forthcoming in *Journal of Institutional and Theoretical Economics*

Incentive Compatible Reimbursement Schemes for Physicians

Winand Emons*
University of Bern, CEPR, CIRPÉE
Version July 2013

Abstract

Physicians choose capacity before demand materializes; actual demand may be higher or lower than capacity. 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 for all demand realizations. 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, http://staff.vwi.unibe.ch/emons. I thank Uwe Dulleck, Rudi Kerschbamer, Albert Ma, Fridolin Marty, Pius Matter, Gerd Mühlheusser, Klaus Neusser, and two referees 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. Furthermore, ex post the patient can typically not determine which treatment was required ex ante. It is often impossible to find out whether provided treatments were necessary or whether necessary treatments were not provided. From ex post observations the patient can never be certain of the quality of the treatments he obtained; therefore, such services have been termed credence goods (Darby and Karni (1973)).

This information advantage may induce physicians to behave opportunistically: they may recommend unnecessary yet profitable treatments or they may not perform urgently needed yet unprofitable treatments.[2] For example, 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.

[1] See, e.g., Brownlee (2007, p. 5) or Reilly and Evans (2009). Likewise, the Canadian Association of Radiologists estimates that 30% of imaging is unnecessary in the Canadian health care system; www.car.ca/uploads/news%20publications/car_cat_scan_eng.pdf (2009).

[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, as she elaborates in the book, the most powerful reason for overtreatment is that doctors and hospitals get paid more for doing more. Interestingly, in a 2009 survey among 627 US primary care physicians, only 3% of the respondents said money influences their practice, but most think money does influence the practice of other physicians: 62% said there would be fewer diagnostic tests if tests didn't create revenue for subspecialists, and 39% think the same of primary care doctors; dartmed.dartmouth.edu/winter11/html/disc_study/.

(1993)). 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. For Germany Jürges (2009) finds no evidence for demand inducement for statutorily insured patients; there is, however, demand inducement among the patients with private insurance which reimburses medical doctors at higher rates than their statutory counterparts. Even in China overprescription is routine; hospitals use these profits to subsidize underfunded operations (Economist 11/07/1998). Gruber et al. (1999) show that in the US the frequency of cesarian deliveries compared to vaginal deliveries positively reacts to fee differentials of health insurance programs. 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; by fraudulent behavior we mean that a physician performs unnecessary treatments or does not perform necessary treatments.[3] 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).[4] Patients are in need of a checkup. Some patients are in good condition and require no further treatment; the rest is in bad condition and needs treatment. After the

[3]Mark Twain describes this behavior as follows: "Well, then, says I, what's the use you learning to do right, when it's troublesome to do right and ain't no trouble to do wrong, and the wages is just the same?", (Huckleberry Finn, (1885, p. 128)), archive.org/stream/adventureshuckle00twaiiala#page/n9/mode/2up%7C. George Bernard Shaw writes: "That any sane nation having observed that you could provide for the supply of bread by giving bakers a pecuniary interest in baking for you, should go on to give a surgeon a pecuniary interest in cutting off your leg, is enough to make one despair of political humanity." (The Doctor's Dilemma: A Tragedy, (1906 p. xiii)), archive.org/stream/doctorsdilemmatr00shawuoft#page/xii/mode/2up/search/any+sane+nation+having+.

[4]The major difference to these papers is that here prices are set right at the outset and cannot adjust to the realizations of demand later on. This implies that prices alone cannot give proper incentives; see Section 4.

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.[5]

We consider a set of physicians, each of whom chooses a fixed capacity: when actual demand realizes, 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.

An insurer sets reimbursement terms. Then physicians choose their capacity levels; we focus on symmetric strategies. Nature then determines total demand; each physician gets an equal share thereof. This modeling implies that all doctors either have excess capacity or excess demand. Doctors decide how many of their patients they want to diagnose; patients who obtain no diagnosis end up with no service. Having diagnosed her patients, a doctor then decides whom to treat.

Physicians choose their capacity and their diagnosis and treatment policy so as to maximize profits. The insurer wants to induce non-fraudulent services. Moreover, he wants to implement some average capacity level which, in turn, implies that with positive probability demand may exceed capacity and vice versa.

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 patients get non-fraudulent services for all possible demand realizations. Consider, for example, equal compensation prices equalizing the profit per diagnosis with the profit per treatment. With these prices doctors are indifferent between diagnosis and treatment and, accordingly, provide honest services if demand exceeds capacity. Yet if doctors have excess capacity, they overtreat to use up their idle capacity.[6] By contrast, with a fully capitated scheme where the treatment

[5] 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.

[6] "Fee-for-service is especially inflationary in the context of physician oversupply; there

price is zero, the incentive to overtreat disappears and doctors behave non-fraudulently when they have excess capacity. But now doctors undertreat when they have excess demand because diagnosis is much more attractive than treatment. It is thus impossible to find fee-for-service schemes that give proper incentives for all possible demand realizations, i.e., when physicians have excess demand or when they have excess capacity.

In the next step we use the fact that the insurer has 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 with a particular doctor. In addition to the fees-for-services, the insurer can thus use a quota that states the maximum fraction of diagnosed patients per physician 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. The level of the prices determines the physicians' capacity choice: the higher the revenue per patient, the higher their capacity choice. Physicians make positive expected profits.

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 inferences about the appropriateness of the treatment he has received.[7] Most of this literature considers experts operating in a market environment. The only model we are aware of incorporating insurance in a credence good

is nothing more expensive than an underemployed specialist," Robinson (2001, p. 4).

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

set-up is Sülzle and Wambach (2005). They take prices as given and analyze the impact of co-insurance 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.[8] 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; they are such that the bad mechanic always wants to change the engine. Ely and Välimäki do not analyze how the bad mechanic's incentives can be aligned with prices. By contrast, we also determine reimbursement prices such that, together with the quota, experts have proper incentives and the outcome is efficient.

In the health economics literature physician-induced demand has been studied in a variety of models. Farley (1986) and De Jaegher and Jegers (2000) are models based on demand-setting and altruism. In Dranove (1988) patients make rational decisions about whether or not to accept a doctor's recommendation; informed patients will be subject to less inducement than less informed patients. Calcott (1999) and De Jaegher and Jegers (1999) model demand inducement as cheap talk games and derive equilibria with or without demand inducement. See McGuire (2000) for a survey of the earlier literature. The more recent literature stresses the gate-keeping role of general practitioners: besides diagnosis and treatment the general practitioner also refers patients to specialists; see, e.g., Brekke et al. (2007) or Allard et al. (2011).

[8]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 restrictions. Section 5 concludes.

2. The Model

An agent needs a medical checkup. During the period to come the individual may fall ill or he may stay healthy. At the time of diagnosis 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 should treat him; after the treatment the consumer is in good condition. A treatment is only possible after diagnosis.

Each physician makes a prior sunk capacity decision determining L units of time that she devotes to her practice. Since we normalize the doctors' reservation wage to 1, L also measures a physician's sunk cost. The capacity L can only be allocated between diagnosis and treatment: $d > 0$ is the time a doctor needs per diagnosis and $t > 0$ the time per treatment; given our normalization, d and t also measure the minimum average costs of diagnosis and treatment. Note that marginal costs are different from average costs. A doctor 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 t.[9] If there are additional variable costs per diagnosis and per

[9]We assume that diagnosing and treating a patient if necessary is efficient. If we normalize the utility of staying healthy to, say, 1 and falling sick to 0 monetary units, efficiency requires $p(q_h - q_\ell) > d + pt$. Treating a patient in bad condition increases his

treatment, the fees-for-service we introduce in the next section are simply the doctor's remuneration net of these variable costs.

There is a continuum of identical consumers the mass X of which is random; it has continuous density $g(X)$ over the support $[\underline{X}, \overline{X}]$. In units of time a capacity of $X(d + pt)$ is necessary to serve the entire market. Each physician gets an equal share of consumers. A doctor's demand is thus $x := X/n$ which is distributed on $[\underline{x}, \overline{x}]$ with density $f(x) := ng(X)$; denote the c.d.f. by $F(x)$. Patients' risks are independent and identically distributed. We assume that a continuum of such random variables sums to a non-random variable.[10] The size of a physician's demand is thus random whereas the fraction of her patients in need of treatment is non-random. Define $\lambda = L/(d+pt)$ as a doctor's capacity in terms of customers given non-fraudulent behavior. Since we look for symmetric strategies, total capacity equals $n\lambda$. According to whether $n\lambda \lessgtr X$, there is too little/sufficient/excess capacity in the market. Due to our symmetry assumption, market conditions translate into the individual physician level: all doctors have either excess capacity or they all face excess demand.

Let us now look at a doctor's incentives. After diagnosis the physician knows the patient's condition. When the patient is in bad condition, she can perform a treatment that turns him into good condition. Yet she can also 'treat' a patient in good condition; in this case the physician wastes t units of time, leaving the patient 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)).

If the patient is in good condition, the medical doctor can recommend

utility by $(q_h - q_\ell)$; with probability p he is in bad condition. Our diagnosis corresponds to Dulleck and Kerschbamer's (2006) cheap treatment; their expensive treatment corresponds to our "diagnosis cum treatment".

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

no treatment. Nevertheless, the same recommendation is also possible when the patient is in bad condition. We will refer to this type of fraud as undertreatment (Dulleck and Kerschbamer (2006)).[11]

Ex post the patient cannot find out whether he was treated unnecessarily or whether necessary treatment 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 got.

Note that we assume diagnosis and treatment to be verifiable. This assumption seems appropriate for physicians whose patients necessarily take part in any treatment. It is not appropriate for, e.g., a consumer who sends his gadget to a service center. When the gadget is returned, the customer is unable to tell whether somebody in the repair center has actually worked on it. 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.[12]

All patients have full insurance from one insurance company. The insurer reimburses the physicians. The sequence of the events is as follows. The insurance company chooses the reimbursement terms. Next physicians choose their capacity. Then nature chooses the total mass X or, equivalently, a physician's mass x of patients. A physician then decides how many patients $\mu \leq x$ she diagnoses; $x - \mu$ patients get no service. After having diagnosed her μ patients, the doctor then decides whom to treat.

Medical doctors maximize profits. The insurer wants to find reimbursement terms that induce non-fraudulent behavior. Moreover, the insurer wants to implement an average aggregate capacity level, meaning at the physician level a capacity $\lambda \in (\underline{x}, \overline{x})$. We thus assume neither aggregate capacity $\underline{X}$ nor capacity $\overline{X}$ is optimal.[13] Accordingly, with positive probability

[11]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).

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

[13]Diagnosis and treatment are efficient; see footnote 9. This does, however, not imply that the maximum capacity $\overline{X}$, which is needed with probability zero, is efficient. Rather

there will be excess capacity and with positive probability there will be excess demand. 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 T per performed treatment. Consider the subgame starting after the physician has chosen a capacity of L units of time which by then is sunk. In terms of patients the physician has capacity $\lambda < \overline{x}$ given honest behavior. Apparently, her behavior depends on the size of her clientele x relative to her capacity λ. According to whether $x \gtreqless \lambda$ 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 physician's incentives also depend on the relative profitability of diagnosis to treatment which, in turn, is determined by the prices D and T. If the doctor has too many patients, she only faces (at the margin) her time constraint. She compares the profit per hour treatment T/t with the profit per hour diagnosis D/d.[14] 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; here we assume that if the doctor is indifferent between non-fraudulent and fraudulent behavior, she opts for the honest one.[15]

Lemma 1:

i) If $x > \lambda$, the physician is honest if and only if $T = tD/d$;
ii) if $x = \lambda$, the doctor is honest if and only if $T \leq tD/d$;
iii) if $x < \lambda$, the doctor is honest if and only if $T = 0$.

than specifying $g(\cdot)$ and derive the optimal capacity level, we show that any interior capacity level can be implemented.

[14] Recall that the capacity cost is sunk; our results, however, do not change if we define profits per hour as $(D-d)/d$ and $(T-t)/t$.

[15] This result corresponds to Lemma 1 in Emons (1997).

Proof: i) If $x > \lambda$, 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 T/t compared to the profit per hour diagnosis D/d. If $T = tD/d$, she is indifferent between diagnosis and treatment and, therefore, behaves honestly. If $T > tD/d$, she prefers treatment to diagnosis and thus overtreats and undertreats if $T < tD/d$.

ii) If $x = \lambda$, the physician fully utilizes her capacity with non-fraudulent behavior. If $T < tD/d$, 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 $L - xd$; honestly treating the patients in bad condition of her clientele just exhausts her capacity. If $T = tD/d$, the argument is along similar lines as i). If $T > tD/d$, the physician strongly prefers treatment to diagnosis. Hence, she will treat all patients she diagnoses.

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

To explain Lemma 1 consider Figure 1. Along the line $T = tD/d$ the profit per hour diagnosis equals the profit per hour treatment: on this equal compensation line the physician is indifferent between diagnosis and treatment so that with too many patients she picks efficient treatment. She diagnoses $\mu = \lambda$ patients and treats the fraction p thereof; the remaining $(x - \lambda)$ patients get no treatment.

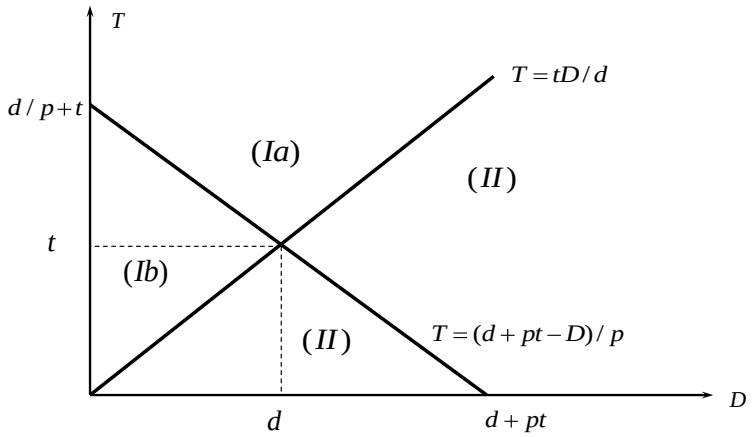

Figure 1: The equal compensation and the zero-profit lines

In regions (Ia) and (Ib) where $T > tD/d$ 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 $T < tD/d$ 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 overtreat as long as $T > 0$. Only when $T = 0$ 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. When we rule out the prices $D = T = 0$ which provide no incentives whatsoever, we can summarize our findings as follows:

Proposition 1: *If the insurer uses simple fee-for-service reimbursement schemes (D, T), there exists no set of prices under which for all demand realizations patients get non-fraudulent services.*

In Figure 1 we have also depicted the line $T = (d + pt - D)/p$. All prices along this line generate zero-profits when a physician serves λ patients non-fraudulently. Suppose, for example, the insurance company reimburses equal compensation prices (d, t). Then if physicians have enough or too many customers, they have proper incentives and they make zero profits. Yet, if a physician has excess capacity, she will overtreat.

If insurers use, say, the fully capitated reimbursement $(d + pt, 0)$, physicians have proper incentives with enough or too few patients.[16] Yet, if there is excess demand, doctors will undertreat.

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).[17]

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

[17] In Dulleck and Kerschbamer (2006) doctors have no capacity constraint, t and d are are marginal costs so that equal compensation (equal markup) prices satisfy $T - t = D - d$. Another set-up with capacity constrained experts can be found in Richardson (1999).

4. Fee-for-service with Quantity Restrictions

Let us now use the fact that the 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.

The insurer offers reimbursement schemes (D, T, 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 efficient 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.

Lemma 2: *Let $z = p$.*
i) If $x > \lambda$, the physician is honest if and only if $T \geq tD/d$;
ii) if $x \leq \lambda$, the doctor is honest for all prices (D, T).

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

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

If $T > tD/d$, 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 $x = \lambda$, the physician fully uses her capacity with non-fraudulent behavior. If $T = tD/d$, she has proper incentives and uses up her capacity by honestly serving all patients. If $T < tD/d$, she prefers diagnosis to treatment. She diagnoses all patients; to use up her remaining time $L - xd$ she has to treat. Honestly treating the patients in bad condition just exhausts her capacity. If $T > tD/d$, 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 $x < \lambda$, the physician has unused capacity with non-fraudulent behavior. As long as $D > 0$, she will diagnose all x patients. As long as $T > 0$, she would like to treat more than px patients to use her idle capacity. Yet she cannot bill more than px patients for treatment. ■

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 $T > 0$, 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.

Define region (Ia) as all prices (D, T) on and above the equal compensation line and *above* the zero-profit line. Lemma 2 implies the following result:

Proposition 2: *Under the reimbursement schemes (D, T, p) with (D, T) in region (Ia) all patients get non-fraudulent service. Each physician chooses the capacity $\lambda^* \in (\underline{x}, \overline{x})$ solving $F(\lambda^*) = 1 - (d + pt)/(D + pT)$; a physician's equilibrium profit is $(D + pT) \int_{\underline{x}}^{\lambda^*} xf(x)dx > 0$.*

Proof: Lemma 2 implies that for reimbursement schemes (D, T, p) with (D, T) in region (Ia) physicians have proper incentives whatever their demand. Given non-fraudulent behavior, physicians choose their capacity λ so as to maximize $(D + pT)[\int_{\underline{x}}^{\lambda} xf(x)dx + \lambda(1 - F(\lambda))] - \lambda(d + pt)$. Solving the first order condition yields $F(\lambda^*) = 1 - (d + pt)/(D + pT)$. ■

Let us first comment on the capacity choice. If the insurer picks prices on the zero-profit line, $D + pT = d + pt$ so that $\lambda = \underline{x}$. Doctors pick the capacity level they can sell for sure and break even. For prices above the zero-profit line (region (Ia)), revenue per customer $D + pT > d + pt$ and physicians choose capacity $\lambda > \underline{x}$; with positive probability a doctor has idle capacity. Physicians make positive expected profits. The capacity level increases with the revenue per customer and is below $\overline{x}$.

There exist thus reimbursement schemes (D, T, z) inducing non-fraudulent behavior for all realizations of demand. With the level of the prices (D, T) the insurer controls the capacity that is provided by physicians. The higher the prices, the more capacity they provide. Physicians' profits increase with the price level.

In our set-up the insurer knows the equilibrium capacity level of a physician. Nevertheless, for our incentive scheme to work once capacity is chosen, the insurer need *not* know the physician's actual capacity level.[18]

A few qualifying remarks are in order. In our model 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 sums to a non-random variable. 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 10. The reimbursement quota $z = p$, therefore, coincides with the actual number of patients in need of treatment.

With a finite population the actual number of patients in need of treatment will typically be different from the expected value. This creates problems at the aggregate insurance and at the individual physician level. If at the insurance level the actual fraction of patients in need of treatment is above p, our quota $z = p$ enforces undertreatment. If the actual fraction is below p, doctors will overtreat. This problem becomes smaller, the more clients the insurance company has.[19]

At the physician level our results change as follows. Assume, for the sake the argument, that at the insurance level the actual realization equals the expected value p. When deciding on how many patients to diagnose the doctor bases her decision on the expected value p as in our set-up; in particular, a physician with excess demand diagnoses λ patients. Nevertheless, when it comes to the treatment decision the physician treats the fraction $z = p$

[18] 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.

[19] In Switzerland Santésuisse, the association of the Swiss health insurers, collects and aggregates the data of the individual insurers about a physician's behavior to check, e.g., that the doctor doesn't bill more than 24 hours a day.

independently of the actual needs of her clientele. Thus, if the actual fraction of patients in need of treatment is below z, the physician overtreats and undertreats if it is above z. To sum up: with a finite population our quota system works best for large insurance companies the clientele of which are treated by relatively few physicians.

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, our results continue to hold. If, however, there is a selection bias such that some physicians get on average less healthy, i.e., higher p 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. With excess capacity a doctor with high p patients does better than her colleague with a low p group; she uses up more of her idle capacity. With excess demand, if treatment is more attractive than diagnosis, a physician also prefers high p patients; with equal compensation prices doctors facing excess demand are indifferent as to the health status of their clientele. Fee-for-services with an adjusted quota thus make less healthy patients in expectation attractive for physicians. This is in stark contrast to capitation where physicians try to skim the healthy and avoid the ill.

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 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. If there are several treatment options and the physician has private information on what the best treatment is for the patient, profits per hour treatment have to be equalized across all treatments and quotas for each treatment have to be implemented.

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. In

the U.S. physicians are paid bonuses to restrict the percentage of patients who are given referrals (Grumbach et al (1998)).[20] Such bonuses may give incentives not to overtreat at the margin. If, however, excess capacity is sufficiently large, overtreating is more profitable than cashing in on the bonus. 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.[21] Here it is unclear what the average actually measures: inefficiencies may be compared with inefficiencies.

In Germany physicians are endowed each quarter with a so called 'budget'. Once they exceed this budget, they get paid less per treatment. At first insurers did not pay at all for any service provided outside this budget; the quota (Praxisbudget) was strict as is our quota z. Poorly set quotas made it, however, difficult for statutorily insured patients to see their doctor at the end of a quarter because her budget was exhausted. Therefore, in 2009 the quota was softened (Regelleistungsvolumen); now the doctor's reimbursement goes down from 100% to 75% to 50% to 25% once she exceeds the budget. It is not entirely clear how the budget is determined; since it is defined for the entire practice, it is perhaps too broad. Furthermore, the budget is based on past behavior. It may thus lead to ratcheting: staying within the budget today may lead the regulator to lower the budget tomorrow. Nevertheless, the Praxisbudget seems to have curbed overtreatment, albeit at the expense of some undertreatment. The new Regelleistungsvolumina try to deal with the undertreatment problem while maintaining the virtues of quotas concerning overtreatment. Our results tend to support the German approach. Yet, a more sophisticated use of treatment records seems warranted.

Besides physician-induced demand, the classic moral hazard problem that consumers with full insurance tend to overconsume health care is a major driver of overtreatment. We have not addressed consumer moral hazard in our model. It is, however, obvious that our treatment quota also curbs excessive treatment demanded by fully insured consumers. An appropriately designed treatment quota, therefore, not only curbs sellers' incentive to overtreat but at the same time also curbs patients' incentives to overconsume. Consumer based instruments like co-payments solve the consumer

[20] Consumers, however, seem to disapprove of cost control bonuses (Gallagher et al (2001)).

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

moral hazard problem but are not effective to deal with physician-induced demand.

A last remark concerns the relation of our results here to our earlier results where the market solves the expert problem without quantity restrictions. In Emons (1997) experts set prices after they have chosen capacity. If capacity falls short of demand, experts charge high equal compensation prices and make positive profits; if capacity exceeds demand, prices are zero and experts make losses. In either case experts provide honest services. The experts' capacity choices (entry decisions) are mixed so that on average they break even. The important difference to the paper at hand is thus when prices are actually set. Whereas in Emons (1997) prices adjust to the capacity/demand realizations and induce proper behavior, this is not possible in the current set-up where prices are set beforehand. This is typically the case in health care markets.[22]

If in our scenario a patient pays the doctor himself, he has only one observation of her treatment policy. Quantity restrictions are not feasible and we encounter fraudulent behavior. We then use the special feature of health care markets, namely that patients have insurance and doctors are reimbursed by the insurance companies. The insurance companies get information about the medical doctor's overall treatment behavior which enables them to employ quantity restrictions.[23]

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 with positive probability. 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 the insurer observes a physician's actions for the entire set of his policy holders. This allows the insurer to set a quota which

[22] In Emons (2001) the monopolistic expert chooses prices and capacity together so that the issue of over- or undercapacity does not arise in equilibrium.

[23] Our approach thus seems to be applicable to other markets where insurers reimburse experts, such as, e.g., the market for legal services.

states the maximum fraction of diagnosed patients for whom he actually pays the treatment. If the insurer sets this quota equal to the fraction of patients in need of treatment, he curbs overtreatment. Therefore, only the undertreatment problem remains which is solved by prices making diagnosis not more attractive than treatment.

References

Allard, M., Jelovac, I., and P. Léger (2011), "Treatment and Referral Decisions under different Physician Payment Mechanisms," *Journal of Health Economics,* 30, 880-893.

Brekke, K., Nuscheler, R., and O. Straume (2007), "Gatekeeping in Health Care," *Journal of Health Economics,* 26, 149-170.

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

Calcott, P. (1999), "Demand Inducement as Cheap Talk," *Health Economics,* 8, 721-733.

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

De Jaegher, K. and M. Jegers (2000), "A Model of Physician Behavior with Demand Inducement," *Journal of Health Economics,* 19, 231-258.

De Jaegher, K. and M. Jegers (2001), "Physician-Patient Relationship as a Game of Strategic Information Transmission," *Health Economics,* 10, 651-668.

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

Dranove, D. (1988), "Demand Inducement and the Physician/Patient Relationship," *Economic Inquiry,* 26, 281-298.

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

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

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

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

Farley, P. (1986), "Theories of the Price and Quantity of Physician Services," *Journal of Health Economics,* 5, 315-333.

Gallagher, T., R. St. Peter, M. Chesney, and B. Lo (2001), "Patients Attitudes Toward Cost Control Bonuses for Managed Care Physicians," *Health Affairs,* 20, 186-192.
Gruber, J., J. Kim, and D. Mayzlin (1999), "Physician Fees and Procedure Intensity: The Case of Cesarian Delivery," *Journal of Health Economics,* 13, 473-490.
Grumbach, K., D. Osmond, K. Vranizan, D. Jaffe, and A. Bindman (1998), "Primary Care Physicians Experience of Financial Incentives in Managed Care Systems," *New England Journal of Medicine,* 339, 1516-1521.
Judd, K. (1985), "The Law of Large Numbers with a Continuum of IID Random Variables," *Journal of Economic Theory,* 35, 19-25.
Jürges H. (2009), "Health Insurance Status and Physician Behavior in Germany," *Journal of Applied Social Science Studies (Schmollers Jahrbuch),* 129, 297-308.
Labelle, R., G. Stoddart, and T. Rice (1994), "A Re-examination of the Meaning of Supplier-Induced Demand," *Journal of Health Economics,* 13, 347-368.
Marty, F. (1998), "Capacity as a Determinant of the Supply for Physicians' Services," Discussion Paper, University of Bern, staff.vwi.unibe.ch/emons/downloads/phy_0598.pdf.
McGuire, T. (2000), "Physician Agency," in *Handbook of Health Economics* (Culyer A. and J. Newhouse, eds.), Elsevier, Amsterdam.
Richardson, H. (1999), "The Credence Good Problem and the Organization of Health Care Markets," Discussion Paper, Texas A&M University.
Reilly B., and A. Evans (2009), "Much ado about (doing) nothing," *Annals of Internal Medicine,* 150, 270-271.
Robinson J. (2001), "Theory and Practice in the Design of Physician Payment Incentives," *Milbank Quarterly,* 79, 1-17.
Roth, H. R. and W. Stahel (2005), "Die ANOVA-Methode zur Prüfung der Wirtschaftlichkeit von Leistungserbringern nach Artikel 56 KVG," www.physicianprofiling.ch/CONGutachtenANOVADrRoth.pdf.
Sülzle, K. and A. Wambach (2005), "Insurance in a Market for Credence Goods," *Journal of Risk and Insurance,* 72, 159-176.
Wennberg, J., B. Barnes, and M. Zubkoff (1982), "Professional Uncertainty and the Problem of Supplier-induced Demand," *Social Science & Medicine,* 16, 811-824.

Winand Emons
Departement Volkswirtschaftslehre
Universität Bern
Schanzeneckstrasse 1
3001 Bern
Switzerland
E-mail:
winand.emons@vwi.unibe.ch