uwe Dulleck, Rudi Kerschbamer, Fridolin Marty, and Pius Matter for helpful comments.

1. Introduction

The US spends between one fifth and one third of its health care expenditures, that is between 500 and 700 billion dollars, on care that doesn't improve anybody's health. These unnecessary tests and treatments aren't just expensive, they can also harm patients.[1]

One factor contributing to this enormous waste is that medical services constitute credence goods: a physician not only provides the medical services; at the same time she also acts as the expert who determines how much treatment is necessary because her patient is unfamiliar with the medical condition.

Aggravating this special feature is the fact that even ex post the patient can hardly determine the extent of the treatment that was required ex ante. It is often difficult, if not impossible, to find out whether treatments were really needed or whether necessary treatments were not provided. Since from ex post observations the patient can never be certain of the quality of the treatments he obtained, such services have been termed credence goods (Darby and Karni (1973)).

This information asymmetry creates obvious incentives for opportunistic physician behavior. On the one hand, if there is plenty of money in treatments, physicians may recommend unnecessary treatments. On the other hand, doctors may not perform urgently needed treatments if other activities are more profitable.[2]

To give a few examples. In the Swiss Canton of Ticino the population average had 33% more of the seven most important operations than medical doctors and their families. Interestingly enough, lawyers and their beloved have about the same operation frequency as the families of medical doctors (Domenighetti et al. (1993)). Gruber et al. (1999) show that the frequency of cesarian deliveries compared to vaginal deliveries positively reacts to fee

[1] See, e.g., Brownlee (2007, p. 5) or The Economist 02/13/1999.

[2] Brownlee (2007, p. 8) mentions a couple of other reasons for overtreatment: doctors simply don't know which treatments are most effective, they want to help patients even when they don't know the right thing to do, malpractice fears drive defensive medicine, medical custom varies from region to region, one doctor often doesn't know that another physician has already ordered a battery of tests, and patients, being insured, ask for fancy treatments (demand-induced supply). Yet, according to her view, the most powerful reason for overtreatment is that doctors and hospitals get paid more for doing more.

differentials of health insurance programs. Marty (1998) shows, using 8000 bills of Swiss general practitioners, that doctors with sufficient demand charge significantly less per patient than doctors with excess capacity. Primary care physicians are squeezed financially so that their numbers dwindle; at the same time the number of the highly profitable specialists continues to rise, leading Brownlee (2007, p. 265) to sigh: "...sometimes what we really need is not a doctor who delivers more care but one who seems to care more..."

In this paper we analyze whether health insurers can design reimbursement schemes so that physicians have no incentives to behave fraudulently. We first show that simple fee-for-service reimbursement schemes do not provide proper incentives. If insurers use, however, fee-for-service schemes with quantity restrictions, they solve the fraudulent physician problem.

As a workhorse we use the basic model of Emons (1997, 2001). Patients are up for a diagnosis. Some patients are in good condition and need no further treatment; the rest is in bad condition and needs treatment. After the diagnosis the physician knows which condition the patient is in. She can then treat him. The physician can only perform the treatment after a diagnosis. We thus have economies of scope between diagnosis and treatment, making the separation of diagnosis and treatment inefficient.[3]

We consider a set of physicians, each of whom has a fixed capacity: a physician may have to ration her patients due to insufficient capacity, or she may also end up with idle capacity. If a physician has excess demand, she may undertreat patients, i.e., she may not provide necessary treatments if diagnosis is financially more attractive than treatment. By contrast, with excess capacity the physician may start to overtreat, i.e., provide unnecessary treatments to use up idle capacity.

Insurers set reimbursement terms. To focus on the incentive effects of the reimbursement terms, we consider the following simple mechanism to allocate patients to physicians. Nature assigns patients randomly to the physicians. Doctors decide how many of their patients they want to diagnose; patients who obtain no diagnosis are referred to a second round. Having diagnosed her patients, a doctor then decides whom to treat.

In the second round nature allocates those patients who obtained no

[3]This separation mechanism is often encountered in the prescription and preparation of drugs: the physician prescribes the drugs and the pharmacist may only sell only what has been prescribed by the doctor.

service so far to the physicians who still have spare capacity, and so on, until either all patients eventually found a physician or until physicians have no more capacity left. The physicians' aggregate capacity is just sufficient to treat all patients non-fraudulently. Therefore, the referral process ensures that all patients are serviced if doctors behave honestly.

Physicians are myopic profit maximizers. When they decide about the first round of patients, they do not anticipate that in a later round they might get referrals from colleagues.

We first analyze simple fee-for-service reimbursement schemes: the physician is paid per diagnosis and per treatment she performs. We show that there exists no fee-for-service scheme under which all patients get non-fraudulent services. Consider, for example, equal compensation prices equalizing the profit per diagnosis with the profit per treatment. With these prices all doctors with excess demand are indifferent between diagnosis and treatment and, accordingly, provide honest services. Yet doctors with excess capacity overtreat to use up their idle capacity. If, by contrast, the treatment price is zero, the incentive to overtreat disappears and doctors with excess capacity behave non-fraudulently. But now doctors with execss demand undertreat because diagnosis is much more attractive than treatment. It is thus impossible to find fee-for-service schemes that give both, physicians with excess demand and physicians with excess capacity proper incentives at the same time.

In the next step we use the fact that insurers have more information than the individual patient. Whereas the patient has only one observation of the physician's behavior, the insurance company has the set of observations for its entire clientele. In particular, the insurer knows how many of its policy holders actually underwent treatment. In addition to the fees-for-services the insurer can thus use a quota that states the maximum fraction of diagnosed patients for which the insurer pays the treatment.

Obviously, this quota needs to be equal to the fraction of patients actually in need of treatment. If the quota is lower, it enforces undertreatment; if it is higher, it opens the door for overtreatment. It turns out that a quota equal to the fraction of patients in need of treatment curbs overtreatment. If a doctor wishes to overtreat to use up idle capacity, she is not reimbursed for these treatments. We are thus only left with the problem of undertreatment if a doctor has excess demand. This problem is solved by prices making

diagnosis not more attractive than treatment. With these prices a doctor prefers providing necessary treatment to diagnosing another patient. Since physicians with excess demand and physicians with excess capacity have proper incentives, the referral process eventually ensures that all patients get non-fraudulent services.

The literature on credence goods as surveyed by Dulleck and Kerschbamer (2006) looks at one-shot relationships between the expert and her customer. The customer has only one observation of the expert's actions. This information together with the outcome of his case does not allow the customer to draw perfect inferences about the appropriateness of the treatment he has received.[4] Most of this literature considers experts operating in a market environment. The only model we are aware of incorporating insurance in a credence good set-up is Sülzle and Wambach (2005). They take prices as given and analyze the impact of coinsurance on the physician's incentives to cheat and on the patients' incentive to search for a second opinion. They do not attempt to find contracts inducing non-fraudulent behavior.

In Ely and Välimäki (2003) short-lived motorists play a repeated game with long-lived mechanics. Good mechanics prefer to act truthfully while bad mechanics prefer to always change the engine. Each motorist observes the repairs performed for preceding customers but has no idea whether these repairs were appropriate.[5] Good mechanics may not do necessary engine replacements early on in the game to separate themselves from the bad mechanics and signal their good type to future motorists. Motorists anticipate this incentive to undertreat to build up a good reputation and may not visit the mechanic in the first place. Similar to us, Ely and Välimäki use the information of the expert's treatment history. In Ely and Välimäki prices are exogenously given. By contrast, we also determine reimbursement prices such that, together with the quota, experts have proper incentives and the outcome is efficient.

[4]Typically, this literature assumes the undertreatment problem via inference away and deals only with the overtreatment issue; see our discussion below.

[5]Ely and Välimäki assume that a motorist finds out ex post whether or not he received the appropriate service. Strictly speaking, they do not analyze a credence good but a horizontally differentiated experience good. Yet, the motorist takes the information about the appropriateness of the repair with him to his grave. Thus, the following motorists know which repair he got but do not know whether it was appropriate.

The rest of the paper is organized as follows. The next section introduces the basic model. In section three we look at fee-for-service reimbursement schemes. In the next section we extend fee-for-services with quantity rationing. Section 5 concludes.

2. The Model

An agent is up for a diagnosis by a physician. During the period to come the individual may fall ill or he may stay healthy. If the agent stays healthy, he receives a monetary utility of 1; if he becomes sick, the utility is 0.

At the time under consideration the agent may be in good or bad condition. If the patient is in good condition, the probability of staying healthy is $q_h \in (0,1)$; if the patient is in bad condition, the probability of staying healthy is $q_\ell \in (0, q_h)$, i.e., lower than when the consumer is in good condition. Let $p \in (0,1)$ be the probability that the patient is in bad condition. The patient does not know in which of the two conditions he is in, nor can he infer it ex post since he may fall ill or stay healthy under both conditions.

The patient visits one of n medical doctors, indexed by $i = 1, \ldots, n$; in what follows we will suppress the index i wherever possible. By diagnosing the agent, the physician detects his true condition. When the patient is in good condition, he needs no further treatment. When the consumer is in bad condition, the doctor can treat him; after the treatment the consumer is in good condition. Let $d > 0$ be the total resource cost of diagnosing and $r > 0$ the resource cost of treating a patient.[6]

The timing of the production decisions, however, is such that these costs are not experienced as genuine marginal costs. The physician has L units of time (say, hours) available. She allocates her L units of time to diagnosis and treatment; d is the time the doctor needs per diagnosis and r the time per treatment. The physician's time cost is sunk.

The physician's reservation wage is normalized to 1. Accordingly, L is the sunk cost of being active; d and r measure the minimum average costs of diagnosis and treatment if, say, the doctor performs either activity exclusively. Note that marginal costs are different from average costs. A doctor

[6]Our diagnosis corresponds to Dulleck and Kerschbamer's (2006) cheap treatment; their expensive treatment corresponds to our "diagnosis cum treatment". The information structure in De Jaegher's (2009) prevention scenario is similar to ours except that he has an additional moral hazard problem in treatment.

has a fixed capacity the cost of which is sunk. Therefore, her marginal costs are 0 except for the capacity margin where marginal costs are "$+\infty$". When, in the following, we talk about minimum average costs we mean d and r.

There is a continuum of identical consumers with total mass 1.[7] In units of time a capacity of $d + pr$ is necessary to serve the entire market non-fraudulently. We assume $L_i < d + pr,\ i = 1, \ldots, n$; each physician does not have the capacity to serve the entire market honestly. Define $\lambda_i = L_i/(d+pr)$ as the doctor's capacity in terms of customers given non-fraudulent behavior. Let $\sum_{i=1}^{n} \lambda_i = 1$; altogether the capacity in the market is just sufficient to serve everybody honestly. This assumption implies first of all that if a physician has excess demand, referrals to a colleague are a efficient given non-fraudulent behavior. Moreover, it allows us to easily determine prices allowing doctors to recover just their sunk cost L with non-fraudulent services.[8]

Consumers are risk neutral and care only about monetary flows. Accordingly, given that we have normalized the utility of staying healthy to 1 monetary unit, without diagnosis and treatment a consumer's expected utility is $\bar{U} = (1-p)q_h + pq_\ell$. With (honest) diagnosis and treatment priced at minimum average costs the consumer's expected utility amounts to $q_h - d - pr$. The consumer incurs the cost of diagnosis in any case. With probability p the consumer is in bad condition and needs treatment. In return, the consumer is in good condition for sure.

It is efficient to diagnose the consumer and treat him if necessary, meaning $q_h - d - pr > \bar{U}$ or $p(q_h - q_\ell) > d + pr$. Treating a consumer in bad condition increases his utility by $(q_h - q_\ell)$. With probability p the consumer is in bad condition. Accordingly, the expected benefit from diagnosing and treating

[7] We make the continuum assumption not only for notational convenience. With a finite number of consumers we run into the following problem. Suppose the physician expects a clientele with $(1-p)$ patients in good and p patients in bad condition. With a finite number of customers, however, the actual realization of her clientele will typically be different from the expected one. Accordingly, at the end of the day she will realize that she has either insufficient or excess capacity and she will start behaving fraudulently (suggesting that it is better to see a doctor in the morning rather than late afternoon). With a continuum of patients we do not encounter this difficulty. If L measures, say, capacity per year, finiteness is less of a problem than if L is the capacity per day because the number of patients is larger.

[8] Referrals are also efficient if $\sum_{i=1}^{n} \lambda_i > 1$; the determination of zero-profit prices is, however, more cumbersome.

the consumer is $p(q_h - q_\ell)$. The surplus the physician's services may generate is, therefore, $p(q_h - q_\ell) - (d + pr)$.

Let us now describe how a doctor may defraud her patients. After diagnosis the physician knows which condition the patient is in. When the patient is in bad condition, she can treat him, i.e., turn him into good condition. Yet she can also 'treat' a patient in good condition; in this case the physician unnecessarily spend r units of time on the patient — leaving him at least in good condition. This kind of behavior has been termed overtreatment (Dulleck and Kerschbamer (2006)) or supplier-induced demand in health economics (Labelle et al. (1994)).

Alternatively, when the patient is in good condition, the medical doctor can recommend no treatment. Nevertheless, she can make the same recommendation when the patient is in bad condition. We will refer to this type of fraud as undertreatment (Dulleck and Kerschbamer (2006)).[9]

Ex post the patient has no way of finding out whether he was treated unnecessarily or whether he needed treatment that was not provided. The physician's services thus constitute 'credence' goods as distinct from search and experience goods — from ex post observations the consumer can never be certain of the quality of the services he has purchased. See Darby and Karni (1973).

Note that we assume diagnosis and treatment to be verifiable. This assumption is appropriate for physicians whose patients necessarily take part in any (un-)necessary treatment. It is not appropriate for, e.g., a customer who sends his gadget to a service center. When the widget is returned the customer is unable to tell whether somebody in the repair center has actually worked on the gadget. Here the expert has yet another possibility to defraud her customers. She can claim to have fixed the widget without having touched it, thus collecting repair fees from an unlimited number of customers.[10]

[9]Most of the credence goods literature assumes the undertreatment problem away by setting $q_h = 1$. Under this assumption a patient knows for sure that he didn't get the necessary treatment when he falls ill. Moreover, the patient's health status is verifiable and a legal rule holds the physician liable if the patient becomes sick; see, e.g., Dulleck and Kerschbamer (2006).

[10]See, e.g., Emons (2001) or Dulleck and Kerschbamer (2006) for set-ups where the expert's actions are not verifiable.

All patients have full insurance from one of m insurance companies, indexed by $j = 1, \ldots, m$. Insurers reimburse the physicians. The sequence of the events is as follows. Insurance companies choose identical reimbursement terms. In the first round all n doctors announce that they have free capacity. Then patients decide which doctor to see. We model this allocation decision as a move by nature determining the fraction of patients η_i visiting physician i, $i = 1, \ldots, n$ with $\sum_{i=1}^{n} \eta_i = 1$. Formally, nature chooses from a continuous density over the $(n-1)$-dimensional simplex. Therefore, $(\eta_1, \ldots, \eta_n) \neq (\lambda_1, \ldots, \lambda_n)$ with probability 1, meaning some doctors have excess demand while others have excess capacity. Physician i then decides how many patients $\mu_i \leq \eta_i$ she diagnoses; $\eta_i - \mu_i$ patients are referred to the second round. After having diagnosed her μ_i patients, doctor i then decides whom to treat.

In the second round all doctors with free capacity announce this fact; physicians who have used up their capacity in the first round drop out. Nature then allocates the patients who obtained no service so far to those doctors with free capacities like in the first round and so on. The process is over when either all patients eventually found a physician or if physicians have no more capacity left. If a patient doesn't find a doctor willing to take him, he ends up with his reservation utility of $\bar{U}$.

Medical doctors maximize profits; since all costs are sunk in our framework, profit maximization boils down to revenue maximization. Moreover, we assume physicians to be myopic. They maximize profits for each round of patients. Specifically, when they decide about the first round of patients, they do not anticipate that in a later round they might get some referrals from colleagues. Doctors thus go shortsightedly go for the quick buck. There is no discounting between rounds.[11] Insurers try to find reimbursement terms that induce non-fraudulent behavior and generate zero-profits for physicians

[11] Alternatively, we could assume that physicians discount profits so that they prefer profits in the first round to profits in the second round. Then a doctor with idle capacity in the first round compares the profit from overtreatment now with the expected profit she can make if she carries over the capacity into the next round. If the discount factor is sufficiently small, she will also go for the quick buck. With discounting, the determination of prices yielding zero profits becomes, however, more cumbersome; moreover, we need to determine expected profits for later rounds. Finally, we consider the assumption that physicians fully anticipate the workings of the market somewhat far-fetched.

with honest services.[12] We will now look at different reimbursement schemes.

3. Fee-for-service

Under a simple fee-for-service reimbursement scheme the physician gets D per performed diagnosis and R per performed treatment.

Recall that a doctor has a capacity of L units of time having a sunk cost L. In terms of patients the physician has capacity $\lambda < 1$ given honest behavior. Apparently, her behavior depends on the size of her clientele η relative to her capacity λ. According to whether $\eta \gtreqless \lambda$ we will say that the physician has too many/enough/not enough patients given non-fraudulent behavior. If, say, the doctor does not have enough patients, she may start 'treating' patients in good condition to utilize her otherwise idle capacities. If she has too many patients, she may, e.g., be tempted not to treat all patients in bad condition given that diagnosis is more profitable than treatment.

The last example indicates that the physician's incentives also depend on the relative profitability of diagnosis to treatment which, in turn, is determined by the prices D and R. If the doctor has too many patients, the only constraint she faces (at the margin) is her precious time. To maximize profits, she compares the profit per hour treatment $(R-r)/r$ with the profit per hour diagnosis $(D-d)/d$. If the former exceeds the latter she will overtreat whereas she will undertreat if diagnosis is more profitable than treatment. We specify these ideas more precisely in the following Lemma.

Lemma 1:
i) If $\eta > \lambda$, the physician is honest if and only if $R = rD/d$;
ii) if $\eta = \lambda$, the doctor is honest if and only if $R \leq rD/d$;
iii) if $\eta < \lambda$, the doctor is honest if and only if $R = 0$.

Proof: i) If $\eta > \lambda$, the doctor has more patients than she can handle with honest behavior. Given her time constraint, she is only interested in the profit per hour treatment $(R-r)/r$ compared to the profit per hour diagnosis

[12] Non-fraudulent behavior maximizes total surplus. Zero profits ensure that physicians are active. The insurers' objective thus coincides with the one of a social planner who maximizes social welfare subject to the constraint that physicians cannot be forced to practice.

$(D-d)/d$. If $R = rD/d$, which implies $(R-r)/r = (D-d)/d$, she is indifferent between diagnosis and treatment and, therefore, behaves honestly. If $R > rD/d$, she prefers treatment to diagnosis and thus overtreats and undertreats if $R < rD_i/d$.

ii) If $\eta = \lambda$, the physician fully utilizes her capacity with non-fraudulent behavior. If $R < rD/d$, she strictly prefers diagnosis to treatment; yet she makes diagnoses for her entire clientele. She has to perform treatments to use up her remaining time $L - \eta d$; honestly treating the patients in bad condition of her clientele just exhausts her capacity. If $R = rD/d$, the argument is along similar lines as i). If $R > rD/d$, the physician strongly prefers treatment to diagnosis. Hence, she will treat all patients she diagnoses.

iii) If $\eta < \lambda$, the doctor has idle capacity with honest behavior. As long as $R > 0$, she makes money by treating some more patients to use her idle capacity. Only when $R = 0$ the incentive for overtreatment disappears. ■

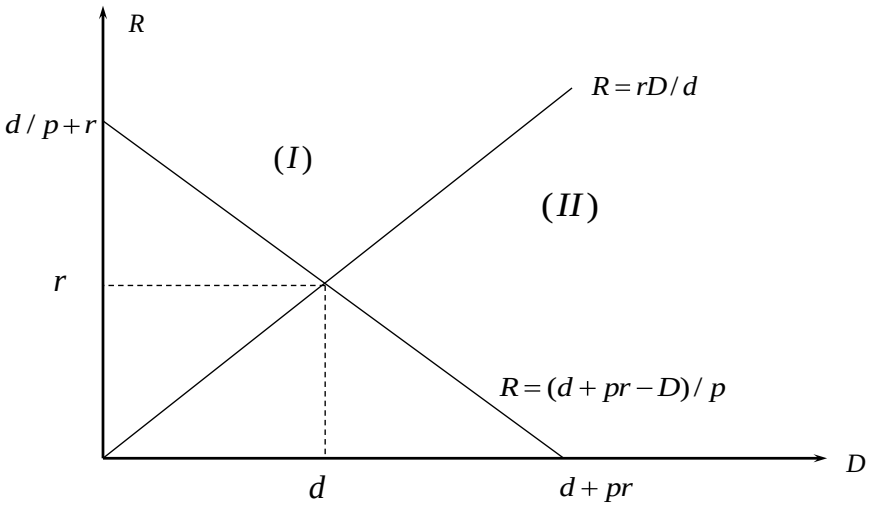

Figure 1: The equal compensation and the zero-profit lines

The message of Lemma 1 can be seen in Figure 1. Consider the line $R = rD/d$ along which $(R-r)/r = (D-d)/d$. Accordingly, on this equal compensation line the physician is indifferent between diagnosis and treatment so that with too many patients she opts for efficient treatment: She diagnoses $\mu = \lambda$ patients and treats the fraction p thereof; she refers the remaining $(\eta - \lambda)$ patients to the second round.

In region (I) where $R > rD/d$ the doctor prefers treatment to diagnosis. Whatever the number of patients, she will 'treat' everybody she diagnoses, i.e., she will overtreat. In region (II) in which $R < rD/d$ the physician prefers diagnosis to treatment so that she wishes to increase the number of diagnoses at the expense of treatments. If the physician has too many patients, we will observe undertreatment. With enough patients, however, she cannot diagnose more patients; she treats efficiently to make some money out of her otherwise unused capacity.

When the physician does not have enough patients, she will treat everybody as long as $R > 0$. Only when $R = 0$ the physician has proper incentives if she does not have enough patients. She does not overtreat to utilize her idle capacity because there is no money in treatment.

Lemma 1 has the following negative implication:

Proposition 1: *If insurers use simple fee-for-service reimbursement schemes (D, R), there exists no set of prices under which all patients get non-fraudulent services.*

Proof: If $R > rD/d$, all doctors overtreat regardless of their demand. $(d + pr)/(d+r)$ patients get a diagnosis and a treatment; $(1-p)$ of these treatment are unnecessary. $(1-(d+pr)/(d+r))$ patients end up with no medical services at all.

If $R = rD/d$, all physicians with enough and too many customers have proper incentives and refer efficiently. Yet those doctors with excess capacity overtreat in each round. They don't have the capacity in later rounds to serve the referrals of their colleagues non-fraudulently. Thus, some patients are overtreated which implies that some patients end up with no service at all.

If $0 < R < rD/d$, physicians with excess demand undertreat, i.e., some of their patients don't get the necessary treatment. By contrast, doctors with excess capacity overtreat. Whether or not some patients end up with no service depends on the demand realization in each round.

If $R = 0$, physicians with excess capacity have correct incentives and refer efficiently. Yet, physicians with excess demand undertreat. All patients are diagnosed, yet some patients are denied the necessary treatment. ■

In Figure 1 we have also depicted the line $R = (d + pr - D)/p$. All

prices along this line generate zero-profits with λ_i patients and non-fraudulent behavior. Suppose, for example, the insurance companies reimburse equal compensation prices (d, r). Then the physicians with enough or too many customer have proper incentives and refer efficiently; furthermore, they make zero profits. Yet, those physicians with excess capacity will overtreat. Therefore, after the first round the market has no longer the capacity to deal with the referred patients non-fraudulently.

If insurers use, say, the fully capitated reimbursement $(d + pr, 0)$, physicians have proper incentives with enough or too few patients.[13] Yet, those doctors with excess demand will undertreat. Accordingly, in the second round there is excess capacity in the market.

Note that this negative result is driven by the physician's fixed capacity. In a set-up where experts only incur variable costs, equal compensation prices always induce honest behavior; see, e.g., Dulleck and Kerschbamer (2006).[14]

4. Fee-for-service with Quantity Restrictions

Let us now use the fact that an insurance company has more information than an individual patient. Whereas the patient has only one observation of the physician's behavior, the insurance company has the set of observations for its entire clientele. In particular, the insurer knows how many of its policy holders actually underwent treatment. To be more specific, let $\eta_{ij} > 0$ be the fraction of physician's i's patients having insurance from firm j with $\sum_{j=1}^{m} \eta_{ij} = \eta_i$.

Insurers offer reimbursement schemes (D, R, z) where z denotes the maximum fraction of diagnosed patients for whom the insurer actually pays the treatment. It turns out that the quota z is a powerful instrument to curb overtreatment.

First note that to implement non-fraudulent behavior we need $z = p$. If $z < p$, the insurer enforces undertreatment. If, by contrast, $z > p$, we run into the problems as described by Lemma 1.

[13] With enough patients the physician makes zero profits; with excess capacity, however, she makes losses.

[14] In Dulleck and Kerschbamer (2006) doctors have no capacity constraint so that equal compensation (equal markup) prices satisfy $R - r = D - d$. Another set-up with capacity constrained experts can be found in Richardson (1999).

Lemma 2: *Let $z = p$.*

i) If $\eta > \lambda$, the physician is honest if and only if $R \geq rD/d$;

ii) if $\eta \leq \lambda$, the doctor is honest for all prices (D, R).

Proof: i) If $\eta > \lambda$, the physician has more patients than she can handle with honest behavior. If $R < rD/d$, she prefers diagnosis to treatment. She diagnoses all η patients and uses her remaining capacity (if any) to treat a few patients. We have thus undertreatment.

If $R = rD/d$, the physician is indifferent between diagnosis and treatment and, therefore, honestly deals with λ patients.

If $R > rD/d$, the physician prefers treatment to diagnosis. She would like to treat all patients she diagnoses. Yet she can bill treatments only for the fraction p of the patients she diagnoses. To use up her capacity, she diagnoses λ patients and treats the fraction p thereof being in bad condition.

ii) If $\eta = \lambda$, the physician fully uses her capacity with non-fraudulent behavior. If $R = rD/d$ she has proper incentives and uses up her capacity by honestly serving all patients. If $R < rD/d$, she prefers diagnosis to treatment. She diagnoses all patients; to use up her remaining time $L - \eta d$ she has to treat. Honestly treating the patients in bad condition just exhausts her capacity. If $R > rD/d$, the doctor prefers treatment to diagnosis. She would like to treat all patients but is curbed by the quota p. Hence, she behaves honestly.

If $\eta < \lambda$, the physician has unused capacity with non-fraudulent behavior. As long as $D > 0$, she will diagnose all η patients. As long as $R > 0$, she would like to treat more than $p\eta$ patients to use her idle capacity. Yet she cannot bill more than $p\eta$ patients for treatment. ■

It is perhaps somewhat surprising that the quota $z = p$ induces honest behavior for all prices if the physician has enough or not enough demand. With excess capacity the physician diagnoses all patients. As long as $R > 0$, she would like to overtreat to use her idle capacity. Yet the reimbursement quota prevents her from doing so. By contrast, if the physician has excess demand, diagnosis may not be more attractive than treatment. If diagnosis is relatively more profitable than treatment, the physician will diagnose all patients she can get hold of and treat less than the fraction p thereof, i.e., we have undertreatment.

Lemma 2 implies the following result:

Proposition 2: *Under the reimbursement schemes* $(D, (d + pr - D)/d, p)$ *with* $D \leq d$ *all patients get non-fraudulent service and all physicians make zero-profits.*

Proof: Lemma 2 implies that for reimbursement schemes $(D, (d+pr-D)/d, p)$ with $D \leq d$ physicians have proper incentives whatever their demand. Hence doctors with excess demand service λ_i patients honestly and refer $\eta_i - \lambda_i$ patients to the second round. Physicians with excess capacity treat η_i patients honestly and enter the second round with capacity $\lambda_i - \eta_i$. Since $\sum_i \lambda_i = 1$, $\sum_{\{i|\eta_i > \lambda_i\}}(\eta_i - \lambda_i) = \sum_{\{i|\eta_i < \lambda_i\}}(\lambda_i - \eta_i)$; the remaining physicians have the capacity to handle the remaining patients honestly, and so on for further rounds.

In each round at least one doctor has excess demand, uses up her entire capacity, and is no longer active in the following round. The referral process thus comes to an end after a finite number of rounds and each doctor has used up her capacity with non-fraudulent services. In the last round only one physician remains whose capacity is just sufficient to serve the remaining patients honestly. Finally, note that the prices $(D, (d+pr-D)/d)$ with $D \leq d$ together with honest behavior and demand λ_i yield revenue L so that doctors end up with zero profits. ■

There exist thus reimbursement schemes (D, R, z) inducing non-fraudulent behavior for all realizations of demand. Any physician with excess demand will refer the patients she cannot deal with honestly to the next round. Any physician with excess capacity behaves non-fraudulently, meaning that she has capacity left to treat the referrals from colleagues. Since we have assumed that $\sum_i \lambda_i = 1$, at the end of referral process each physician has enough customers. Thus, incentive compatible prices on the line $R = (d + pr - D)/p$ together with the quota $z = p$ indeed yield zero profits. Note that for our scheme to work, an insurer need *not* know the physician's capacity level.[15]

[15] Assessing a physician's capacity is a tricky task. For example, in Switzerland a lot of -in particular female- physicians prefer to work part- rather than full-time, making her capacity level her private information. Any reimbursement scheme that builds on a physician's capacity level, therefore, has to deal with the issue how this information is revealed.

A few qualifying remarks are in order. We have assumed that the number of patients is a continuum. Each physician serves a fraction of the market. Therefore, a doctor's clientele is also a continuum. We assume that a continuum of independent and identically distributed random variables sum to a non-random variable.[16] To be more specific, a physician has continuum of patients, the fraction p of which is in need of treatment; see also the discussion in footnote 7. The reimbursement quota $z = p$, therefore, coincides with the actual number of patients in need of treatment.

With a finite number of patients, the actual number of patients in need of treatment will typically be different from the expected value. If the actual number exceeds z, the quota enforces undertreatment so that some patients do not get necessary treatment; if it is lower, the quota opens the door to overtreatment. With a finite number of patients the optimal quota minimizes the expected costs from under- and overtreatment. By working with the continuum we have avoided this technical difficulty. Nevertheless, the larger the number of patients a doctor handles, the less important this problem becomes.

Another difficulty arises if patients are not identical as in our setup. Suppose the probability of being in need of the treatment is distributed in the population on $[0, 1]$ with mean p, the density having full support. As long as each physician gets a random sample of the population as patients, our results continue to hold. If, however, there is a selection bias such that some physicians get on average less healthy patients than others, our one-size-fits-all quota no longer gives proper incentives for all physicians. The quota then has to be adjusted to the group of patients seeing the doctor.

We have assumed that only one treatment is available and thus that the fraction of patients in need of treatment is well defined. Often there are, however, professional disagreements covering the diagnosis and treatment of illness. For example, Wennberg et al. (1982) show that the wide range of acceptable diagnoses and therapies are a major factor in the wide variation in rates of of utilization and costs of medical services among neighboring medical markets. Our analysis, therefore, applies to diseases where there are no professional disagreements, or to cases where insurers enforce the most effective way of dealing with the illness.

[16] See Judd (1985) for a discussion of this assumption.

Despite these shortcomings of our simple model, we think that treatment quotas are a useful instrument for insurers to curb overtreatment incentives. As to our knowledge, insurers tend to make little use of this instrument. For example, in Switzerland insurers start an investigation if a physician's actual billing per patient is 30% higher than the average for this group of doctors.[17] Given our results, a more sophisticated use of treatment records seems warranted.

5. Conclusions

The purpose of this paper is to develop incentive compatible reimbursement schemes for physicians. We have chosen a framework where due to the physicians' fixed capacity levels both, the problem of under- and of overtreatment arise. Simple fee-for-service schemes do not solve the incentive problems. Either physicians with excess capacity or physicians with excess demand have the wrong incentives.

We then use the fact that insurers observe a physician's actions for the entire set of their policy holders. This allows insurers to set a quota which states the maximum fraction of diagnosed patients for whom insurers actually pay the treatment. If insurers set this quota equal to the fraction of patients in need of treatment, they curb overtreatment. Therefore, only the undertreatment problem remains which is solved by prices making diagnosis not more attractive than treatment.

[17] For more on this so called ANOVA-method see, e.g., Roth and Stahel (2005).

References

Brownlee, S., Overtreated: Why too much Medicine is Making us Sicker and Poorer, New York: Bloomsbury (2007).

Darby, M. R. and E. Karni, Free Competition and the Optimal Amount of Fraud, *Journal of Law and Economics* 16 (1973), 67-88.

De Jaegher, K., Physician Incentives: Cure versus Prevention, *Journal of Health Economics* forthcoming (2010).

Domenighetti, G., Casabianca, A., Gutzwiller, F., and S. Martinoli, Revisiting the most Informed Consumer of Surgical Services, *International Journal of Technology Assessment in Health Care* 9 (1993), 505-513.

Dulleck, U. and R. Kerschbamer, On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods, *Journal of Economic Literature* 44 (2006), 5-42.

Ely, J. and J. Välimäki, Bad Reputation, *Quarterly Journal of Economics* 118 (2003), 785-814.

Emons, W., Credence Goods and Fraudulent Experts, *Rand Journal of Economics* 28 (1997), 107-119.

Emons, W., Credence Goods Monopolists, *International Journal of Industrial Organization,* 19 (2001), 375-389.

Gruber, J., J. Kim, and D. Mayzlin, Physician Fees and Procedure Intensity: The Case of Cesarian Delivery, *Journal of Health Economics* 13 (1999), 473-490.

Judd, K., The Law of Large Numbers with a Continuum of IID Random Variables, *Journal of Economic Theory* 35 (1985), 19-25.

Labelle, R., G. Stoddart, and T. Rice, A Re-examination of the Meaning of Supplier-Induced Demand, *Journal of Health Economics* 13 (1994), 347-368.

Marty, F., Capacity as a Determinant of the Supply for Physicians' Services, Discussion Paper, University of Bern (1998), www.staff.unibe.ch/emons/downloads/phy_0598.pdf.

Richardson, H., The Credence Good Problem and the Organization of Health Care Markets, Discussion Paper, Texas A&M University (1999).

Roth, H. R. and W. Stahel, Die ANOVA-Methode zur Prüfung der Wirtschaftlichkeit von Leistungserbringern nach Artikel 56 KVG, (2005), www.physicianprofiling.ch/CONGutachtenANOVADrRoth.pdf.

Sülzle, K. and A. Wambach, Insurance in a Market for Credence Goods, *Journal of Risk and Insurance* 72 (2005), 159-176.

Wennberg, J., B. Barnes, and M. Zubkoff, Professional uncertainty and the problem of supplier-induced demand, *Social Science & Medicine* 16 (1982), 811-824.