TILBURG UNIVERSITY

Provider Competition and Over-Utilization in Health Care

Boone, Jan; Douven, R.C.M.H.

Document version:
Early version, also known as pre-print

Publication date:
2014

Link to publication

Citation for published version (APA):
Boone, J., & Douven, R. C. M. H. (2014). Provider Competition and Over-Utilization in Health Care. (CentER Discussion Paper; Vol. 2014-055). Tilburg: Economics.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright, please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 17. Jul. 2018

CentER

Discussion Paper

No. 2014-055

PROVIDER COMPETITION AND OVER-UTILIZATION IN HEALTH CARE

By

Jan Boone, Rudy Douven

29 September, 2014

ISSN 0924-7815
ISSN 2213-9532

TILBURG UNIVERSITY

Provider competition and over-utilization in health care*

Jan Boone† and Rudy Douven‡

September 22, 2014

Abstract

This paper compares the welfare effects of three ways in which health care can be organized: no competition (NC), competition for the market (CfM) and competition on the market (CoM) where the payer offers the optimal contract to providers in each case. We argue that each of these can be optimal depending on the contracting environment of a speciality. In particular, CfM is optimal in a clinical situation where the payer either has contractible information on provider quality or can enforce cost efficient protocols. If such contractible information is not available NC or CoM can be optimal depending on whether patients react to decentralized information on quality differences between providers and whether payer's and patients' preferences are aligned.

Keywords: competition, health care, selective contracting, over-utilization, mechanism design

JEL classification: D82, L5, I11

*We benefited from discussions with Rein Halbersma and seminar participants at CPB. Jan Boone gratefully acknowledges financial support from the Netherlands Organisation for Scientific Research (NWO) through a Vici grant. Douven worked on the paper as a Harkness Fellow 2013/2014 at Harvard Medical School, supported by the Commonwealth Fund, a private independent foundation based in New York City, and the Dutch Ministry of Health. The views presented here are those of the authors and not necessarily those of their sponsors.

†CentER, TILEC, Department of Economics, Tilburg University, CPB and CEPR. Email: j.boone@uvt.nl.

‡CPB Netherlands Bureau for Economic Policy Analysis, Erasmus University Rotterdam and Harvard Medical School. Email: r.douven@cpb.nl

1. Introduction

Due to numerous imperfections in health care markets, it is not clear what form provider competition should take. Countries have different modes of provider competition and we do not know which one is actually optimal. This is becoming more important as policy makers are struggling with the question how to organize health care (McClellan, 2011). In particular, with health costs rising, it has become a priority to reduce over-utilization (Chandra and Skinner, 2012; Chandra et al., 2011).[1] As competition tends to raise production, it is not clear that provider competition is desirable in this context. Finally, provider contracts are changing from fee-for-service and capitation to contracts explicitly using quality indicators.[2] How does the availability of quality information affect optimal provider interaction?

Our starting point is a health care market in which a welfare maximizing payer contracts providers that can differ in their quality of health services. In this market we introduce two imperfections. First, asymmetric information between the payer and the provider can result in over-utilization of health services. Insured patients face few constraints for consuming health care and physicians tend to focus on the effectiveness of treatments. Thus, both actors tend to pay too little attention to cost efficiency (Chandra and Skinner, 2012). Second, preferences of payer, providers and patients are not aligned (Arrow, 1963). As mentioned, patients and physicians prefer effective treatments irrespective of costs while cost-effectiveness is important for the payer. We confront this setting with three archetypes of provider competition.

The first mode is no competition (NC) between providers: this is the situation where each provider is a monopolist. A practical way to implement NC is to divide a country into regions, each region has one provider and patients have to visit the provider in their region. NC captures regionally organized health care by a public payer, such as in Portugal or Sweden (this is sometimes –dismissively– referred to as postal code health care). With provider competition, there are two modes to consider: payer-driven and patient-driven competition (Dranove et al., 1993; McGuire, 2011b). With payer-driven competition, or competition for the market (CfM), the payer bargains in each region with providers and contracts a (strict) subset of these providers. This is referred to as the payer's network. Providers have to compete to become part of the network. This resembles competition in the US, where in the employer-based insurance market employers contract pre-selected providers for their employees. We

[1] Suggestive evidence of over-utilization comes from differences in utilization rates across providers, geographical regions or countries. See e.g. the work by Wennberg (2010) and Skinner (2012) who consider several treatment categories, such as treatments for heart attacks, back surgery or tonsillectomy. The phenomenon of over-provision is also referred to as "flat of the curve medicine" (Getzen, 2007).

[2] For example, Accountable Care Organizations in Medicare (McClellan et al., 2010) and the Alternative Quality Contract in the employer based insurance market (Song et al., 2012).

denote patient-driven competition: competition on the market (CoM). This is the way competition has been introduced in Europe in countries like Belgium, the Netherlands and the UK (Propper and Leckie, 2011). Patients have full provider choice and the payer pays providers based on the number of patients treated. Providers compete on the market for patients.

In the real world, health care systems tend to be combinations of these three archetypes.[3] To illustrate, both NC and CfM usually leave some choice for patients–but less choice than CoM. To facilitate the exposition, we focus on the extremes where there is no patient choice in NC and CfM.

It turns out that the optimal way to organize provider competition depends on the information available to payer and patients. We distinguish two information scenarios: first and second best contracting opportunities. Under *first best contracting* one (or both) of the following is possible: (i) payer can enforce medical protocols limiting physicians' decisions to cost efficient treatments; (ii) quality indicators make a physician's quality contractible for the payer. Under *second best contracting*, neither of these two options is available. With second best contracting, we consider the role of decentralized information. Patients may have an idea about physicians' quality differences by learning from each others' experiences ("word of mouth"). This captures the traditional role of markets as aggregating decentralized information. This information is not contractible at the centralized (payer) level but patient streams (under CoM) can signal quality. Payer can use patient streams when paying providers but should not encourage ineffective treatments.

The main goal of our paper is to introduce a workhorse model in which the welfare effects of the three competition modes can be analyzed. Our results can be summarized as follows (where the main issues are indicated by italics). With first best contracting opportunities, CfM is optimal if patients' *travel costs* are negligible. Payer chooses the best provider and contracts only with her. With second best contracting opportunities, CfM leads to a welfare loss because it is *biased in favor of low quality* providers (lemma 2). CfM can still be optimal in this case if patients' and payer's *preferences are misaligned* (propositions 2 and 3). If, on the other hand, patients react to *decentralized information* by choosing the best provider, CoM is optimal. If patients have a preference to visit the closest provider, NC is optimal. By preventing that *patients shop around* for a provider (lemma 3), NC (compared to CoM) reduces over-treatment by low quality providers. The obvious drawback of NC is *mis-allocation*: half the market is bound to an inferior provider if providers differ in quality.

To the best of our knowledge, this is the first paper to analyze the welfare effects of different modes of provider competition. Somewhat related is Gaynor et al. (2000) who analyze whether an exogenous increase in provider prices (due to provider market power) can raise welfare by reducing over-treatment.

[3] An example of a combination of CfM and CoM in one system is Medicare in the US. Traditional fee-for-service Medicare resembles CoM while Medicare Advantage resembles CfM.

Their main result is that lower provider prices cannot reduce welfare if insurers adapt their coinsurance rates. However, an exogenous price change is different from a change in mode of competition. In our model, prices are endogenous and change with the competitive setting. Further, we allow for non-linear contracts (instead of focusing on fee-for-service contracts). On the other hand, we keep the demand side (coinsurance) fixed.

As we derive optimal contracts in each setting, our paper is related to two strands of the literature analyzing contracts. First, there is a health economics literature that analyzes payment systems (for a given organization of the health care market; e.g. the physician is a monopolist). Excellent summaries of this literature include Chalkley and Malcomson (2000); McGuire (2000, 2011a). Findings include that over-treatment can be reduced by combining a capitation fee with a fee-for-service that is below marginal treatment cost. Chalkley and Malcomson (2000) criticize this literature as it does not consider a provider's decision to treat only some patients (in most models either demand is fixed ex-ante or a physician treats either all patients or none). Chalkley and Malcomson (1998) argue that a payer can prevent over-treatment by including the number of treated patients in the payment contract. We do allow for such contracts. Mougeot and Naegelen (2005) analyze so-called global budgets: total health costs summed over all providers is fixed ex ante. They argue that in a competitive market global budgets do not implement the first best outcome. Our set-up allows for such budgets and we confirm that they are not optimal (neither in first nor in second best).

Second, whereas the payment literature considers combinations of existing arrangements like fee-for-service and capitation fees, we use mechanism design to derive properties of the optimal contract.[4] Compared to the mechanism design/optimal regulation literature (see e.g. Laffont and Tirole, 1993), our analysis has two features that differ from standard models. First, usually in this literature high quality agents want to mimic low quality agents (say, to reduce their effort costs). The high quality agent then receives an information rent to prevent this. In our model with over-treatment, low quality physicians have an incentive to mimic high quality physicians (in order to raise production). Hence, low quality providers receive an information rent. Second, we analyze different ways in which the principal (payer) lets the agents (providers) interact in the health care market.

This paper is organized as follows. Section 2 introduces patients' and physicians' preferences. It defines the (health economic) concept "treatment efficiency". In section 3, we consider the case where the payer faces a monopolist provider. We define the (mechanism design) concepts of first and second best outcomes. We show that low quality providers receive an information rent if physicians focus on effective treatments. Then we compare two situations in which patients have no provider choice:

[4] Use of mechanism design in health economics is rare. To illustrate, in the recent Handbook of Health Economics, Volume 2, "mechanism design" is not mentioned in the text.

NC and CfM. Section 5 analyzes the case where patients are free to choose their provider (CoM). Section 6 analyzes two additional reasons why patients' and payer's preferences diverge: providers have different costs and patients face travel or switching costs to visit a provider. Finally, we discuss policy implications: for each competition mode we give examples of specialties where this mode is optimal. Proofs can be found in the appendix.

2. Utility

This section introduces patients, cost and benefits of treatment and physician's intrinsic motivation to treat (financial motivation is introduced in the next section).

2.1. Costs and benefits of treatment

We index patients by ω, which is uniformly distributed on $[0,1]$. The health benefit of a treated patient is denoted by $v(\omega)$, which can be interpreted in terms of gained qalys (quality adjusted life years) from treatment. We assume that higher ω cases gain more from treatment: $v'(\omega) > 0$.[5]

Next we introduce differences in practice styles, and hence quality, across physicians. For example, some hospitals may have better trained physicians or may use more technologically advanced equipment. There is a growing literature that documents quality differences in hospitals (see e.g Gowrisankaran, 2008). We model quality differences by simply assuming that physicians in a hospital can be of low (basic)[6] or high (i.e. higher than basic) quality. If a patient ω is treated in a low, respectively high, quality hospital, her benefit from treatment is given by $v_l(\omega)$, respectively $v_h(\omega)$, with $v_h(\omega) > v_l(\omega)$ for each $\omega \in [0,1]$. Hence, the utility for a patient being treated by a high quality physician is higher (for each ω) compared to a low quality physician. To simplify the exposition, we assume treatment costs c are the same for each ω and across hospitals. In section 6 we relax the latter assumption.

This paper is on the intersection of health economics and mechanism design and hence we will define benchmarks that relate to both these literatures. The first benchmark is treatment efficiency: treatment is efficient if and only if patient's utility of the treatment exceeds the treatment cost.

[5] Although we will assume later in the text $v(0) \geq 0$, one could allow for the fact that some patients receive negative outcomes (negative qaly's). Thus, assuming that for low ω, $v(\omega) < 0$ allows a physician to produce negative benefits (Evans, 1974).

[6] Low quality here does not refer to a situation where quality is so low that the hospital should be closed.

Definition 1 *We define* ω_i^* *by*

$$\begin{aligned} v_l(\omega_l^*) &= c \\ v_h(\omega_h^*) &= c \end{aligned} \tag{1}$$

Under treatment efficiency, patient ω *is treated by i-provider* ($i \in \{l, h\}$) *if and only if* $\omega \geq \omega_i^*$.

As illustrated in figure 1, $v_h(\omega) > v_l(\omega)$ implies that $\omega_h^* < \omega_l^*$: it is efficient for a high quality physician to treat more patients than a low quality provider. To rule out trivial cases, we assume $v_l(0) < c < v_h(1)$.

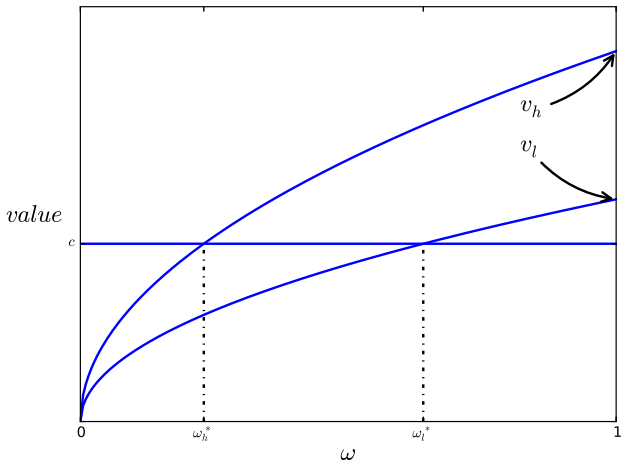


Figure 1: Treatment efficiency: treat all patients with $\omega \geq \omega_i^*$.

If a physician treats patients $\omega < \omega_i^*$ we say that there is over-treatment (compared to treatment efficiency). If a physician does not treat (some) patients with $\omega > \omega_i^*$, we say that there is under-treatment.

The fundamental problem in health care is that ω is not contractible for the payer. Only the physician knows the value of a treatment for the patient and she decides whether to treat or not based on her intrinsic and financial motivation.

2.2. Physicians

Since we are ultimately interested in welfare, we need to be specific about the effects of contracts and competition modes on physician utility. An important aspect here is the physician's intrinsic motivation. Appendix B derives a dis-utility function for a physician (not) treating patients with the following underlying idea. The more value a physician can create for the patient, the lower the dis-utility of treating this patient and the higher the dis-utility of not treating this patient. Consider a

patient walking into a physician's office where the physician believes that her treatment can make a big difference. If the physician for some reason is not allowed to treat this patient, her dis-utility –of not treating– will be big (compared to not being able to treat a patient where she can hardly help). If she is allowed to treat the patient, her dis-utility –of treatment– is small (compared to a patient whom she treats but who hardly experiences a health benefit due to treatment).

This utility structure implies that a physician uses a cut-off value $\bar{\omega}_i$ such that all $\omega > (<)\bar{\omega}_i$ are (not) treated. In other words, physicians ration efficiently. There is some empirical support for efficient rationing: in case of heart attacks (Chandra and Staiger, 2007) and Cesarean sections (Baicker et al., 2006). Other support comes from the Netherlands. In 2001, the government stopped using budgets. This led to strong growth of admissions without a clear medical diagnosis (Vijsel and Westert, 2011). This was interpreted as an increase in less beneficial treatments. Put differently, with the budget in place, the most deserving patients were treated.[7]

The appendix derives that a physician has an intrinsic optimum $\tilde{\omega}_i$ where her dis-utility is lowest. We normalize dis-utility such that if a physician can treat all patients with $\omega \geq \tilde{\omega}_i$ her dis-utility equals zero. Any other cut-off value $\omega_i \neq \tilde{\omega}_i$ leads to strictly positive dis-utility for the physician.[8] For concreteness we distinguish between three physician types in terms of their intrinsic motivation.

Definition 2 *Let $D_i(\omega_i) \geq 0$, $i = h, l$ denote the dis-utility of an i-quality physician who treats all patients with $\omega \geq \omega_i$ (and no patient with $\omega < \omega_i$). We assume that $D_i(.)$ is convex and*

$$D'_h(\omega) > D'_l(\omega) \tag{2}$$

for all $\omega \in [0, 1]$.

A physician's intrinsic optimum, $\tilde{\omega}_i$, is defined as $D_i(\tilde{\omega}_i) = 0$.

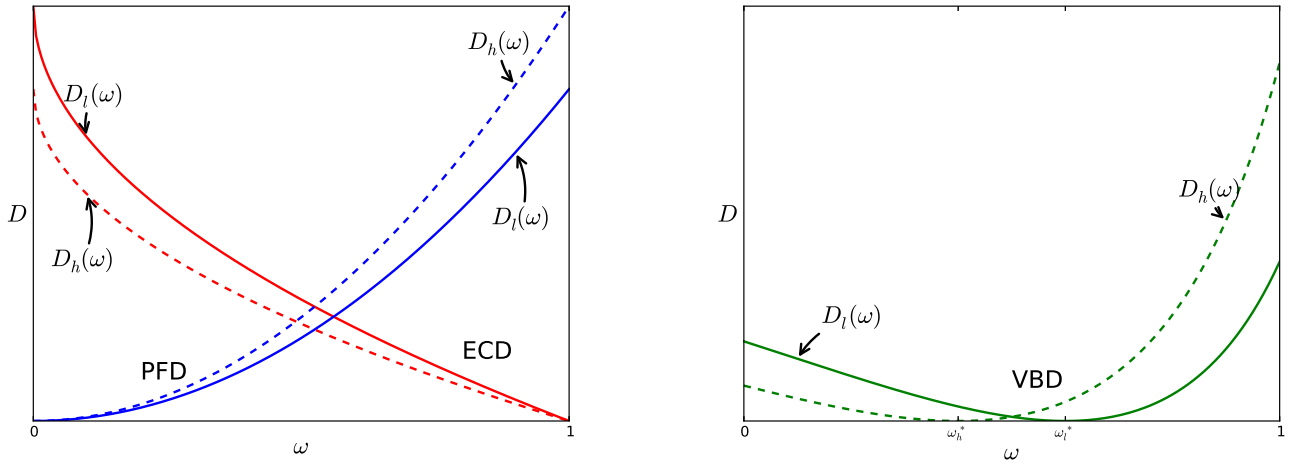
We distinguish three different types of physicians:

- **PFD-physician** *Patient Focused Disutility physician: $\tilde{\omega}_h = \tilde{\omega}_l = 0$.*
- **VBD-physician** *Value Based Disutility physician: $\tilde{\omega}_h = \omega^*_h < \omega^*_l = \tilde{\omega}_l$.*
- **ECD-physician** *Effort Cost Disutiliy physician: $\tilde{\omega}_h = \tilde{\omega}_l = 1$.*

[7] The results below depend on physicians' efficient rationing. If physicians do not ration efficiently –say they randomize to ration– the payer's decision is, in fact, quite simple. If the average (over different patients ω) value of the treatment is below the cost, the treatment should not be covered at all. Otherwise, the payer can cover the treatment but setting a budget does not raise the average value of the treatment.

[8] This is reminiscent of Giuffrida and Gravelle (2001) where a physician experiences dis-utility of patient demand management –either by increasing or decreasing demand.

Appendix B derives these properties of $D_i(\omega)$ based on a physician's dis-utility described above. Equation (2) says that moving the threshold ω_i to the right (treating less patients) is more costly for a h-physician because h-physician can create more value $v_h(\omega)$ for a patient than l-physician. The three physician types are illustrated in figure 2.



(a) Physician disutility under PFD and ECD

(b) Physician disutility under VBD

Figure 2: Disutility functions $D_l(\omega)$ (solid) and $D_h(\omega)$ (dashed).

A PFD-physician seeks to maximize the perceived health of a patient irrespective of the cost c. She has no dis-utility from treating a patient but obtains dis-utility from withholding effective treatment from a patient. Assuming $v_i(0) \geq 0$ for $i = l, h$, her dis-utility equals zero if she treats all patients $\omega \geq 0$. PFD reflects conventional physicians who do their best to improve health outcomes of all patients and to save as many lives as possible.[9] According to Chandra and Skinner (2012) the vast majority of physicians in the US do exactly this. PFD-physicians are likely to be found in situations where patients are fully insured or face only low out-of-pocket payments. These patients may push physicians towards the point of treating as much as possible, as long as treatments are effective ($v_i(\omega) \geq 0$). PFD physicians have more room for maneuver in clinical situations where the grey area of medicine is large and there are no authoritative guidelines or consensus treatment recommendations (section 7 discusses some examples). A high quality PFD-physician can create more benefit for each patient than a low quality PFD-physician and therefore her dis-utility (of not treating a patient) is higher for each ω (see figure 2a).[10]

Whereas the PFD-physician does not regard treatment costs, the VBD physician takes these costs

[9] If we would allow for $v(0) < 0$, PFD stands for Production Focused Disutility: such a physician wants to treat any patient walking into her office even if the patient suffers from the treatment ($v(\omega) < 0$).

[10] Although we postpone physicians' financial incentives till the next section, another possible interpretation of our

into account. Hence her intrinsic optimum coincides with treatment efficiency and obtains dis-utility if she has to treat (cannot treat) patients with $\omega < (>)\omega_i^*$. We think of all physicians as being PFD, except when hard information like protocols forces them to behave like VBD. Thus, VBD-physician is more likely in a clinical situation where there is less ambiguity about medical guidelines and consensus treatments. Alternatively, this physician type is more common in a world were patients face high co-payments (or have no insurance at all) and she focuses on patient welfare. Whereas this paper focuses on supply side incentives (provider payment), the model accommodates demand side incentives by affecting whether the physician behaves like VBD or PFD. However, we do not consider the trade off between higher risk for insured due to co-payments and more efficient treatment decisions.

As shown in figure 2b: with VBD we have that $D_i(\omega_i^*) = 0$ and dis-utility is positive if (i) physicians cannot treat patients with $v_i(\omega) > c$ or (ii) physicians have to treat patients with $v_i(\omega) < c$.

At the other extreme, we have the ECD-physician for whom treating patients is always costly. She has a preference for avoiding effort which reflects a preference for leisure. This is the standard "homo economicus" from the literature on optimal regulation: the agent prefers leisure time over working (see, for instance, Laffont and Tirole, 1993). Based on intrinsic motivation, the ECD-physician prefers to treat no patients at all. Hence, financial incentives are needed to make sure that this physician treats anyone. The ECD-physician behaves in line with the optimal regulation literature. As we show in appendix D, this leads to under-treatment of patients. However, as explained in the introduction we are interested in problems due to over-treatment. Therefore, we focus on the PFD and VBD types. Although in reality the difference is one of degree, it facilitates the exposition to focus on the extremes.

3. One provider

This section adds financial incentives to physicians' utility function. We model the contractual arrangements between a physician/provider[11] and a payer. In our context, the payer can be a health insurer or the government. As there is only one payer, for concreteness we refer to this payer as the government.[12]

As a first step, we analyze the case where there is only one provider. This characterizes the NC

dis-utility function is the following. High quality physicians have made higher investments. For instance, they took extra training and/or invested in higher quality equipment. Then a sunk cost fallacy (Thaler, 1980) may induce them to treat more people compared to a l-physician as they invested more to treat people. In other words, if the cut-off level ω is increased, this is more costly for h than for l-physicians (equation (2)).

[11] In this paper we do not focus on the governance of hospitals, hence we use physician/provider/hospital interchangeably.

[12] In other words, we focus on provider competition and leave insurer competition for future research.

outcome. We consider two contractual environments: (i) complete information where quality is hard information contractible for the payer and (ii) asymmetric information where the provider knows her quality but the payer does not. As this is a fairly standard mechanism design problem, we go over it quite quickly. The relation with health economics concepts like fee-for-service and capitation is explained in Appendix C.

The payer offers the provider two payment options; one aimed at l-type and one at h-type provider. A payment option consists of a budget R_i and the number of patients that needs to be treated y_i. We assume that the number of patients treated is contractible (Chalkley and Malcomson, 1998). As physicians ration efficiently, there is a direct link between y_i and the cut-off: $y_i = 1 - \omega_i$.

We write the contracts offered by the payer as $(\omega_h, R_h), (\omega_l, R_l)$. Hence, the provider needs to treat $1 - \omega_i$ patients after accepting the i-contract. In case of asymmetric information, we allow for the possibility that a physician chooses a contract that is not intended for her type. The payer designing the contracts needs to take the following individual rationality (IR) and –if quality is not contractible– incentive compatibility (IC) constraints into account:

$$R_l - c(1-\omega_l) - D_l(\omega_l) \geq 0 \qquad (IR_l)$$

$$R_h - c(1-\omega_h) - D_h(\omega_h) \geq 0 \qquad (IR_h)$$

$$R_l - c(1-\omega_l) - D_l(\omega_l) \geq R_h - c(1-\omega_h) - D_l(\omega_h) \qquad (IC_l)$$

$$R_h - c(1-\omega_h) - D_h(\omega_h) \geq R_l - c(1-\omega_l) - D_h(\omega_l) \qquad (IC_h)$$

The IR constraints imply that the l and h physicians prefer to treat patients rather than close down and get the outside payoff (normalized to 0). The (IC_l) constraint makes it incentive compatible for l-physician to choose the intended contract (ω_l, R_l), instead of choosing the option for h-provider. The last IC constraint makes it incentive compatible for h to choose (ω_h, R_h) instead of (ω_l, R_l).

We assume that the payer maximizes the benefit from treatments minus expenditure R and can put some weight $\beta \in [0, 1\rangle$ on provider payoffs. We follow here the regulation literature (Baron and Myerson, 1982; Laffont and Tirole, 1993). With probability $F(1-F)$, the payer faces a $l(h)$ physician. Expected welfare is given by

$$\begin{aligned} W = F &\left[\int_{\omega_l}^{1} v_l(y)dy - R_l + \beta(R_l - c(1-\omega_l) - D_l(\omega_l)) \right] \\ + (1-F) &\left[\int_{\omega_h}^{1} v_h(y)dy - R_h + \beta(R_h - c(1-\omega_h) - D_h(\omega_h)) \right] \end{aligned} \qquad (3)$$

The first term (integral) in square brackets reflects the benefit from treatments in case all patients with $\omega \geq \omega_i$ are treated. The second term is the payer's transfer to the physician. The third term reflects provider payoffs and the weight $\beta \in [0, 1\rangle$ that the payer attaches to this. The effect of $\beta < 1$ is that

the payer tries to reduce the transfers to the physician.[13] If $\beta = 0$, the payer maximizes consumer value (total benefits minus payments).

3.1. complete information

Here we assume that the payer has complete information in the sense that he knows the type h, l of the provider. That is, there are contractible quality indicators that allow the payer to separate high from low quality physicians. An example of this in practice is the Alternative Quality Contract (AQC). Such a contract pays provider organizations a budget that depends on predefined quality indicators (Song et al., 2012). In some cases good quality indicators are available but for other specialties this is not yet possible.

We denote the first best outcome as $\hat{R}_i^1, \hat{\omega}_i^1$ with $i = l, h$. First best is used here in the sense of the mechanism design literature: in this outcome the payer observes the quality of the physician. Hence the IC constraints can be ignored and we only work with the IR constraints. First-best outcomes are characterized as follows.

Definition 3 *First best cut-off level* $\hat{\omega}_i^1$ *is the solution to*

$$v_l(\hat{\omega}_l^1) = (c - D_l'(\hat{\omega}_l^1)) \tag{4}$$

$$v_h(\hat{\omega}_h^1) = (c - D_h'(\hat{\omega}_h^1)) \tag{5}$$

In first best, $\hat{\omega}_i^1$ is determined by the marginal patient where the value of treatment (left hand side) equals the marginal cost of treatment (right hand side). The marginal cost of treatment equals the monetary cost c plus the physician's dis-utility $(-D')$. Both these costs are reimbursed by the planner because of the IR constraint.

We observe here a standard result in the health economics literature that even with complete information the first best outcome may differ from treatment efficiency ω_i^* (see, for instance, Cutler, 2006). For the PFD-physician, the first best outcome moves from its intrinsic optimum towards treatment efficiency. Moving away from the intrinsic optimum is costly for the payer but this is compensated by the production of fewer cost inefficient treatments. However, there is still over-treatment (compared to treatment efficiency) because $D_i' > 0$.

For the VBD-physician, the three benchmarks –treatment efficiency, intrinsic optimum and first best– coincide because $D_i'(\omega_i^*) = 0$.

[13] If the payer would not mind paying substantial transfers to the provider ($\beta = 1$), he can implement the first best outcome (see, for instance, equation (22) in the appendix).

3.2. asymmetric information

Here we assume that the payer does not have information on physician quality. He only knows that the probability of a $l(h)$-physician is $F(1-F)$. This we call second best. The contracts have to be chosen such that a physician reveals her type; that is, the contracts have to satisfy the IC constraints.

The optimal contracts solve

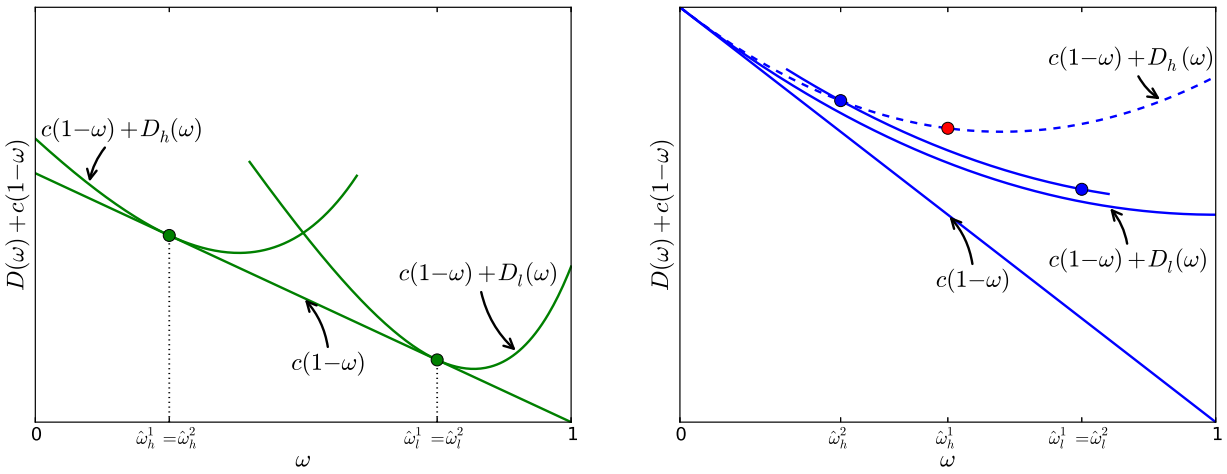
$$\max_{R_h,R_l,\omega_h,\omega_l} W \text{ subject to } (IR_l),\ (IR_h),\ (IC_l),\ (IC_h) \tag{6}$$

At first sight one might think that there are 16 cases to consider: each of the four constraints can bind or not. We prove the following result in the appendix.

Lemma 1 *To find all solutions to the planner's problem (6), we only need to consider the following three cases:*

1. *IR_h and IR_l are binding,*
2. *IR_h and IC_l are binding and*
3. *IR_l and IC_h are binding*

The solution to (6) is denoted by $\hat{R}_i^2, \hat{\omega}_i^2$. It turns out that the first case corresponds to VBD, the second to PFD and the third to ECD. The first two cases are analyzed below. For the ECD case, see Appendix D.



(a) VBD (b) PFD

Figure 3: Second best outcomes in the VBD and PFD case.

Proposition 1 *For a VBD-physician:*

$$\hat{\omega}_i^2 = \hat{\omega}_i^1 = \tilde{\omega}_i = \omega_i^* \text{ and } R_i = c(1-\omega_i^*) \text{ for } i = l, h$$

For a PFD-physician:

$$0 = \tilde{\omega}_l < \hat{\omega}_l^2 = \hat{\omega}_l^1 < \omega_l^* \text{ and } R_l = c(1-\hat{\omega}_l^1) + D_l(\hat{\omega}_l^1) + [D_h(\hat{\omega}_h^2) - D_l(\hat{\omega}_h^2)]$$
$$0 = \tilde{\omega}_h < \hat{\omega}_h^2 < \hat{\omega}_h^1 < \omega_h^* \text{ and } R_h = c(1-\hat{\omega}_h^2) + D_h(\hat{\omega}_h^2)$$

The VBD case is illustrated in figure 3a, which has patient severity ω on the horizontal axis and provider's (treatment plus dis-utility) cost on the vertical axis. A VBD physician (l and h type) in the first best outcome is on her IR constraint and cannot profitably mimic the other physician. So IC constraints are not binding. Further, physicians are only reimbursed for their treatment cost. As they implement treatment efficiency they have no dis-utility ($D_i(\omega_i^*) = 0$) and they do not receive an information rent: $R_i^* = c(1-\omega_i^*)$.

The PFD h-physician gets her overall (treatment and dis-utility) costs reimbursed. As the h-physician treats more patients than the l-physician and l would like to over-treat, she has an incentive to mimic the h-physician. To prevent l from mimicking h, l receives an information rent. The payer distorts $\hat{\omega}_h^2$ below first best $\hat{\omega}_h^1$ to reduce the information rent. The more patients h treats, the smaller the difference in costs between h and l and the smaller the rent. Hence, the h-physician over-treats not only compared to efficiency but even compared to first best.

To see the intuition for this, let's consider two possible alternative contracts: (i) payer does not take quality into account and offers one contract (ω, R) for both the l and h-physician and (ii) payer implements first best $\hat{\omega}_l^1, \hat{\omega}_h^1$.

With figure 3a we explain why offering only one contract reduces welfare. If both providers receive the same contract (ω, R), ω will tend to be somewhere in between ω_h^* and ω_l^*. That is, ω is set such that l-physicians treat too many, and h-physicians too few patients. Moreover, each type needs to be compensated for her dis-utility of not being able to treat at her intrinsic optimum $\tilde{\omega}_i = \omega_i^*$. Hence, offering only one contract (ω, R) is not optimal in the VBD case. Offering only one contract (ω, R) in the PFD case is not optimal either. As illustrated in figure 3b, the budget R needs to be high enough for the h-type to accept the contract. However, if the physician turns out to be an l-type then she will make a big profit on this contract. This outcome can be improved by differentiating the contracts.

Considering alternative (ii), figure 3b explains why implementing first best is not optimal in the PFD case. Suppose we would implement the first best outcome while keeping each type on her IR constraint. Then as shown in the figure, the h-type's contract lies above the l-type's indifference curve (corresponding to her IR constraint). Hence, the l-type will mimic the h-type and treat patients at

$\omega = \hat{\omega}_h^1$. This violates her IC-constraint. To prevent mimicking, the planner needs to leave the l-type a considerable rent. For $\beta < 1$ this rent is costly and the planner would like to reduce it. As can be seen in the figure, by reducing $\hat{\omega}_h^2$ below $\hat{\omega}_h^1$, mimicking the h-type becomes less attractive for the l-type. This allows the planner to reduce the rent paid to the l-type. Starting from first best, the loss in efficiency due to reducing ω_h is a second order loss while the reduction in rents is a first order gain (for $\beta < 1$).

Summarizing, we show that, with one provider, treatment efficiency is attainable only in a rather exceptional case. Only, if protocols can force physicians to take both the benefits and costs of treatments fully into account, it is optimal to implement treatment efficiency. If physicians focus on treatment effectiveness, there will be over-treatment (compared to treatment efficiency) in both the first and second best outcomes. This contractual inefficiency is the starting point to compare three different ways in which health care can be organized (NC, CfM, CoM).

4. No provider choice

Now assume that there are two providers. This section considers two situations where patients cannot choose their provider: NC and CfM. The main results are the following: (i) with first best contracting opportunities, CfM leads to higher welfare than NC, (ii) with second best contracting, NC (CfM) leads to higher welfare if payer's and patients' preferences are (not) aligned.

Assume that the size of the market is 2 and that there are two providers P_1, P_2. We consider two cases: either both providers are VBD or both are PFD. Hence we do not analyze the situation where a VBD physician faces a PFD physician. As we explained in section 2, we think of VBD and PFD as driven by the clinical practice of the physician. For a given condition, this is the same for each physician.

4.1. NC

With NC, payer contracts both providers and determines which provider a particular patient has to visit. For concreteness, we think here of geographical differentiation between providers. Say, each province or each county has one hospital. People living in this geographical area visit the local hospital. Hence a provider faces no competition.

Other forms of differentiation can be kept in mind as well. The payer may send elderly patients to one hospital and young patients (with the same condition) to another hospital. If a patient needs a routine operation, he is treated in a local hospital. The same operation with complications (say,

co-morbidities) is performed in an academic hospital.

With NC, we have the same situation as in section 3 but then "multiplied by two" as there are two providers and the size of the market is two.

4.1.1. first best contracting opportunities

Recall that under first best contracting, either cost efficient treatment protocols can be enforced (VBD) or quality indicators make quality contractible for the payer (complete information as in section 3.1).

To simplify the expression for welfare, we define

$$V_i(\omega) = \int_{\omega}^{1} v_i(y)dy - (c(1-\omega) + D_i(\omega)) \tag{7}$$

Using the result from section 3.1, expected welfare under NC can be written as:

$$W^{NC} = F^2 2V_l(\hat{\omega}_l^1) + (1-F)^2 2V_h(\hat{\omega}_h^1) + 2F(1-F)(V_h(\hat{\omega}_h^1) + V_l(\hat{\omega}_l^1)) \tag{8}$$

This expression is correct in both the PFD and VBD case.

With asymmetric information and VBD, we find the following. Proposition 1 implies that with VBD welfare W^{NC} is given by equation (8).

4.1.2. second best contracting opportunities

Second best contracting is characterized by PFD and asymmetric information. Proposition 1 implies

$$\begin{aligned} W^{NC} =& 2F^2(V_l(\hat{\omega}_l^1) - (1-\beta)(D_h(\hat{\omega}_h^2) - D_l(\hat{\omega}_h^2))) + 2(1-F)^2 V_h(\hat{\omega}_h^2) + \\ & 2F(1-F)(V_h(\hat{\omega}_h^2) + V_l(\hat{\omega}_l^1) - (1-\beta)(D_h(\hat{\omega}_h^2) - D_l(\hat{\omega}_h^2))) \end{aligned} \tag{9}$$

4.2. CfM

With CfM, the payer contracts only one provider to treat patients. The other provider is closed down (or serves another payer; not modelled here). Hence, patients cannot choose their provider. As with NC, the payer chooses a provider for them. That is, if patients want their treatment costs reimbursed, they have to visit the contracted physician. We assume that all patients prefer to visit the contracted physician instead of paying a provider out of pocket.

Compared to NC, CfM introduces competition between providers: only one of them will be chosen to treat the payer's patients.

4.2.1. first best contracting opportunities

Under complete information, the payer observes quality and only IR constraints are binding. The payer randomly chooses one physician if both have the same quality and contracts physician with the

highest surplus $V_i(\hat{\omega}_i^1)$ in the mixed case. We can write welfare under CfM as follows.

$$W^{CfM} = 2F^2 V_l(\hat{\omega}_l^1) + 2(1-F)^2 V_h(\hat{\omega}_h^1) + 4F(1-F)\max\{V_l(\hat{\omega}_l^1), V_h(\hat{\omega}_h^1)\} \tag{10}$$

If good quality indicators are available such that quality becomes contractible for the payer, CfM leads to higher welfare than NC. Using the quality information, the payer contracts with the provider yielding the highest surplus. With VBD, h yields the highest surplus. Since $D_h(\omega) > D_l(\omega)$, this is not necessarily the case under PFD.

Comparing (10) to (8) shows the *mis-allocation effect* of NC: half the market gets surplus $\min\{V_l(\hat{\omega}_l^1), V_h(\hat{\omega}_h^1)\}$.

With asymmetric information, CfM should be organized such that providers truthfully reveal their types. In the VBD case, this is straightforward. In the hh case, the payer implements $\hat{\omega}_h^1$, randomly chooses a provider and pays this provider $R_h = 2c(1-\hat{\omega}_h^1)$; the other provider gets nothing. In the ll case, payer implements $\hat{\omega}_l^1$, randomly chooses one provider who receives $R_l = 2c(1-\hat{\omega}_l^1)$. In the lh case, payer chooses physician with highest surplus $V_h(\hat{\omega}_h^1)$, implements $\hat{\omega}_h^1$ and pays $R_m^h = 2c(1-\hat{\omega}_h^1)$ (the l-physician gets nothing). It is straightforward to verify that this outcome is incentive compatible. W^{CfM} can be written as (10). Hence in the VBD case, CfM yields higher welfare than NC.

Corollary 1 *With first best contracting, CfM implements first best and* $W^{CfM} \geq W^{NC}$. *This inequality is strict if* $V_l(\hat{\omega}_l^1) \neq V_h(\hat{\omega}_h^1)$.

With first best contracting opportunities, CfM implements first best and thus leads to higher welfare than NC (and higher than CoM, as we see below).

4.2.2. second best contracting opportunities

In the PFD case with asymmetric information, we know that l-physicians tend to mimic h-physicians due to their tendency to over-treat. Hence IR constraint is binding for h-type and IC constraint for l-type. The payer needs to choose which physician wins in the mixed case. The following lemma shows that in any incentive compatible CfM outcome, the l-physician is contracted in the mixed case.

Lemma 2 *Under CfM with PFD, there is no incentive compatible outcome where the* h*-type is contracted in the* lh*-case.*

This result is interesting in the light of the negative press that selective contracting sometimes gets. People worry that a payer is *biased against high quality providers* due to costs concerns. Indeed, we show that with CfM (and second best contracting) a payer can only contract with the low quality provider in the lh-case. The intuition is the following. With CfM the excluded provider treats none

of the payer's patients leading to the same (outside) payoff to h and l physicians. Further, l has lower costs than h; hence l can always mimic h but not the other way around. This forces the payer to contract l in the lh-case.

The welfare consequences can be summarized as follows.

Proposition 2 *With asymmetric information and PFD:*

$$W^{CfM} = 2F^2 V_l(\hat{\omega}_l^1) + 2(1-F)^2 V_h(\hat{\omega}_h^2) + 2F(1-F)(2V_l(\hat{\omega}_l^1) - (1-\beta)(D_h(\hat{\omega}_h^2) - D_l(\hat{\omega}_h^2))) \quad (11)$$

Hence $W^{CfM} > W^{NC}$ *if and only if*

$$V_l(\hat{\omega}_l^1) > V_h(\hat{\omega}_h^2) - (1-\beta)\frac{F}{1-F}(D_h(\hat{\omega}_h^2) - D_l(\hat{\omega}_2^h)) \quad (12)$$

Equation (12) shows that CfM leads to higher welfare than NC if and only if the l-physician yields a higher surplus for the payer than the h-physician. The l-physician yields the payer a surplus equal to the left hand side of the equation. The h-physician yields surplus $V_h(\hat{\omega}_h^2)$ minus l's information rent that is associated with the h-physician's contract. Since $v_h(\omega) > v_l(\omega)$, patients' preference is always for the h-physician. Hence, proposition 2 says that welfare under CfM exceeds welfare under NC if and only if the patients' and payer's *preferences are not aligned*. If these preferences are aligned, NC yields higher welfare because CfM is biased towards the l-provider.

5. Provider choice

In this section, the payer contracts both providers and patients are free to choose. This we call competition on the market (CoM). Argument in favor of patient choice is that choice tends to improve market functioning. But in health care the consumer may not be well equipped to choose the best provider (McGuire, 2011b; Pope, 2009). The main results are as follows. With first best contracting, CfM leads to highest welfare. With second best contracting, CoM leads to higher welfare than CfM if payer's and patients' incentives are aligned and decentralized quality information induces patients to visit the best provider.

Allowing for provider choice implies that we need to be specific about a number of things that were swept under the rug before. We allow for the case where patients know more about provider quality than the planner. That is, patients can tap into decentralized information that is not available/not contractible at the centralized payer level. Intuitively, you hear from neighbours and friends how they were treated by the physicians and based on this information you decide which physician to visit yourself.

With CoM, patients are free to travel to the provider of their choice. As mentioned above, for concreteness think of geographical differentiation of providers. We assume that there is a group of $2-2x$ patients that are located in between the two providers (being indifferent to travel either way if both providers have the same quality). This group always visits the higher quality physician if quality levels differ (as $v_h(\omega) > v_l(\omega)$). In addition, there are two groups each of x patients that prefer to go to the closest physician (even if she has lower quality than the other).[14] They only prefer to be treated by the far-away provider if their close-by provider does not want to treat them. If both providers have the same quality, patients split 1:1. If providers differ in quality then we assume patients split $x : 2-x$ with $x \leq 1$ patients (initially) going to the low quality provider.

We call x a measure of patients' responsiveness to quality (differences). When $x \leq 1$ is low (high), patients are (not) responsive to quality differences. That is, with high x, patients see the quality difference but are not willing to travel to the highest quality provider.[15]

Here we compare welfare levels W^{CoM}, W^{NC} and W^{CfM} while ignoring the welfare loss due to travel costs (or other (switching) costs caused by patients visiting a provider which is not their first choice). We come back to these travel costs in section 6.

Although $2x$ patients are not willing to travel in response to a quality difference, we assume that the value of treatment (at any provider) is higher than the travel cost. Hence, patients are willing to travel to make sure that they get treatment at all. If t denotes the cost to travel to provider 1 for the x patients close to 2, we assume that $v_l > t > v_h - v_l$. We assume that a patient does not know his own ω, thus v_i denotes the expected value of treatment. Expected value of treatment by l-physician exceeds the travel cost but the additional value of being treated by h-physician does not exceed t for these patients.

From the assumption that patients are willing to travel to get treatment (at all) and physicians' efficient rationing, we have the following no-arbitrage result. This is true under both complete and asymmetric information.

Lemma 3 *With CoM, the threshold value ω set by the payer is the same for both providers. In the ll case, this threshold is denoted ω_l, in the hh case ω_h and in the mixed lh case ω_m.*

[14] Alternatively, they are already under treatment with this provider and face switching costs (like transferring medical files, building a relationship with the new physician) to go to another. Empirical evidence shows that consumers respond to hospital quality, but travel distance is a dominant determinant of hospital choice (Pope, 2009).

[15] Alternatively, x close to 1 can also be interpreted as a situation where patients hardly observe quality differences. Hence, they do not respond to such differences. On the other hand, x close to zero implies that people can figure out quality differences very well and are willing to travel based on this information. Clearly, x will differ per treatment. Many people are willing to travel to a better hospital for brain or eye surgery but fewer people would be willing to travel far for a couple of stitches.

It is natural for the payer to use the same threshold for both providers in the hh and ll cases. The same threshold ω_m in the mixed case is not so intuitive at first sight.

To see why it is correct, consider the case where the payer tries to implement $\omega_h < \omega_l$ in the mixed case: the h-physician can treat more cases ω than the l-type. If $x > 0$, there are patients close to the l-provider who go there but are not treated because their $\omega < \omega_l$. Since these people prefer to be treated (rather than not being treated at all), they travel to the h-provider. If their $\omega < \omega_h$, they are not treated by the h-physician either. However, types $\omega \in \langle \omega_h, \omega_l \rangle$ "crowd out" types ω' with $\omega_h < \omega' < \omega$ since physicians ration efficiently. If the h-physician is supposed to treat y_h patients, she ends up treating patients close to the l-provider who are not treated by their "home"-physician. This arbitrage behaviour by patients makes sure that all patients with $\omega \geq \omega_m$ are treated independently from where they live. The payer –understanding this– implements the same threshold ω_m for both providers in the mixed case.

Here we see an important feature of CoM: it tends to lead to over-treatment by low quality physicians. Indeed, as we will see below, ω_m is below the optimal treatment threshold ω_l for the l-type. This brings us to the following disadvantage of provider choice (both under complete and asymmetric information).

Corollary 2 *With CoM and $x > 0$, it is impossible to implement the thresholds of section 3 in the mixed case when there is one h and one l-physician.*

This is a negative efficiency effect of CoM caused by patients *shopping around* for a physician who is willing to treat them. In the hl-case, the l-provider tends to over-treat patients while the h-provider tends to under-treat (compared to section 3) both in first and second best. Hence, patients shopping around make it harder for the payer to control over-treatment by the l-provider. With CfM, there is no shopping around as only one provider is active. Under NC, it is possible to implement different thresholds for different physicians because a physician is not allowed to treat patients that "belong" to the other physician. To illustrate, under NC with VBD, the payer can implement treatment efficiency. This is not possible with CoM.

5.1. first best contracting opportunities

Under complete information, we focus on the IR constraints. Let $C_h(\omega), C_l(\omega), C_m^h(\omega), C_m^l(\omega)$ denote the costs for a h provider in the hh-case, l provider in the ll-case, h provider in the mixed case and l

provider in the mixed case resp. when the planner varies the threshold ω. Then we have the following.

$$\begin{aligned} C_h(\omega) &= c(1-\omega) + D_h(\omega) \\ C_l(\omega) &= c(1-\omega) + D_l(\omega) \\ C_m^h(\omega) &= (2-x)(c(1-\omega) + D_h(\omega)) \\ C_m^l(\omega) &= x(c(1-\omega) + D_l(\omega)) \end{aligned} \tag{13}$$

Hence, the IR constraints take the form $R_i = C_i(\omega_i)$ and $R_m^i = C_m^i(\omega_m)$.

Since corollary 1 implies that CfM implements first best with complete information, it is clear that CoM cannot raise welfare compared to CfM. In fact, $W^{CfM} > W^{CoM}$ in this case due to corollary 2, unless $x = 0$ and all patients travel to the provider with the highest social surplus. In the latter case, we have $W^{CfM} = W^{Com}$ as under CoM the l-physician is de facto closed down in the mixed case. Proposition 4 in the appendix proves this formally and compares welfare under CoM and NC in this case (both below W^{CfM}). If patients react to decentralized information such that most patients visit the provider with highest social surplus under CoM, CoM leads to higher welfare than NC: the inefficiency of patients shopping around (corollary 2) is small while welfare increases with the number of patients visiting the best provider. If, on the other hand, patients tend to go to the closest provider ($x = 1$), there is no efficiency gain from CoM compared to NC and the inefficiency of patients shopping around dominates.

With asymmetric information on quality but cost efficient treatment protocols (VBD), corollary 1 implies that CfM implements first best. Again, CoM cannot implement something better than first best and CfM leads to highest welfare W.

Hence with first best contracting opportunities, CfM is the optimal way to organize health care.

5.2. second best contracting opportunities

To analyze the effects of CoM with asymmetric information and PFD,[16] we have 4 IR and 4 IC constraints. The IC constraints here are more involved than in the monopoly case in section 3. To see why, note that figure 3 in the monopoly case has the following simplifying feature. Consider an indifference curve of a high quality provider. By varying ω, we see the value of R that the payer needs to grant the provider to keep her indifferent for different values of ω_h. But for a given choice of ω_h by the payer, the indifference curve also indicates combinations (ω, R) that keep the h-provider

[16] In fact, if providers' know each other's quality one can design games in which truthful revelation is the unique equilibrium with zero information rent (see, for instance, Maskin, 1999; Palfrey and Srivastava, 1991). Moore (1999) gives an overview and discussion of the literature on implementation under complete information. However, in practice, it may not be obvious to implement such a mechanism and therefore we also consider the asymmetric information case.

indifferent when deviating to another combination (ω, R), say (ω_l, R_l). Hence the indifference curve is the same whether the payer considers different values for ω_h or the h-physician considers deviating to $(\omega, R) \neq (\omega_h, R_h)$. This is not true in the case of provider choice. This difference determines whether provider choice makes it easier or harder to satisfy the IC constraints.

Formally, the cost functions in equation (13) are not the relevant ones when a provider decides to deviate and mimic a different type than she actually is. To illustrate, such a deviation may trigger different (out-of-equilibrium) thresholds for competing physicians. Consequently, patients travel to find a physician who is willing to treat them. Such travelling leads to an increase in the number of patients that a physician needs to turn down thereby affecting her dis-utility. Lemma 8 in appendix B derives the cost functions for a deviating physician from first principles for $x \in [0,1]$. For physician i who faces a competitor j, we denote this deviation cost function by $\tilde{C}_{ij}(\omega)$. We specify the relevant functions $\tilde{C}_{ij}$ when we need them below.

Given the cost functions $C_{ij}(\omega)$ and $\tilde{C}_{ij}(\omega)$, the planner's problem can be written as

$$\begin{aligned}\max_{R_i^j, \omega_j} & 2F^2 \left(\int_{\omega_l}^1 v_l(\omega) d\omega - R_l + \beta(R_l - C_l(\omega_l)) \right) \\ & + 2(1-F)^2 \left(\int_{\omega_h}^1 v_h(\omega) d\omega - R_h + \beta(R_h - C_h(\omega_h)) \right) \\ & + 2F(1-F) \left(x \int_{\omega_m}^1 v_l(\omega) d\omega + (2-x) \int_{\omega_m}^1 v_h(\omega) d\omega - (R_m^l + R_m^h) + \beta(R_m^l - C_m^l(\omega_m) + R_m^h - C_m^h(\omega_m)) \right) \end{aligned} \tag{14}$$

subject to the following IR constraints

$$R_h - C_h(\omega_h) \geq 0 \qquad (IR_h)$$

$$R_l - C_l(\omega_l) \geq 0 \qquad (IR_l)$$

$$R_m^h - C_m^h(\omega_m) \geq 0 \qquad (IR_m^h)$$

$$R_m^l - C_m^l(\omega_m) \geq 0 \qquad (IR_m^l)$$

and the following IC constraints

$$R_h - C_h(\omega_h) \geq R_m^l - \tilde{C}_{hh}(\omega_m) \qquad (IC_h)$$

$$R_l - C_l(\omega_l) \geq R_m^h - \tilde{C}_{ll}(\omega_m) \qquad (IC_l)$$

$$R_m^h - C_m^h(\omega_m) \geq R_l - \tilde{C}_{hl}(\omega_l) \qquad (IC_m^h)$$

$$R_m^l - C_m^l(\omega_m) \geq R_h - \tilde{C}_{lh}(\omega_h) \qquad (IC_m^l)$$

Analyzing this optimization problem with 8 constraints for general $x \in [0,1]$ is not informative. The number of possible combinations of binding constraints is too large. Hence, we consider two benchmark cases: $x = 1$ and $x = 0$.

We find the following result. Recall that payer's and patients' preferences are aligned if and only if inequality (12) does not hold.

Proposition 3 *Consider two PFD physicians with asymmetric information. Assume* $D_h(\hat{\omega}_h^1) \leq 2D_l(\hat{\omega}_h^1)$.

- *If payer's and patients' preferences are aligned, then* $x = 1$ *implies that* W^{NC} *is highest and* $x = 0$ *implies that* W^{CoM} *is highest.*

- *If payer's and patients' preferences are not aligned, then* $x = 1$ *implies that* W^{CfM} *is highest; if* $x = 0$ *we have* $W^{CoM} \geq W^{CfM}$ *if and only if*

$$\begin{aligned}(1-F)^2\left(V_h(\hat{\omega}_h^1) - [V_h(\hat{\omega}_h^2) - (1-\beta)\frac{F}{1-F}(D_h(\hat{\omega}_h^2) - D_l(\hat{\omega}_h^2))]\right) \geq \\ 2F(1-F)\left(V_l(\hat{\omega}_l^1) - [V_h(\hat{\omega}_h^2) - (1-\beta)\frac{F}{1-F}(D_h(\hat{\omega}_h^2) - D_l(\hat{\omega}_h^2))]\right)\end{aligned} \tag{15}$$

Although we assume that $D_h(\omega) > D_l(\omega)$, the proposition assumes that D_h is not twice as high as D_l evaluated at $\hat{\omega}_h^1$. As we see below, this assumption leads to higher welfare under CoM than NC in case $x = 0$ as it reduces the information rent.

First, consider the case where payer's and patients' preferences are aligned. Then we know from proposition 2 that $W^{NC} > W^{CfM}$. Hence the question is whether W^{CoM} exceeds W^{NC}. This depends on patients' responsiveness to decentralized quality information.

If every patient prefers to visit the closest physician instead of the best one ($x = 1$), each provider gets the required number of patients in a straightforward way. Patient streams do not affect information rents. Information rents under CoM are then at least as large as under NC.[17] There is no reduction in information rents due to CoM and there is the disadvantage of ω_m in the lh-case (lemma 3): patients shop around to get treatment.[18] Hence, in this case, we find $W^{NC} > W^{CoM}$.

If patients are responsive to quality ($x = 0$), the payer can reduce information rents because of patient streams. To illustrate, consider the lh-case. If the l-physician would claim to be h, the planner would implement $\hat{\omega}_h^1$ and each physician serves half the market. But patients observe the

[17] That is, with $x = 1$, the cost functions are the same under CoM and NC. This follows from equation (13) and lemma 8 in the appendix: costs of a physician i who faces competitor j are given by

$$C_{ij}(\omega) = \tilde{C}_{ij}(\omega) = C_i(\omega) \tag{16}$$

where C_i denotes the cost function of a monopolist physician as analyzed in section 3.

[18] In fact, due to ω_m information rents increase. Intuitively, ω_m is somewhere in between the optimal ω_h and ω_l (lemma 4 in the appendix). As explained in figure 3b, the way the payer reduces rents in the monopoly case is to reduce ω_h. The lower ω_h, the less attractive it becomes for an l-type to mimic a h-type. But with $\omega_m \geq \omega_h$, it is more attractive to deviate under CoM than under monopoly. Hence higher rents are needed to prevent such a deviation.

quality difference and all patients initially go to the h-physician who then treats the highest ω's. The mimicking l-physician treats the lower ω's and hence has a lower threshold than the h-physician. This implies that everyone who is not treated by the h-physician visits the l-physician who then gets dis-utility $2D_l$.[19] Under the assumption in the proposition, this double dis-utility makes mimicking unattractive and the planner can implement $\hat{\omega}_h^1$ in case hh. This is not possible under NC where $\hat{\omega}_h^2$ is implemented in hh-case.

Next consider the case where payer's and patients' preferences are not aligned. With $x = 1$, information rents are at least as high under CoM as under NC and CoM has the inefficiency of ω_m (lemma 3). Hence the relevant comparison here is between NC and CfM. Proposition 2 implies $W^{CfM} > W^{NC}$ and W^{CfM} is highest.

Finally, consider $x = 0$ with preferences that are not aligned. Proposition 2 implies $W^{CfM} > W^{NC}$ (as this does not depend on x). Hence, the relevant comparison is between W^{CoM} and W^{CfM}. Equation (15) is necessary and sufficient for CoM to be welfare maximizing. As explained above, CoM allows for the implementation of $\hat{\omega}_h^1$ in case hh without information rent to be paid in case lh. CfM implements $\hat{\omega}_h^2$ in case hh and pays an information rent in case lh. Hence the left hand side of (15) gives the advantage of CoM in case hh. In contrast, the right hand side gives the advantage of CfM. Because inequality (12) holds, the right hand side is positive. In case lh, CfM implements $\hat{\omega}_l^1$ and pays no information rent in case ll. CoM implements $\hat{\omega}_h^2$ and pays an information rent in case ll. For F small enough, W^{CoM} is highest.

Summarizing, with first best contracting opportunities CfM is the best way to organize health care. With second best contracting and patients' and payer's preferences not aligned, CfM can still be optimal. In this case, CfM can be used to steer patients to the l-provider. If preferences are aligned and patients react strongly to decentralized quality information, CoM is optimal. If, instead, patients prefer to visit the closest provider independent of quality, NC is optimal.

6. Travel costs and cost differences

Up till now, when considering welfare we have ignored travel costs by patients who cannot visit their preferred provider. The implicit assumption has been that the payer deems travel costs to be unimportant. As patients themselves do react to travel costs (with $x > 0$), some form of paternalism plays

[19] As shown in lemma 8 in appendix B, the cost function for a deviating l-physician in lh-case with $x = 0$ is given by

$$\tilde{C}_{lh}(\omega_h) = c(1 - \omega_h) + 2D_l(\omega_h) \tag{17}$$

a role on the payer's side.

This section considers the case where the payer views travel costs as being significant. For concreteness we assume that travel costs take the form of geographical travel cost. This implies that both CoM and NC avoid the travel cost for the group of x patients living close to a provider. However, these people are forced to travel under CfM. Let $t > 0$ denote the travel costs that these people incur when they are forced to travel. The payer's welfare function now takes t into account. How does this affect the results on the optimal form of provider competition?

First, consider VBD. Then the difference in social value between two providers is driven only by their difference in quality. A patient then makes the socially optimal trade off: t versus the difference in quality $v_h - v_l$. Hence CfM is not optimal in this case (whereas it is optimal in this case without travel costs). NC and CoM avoid the travel costs. As shown in proposition 4 in the appendix, CoM (NC) dominates if patients (do not) react to decentralized information on quality differences.

With PFD, this reasoning is no longer complete. To illustrate, with second best contracting there is an information rent that needs to be paid to providers. Fully insured patients do not take this rent into account. Patients just compare $v_h - v_l$ with their travel cost, ignoring the information rent. If the information rent reduction due to CfM (compared to NC and CoM) exceeds the travel cost t, it is optimal for the payer to force the x patients to travel to another physician. This is in line with the intuition in proposition 3: CfM is optimal if payer's and patients' preferences are not aligned.

Another reason why payer's and patients' incentives are not aligned occurs when treatment costs c differ between h and l-providers ($c_h \neq c_l$). This works in the direction of CfM becoming the optimal way to organize health care markets.

We consider two cases. First, assume $c_h < c_l$. This happens when some providers are better organized than others: they are able to offer higher quality at lower costs. Alternatively, some treatments require learning by doing. If one provider has more experience than the other, she can do the treatment better and at lower costs (e.g. because there is a lower probability that a patient needs to return to hospital to be treated again). In this case, the difference in social surplus between h and l providers equals $v_h - c_h - [v_l - c_l] > v_h - v_l$. But patients compare $v_h - v_l$ to their travel costs. Hence, CfM is optimal if $v_h - c_h - [v_l - c_l] > t > v_h - v_l$.

Second, consider $c_h > c_l$. This is the case where higher quality goes hand in hand with higher costs. Again, CfM becomes welfare maximizing when $v_h - c_h < v_l - c_l$: payer's and patients' preferences are not aligned. Here one can think of the medical arms race where providers tend to buy the latest technologies just to attract patients; even if such new technologies are not socially optimal. Problem is caused here by patients being attracted by new technology while they do not take the (full) cost into

account. CfM does two things in this situation. First, the x patients close to the h-provider are forced to visit the l-physician who yields higher social surplus. Second the $2-2x$ "mobile" patients visit the l instead of the h-provider. Both effects tend to raise welfare.

Note that CfM is done here by a payer maximizing total welfare (W and taking travel costs into account). This is different from competing insurers deciding to use selective contracting. Boone and Schottmueller (Boone and Schottmueller) analyze whether competing insurers implement selective contracting if and only if it maximizes total welfare.

Finally, CfM and NC have an advantage (compared to provider choice) that is often relevant in practice. Both allow the payer to risk rate the budget. In the model above, patients are uniformly and symmetrically distributed in terms of severity ω. But in practice, one provider may face patients nearby that are healthier than the nearby patients of her competitor. This implies that for given physician quality, one provider should treat more patients than the other and hence receive a larger budget. With CfM and NC, each provider has a well defined population "belonging to this provider". Hence this population can be risk rated and the budget can be adjusted accordingly. With CoM, payer does not know ex ante which patients visit which provider. Hence risk rating the budget becomes harder.

7. Policy implications

When deciding on provider competition in the health care market, the first lesson is that "one size fits all" will not work. The contracting opportunities available to the payer determine the optimal way to organize health care. These opportunities depend on the clinical situation, which differs between specialties. Hence, the optimal organization differs between specialties.

To illustrate, first best contracting opportunities exist if provider quality information is available. Examples of contracts that go in this direction include budgets that incorporate provider quality (Song et al., 2012) and pay for performance contracts (Werner et al., 2011). Protocols that limit physician discretion also create first best contracting opportunities. Examples of limited discretion include hip fracture repair, cancer and dialysis treatments (Clemens and Gottlieb, 2014, pp. 1336).

With first best contracting opportunities, CfM is optimal if the payer considers travel costs for patients to be small. If travel costs are significant in this case, either NC or CoM is optimal depending on whether decentralized information helps patients to visit the better provider. With second best contracting opportunities, CfM tends to be optimal if payer's and patients' preferences are not aligned. If payer's and patients' preferences are aligned, allowing patients to choose their own provider is optimal

if patients are responsive to quality. Then patient streams help the payer to infer quality differences between providers. There are examples where provider choice has led patients towards high quality providers (Pope, 2009).

The disadvantage of CoM is that patients shop around to find a provider who is willing to treat them. This leads to over-treatment by low quality providers. If patients are not responsive to quality differences (e.g. they tend to visit the closest provider), this disadvantage is not compensated by lower information rents. Then NC is optimal. Although NC is extreme, more generally, segmenting the market into regions, and mandating that patient visit providers in the region weakens provider competition and reduces the disadvantage of shopping around. CoM would then allow a patient to visit any provider in the country.

To illustrate how our framework can be applied in practice, we consider some cases. First, emergency care. In this case, distance to the hospital is most important ($x = 1$: every patient wants to be taken to closest provider) and travel costs are significant. Hence NC is optimal.

Second, consider primary care. Again travel costs are important with a primary care physician; this rules out CfM. If the physician also acts as gatekeeper, shopping around by patients may lead to over-utilization and is not desirable: NC may again be optimal.

Third, obstetrics; with childbirth it is important that a woman fully trusts her gynecologist or midwife. This can differ for each woman and is not necessarily captured by quality indicators. Travel costs, in terms of horizontal differentiation, are important here and payer and patients' preferences are likely to be aligned: CoM is optimal. Each woman chooses the provider she prefers.

Finally, consider treatments where protocols cannot rule out inefficient use of new technologies. A recent example here could be proton beam therapy to treat cancer. Some claim that the additional benefits of proton beam therapy, are small compared to its costs.[20] However, insured patients tend to prefer the latest technology even if it is only marginally better (at a high cost). As patients' and payer's preferences are not aligned, CfM is optimal in this case.

We have derived the optimal payment contract in each case (NC, CfM and CoM). Although implementing the optimal payment scheme in each case is non-trivial, the model does indicate that some policies observed in practice may not be optimal. To illustrate, a recent Dutch policy change imposes a cap (macro or global budget) on the total Dutch hospital expenditure, without any reference to value or quality (see e.g. Schut et al., 2013). In terms of our model this boils down to $R_1 + R_2 = constant$ for some positive constant. This is clearly not optimal. There is no reason why the sum of transfers to providers, $R_1 + R_2$, should be constant over the different states ll, lh and hh. In fact, more patients

[20] http://www.medscape.com/viewarticle/778466A

should be treated and more money should be spent with h than with l-physicians.[21] Moreover, if in the Dutch case total spending (for all providers together) exceeds the macro budget with an amount y, a provider with market share x has to pay the government $x*y$ as a fine for going over the macro budget. This implies that providers with a high market share are punished more than providers with a low market share. In other words, h-providers are punished more than l-providers; reducing h-providers' incentives to treat patients. As illustrated in figure 3b, the optimal response to over-treatment is actually to let the h provider treat more patients than in first best (not less).

We introduced a framework to analyze different competition modes in the face of contractual problems leading to over-utilization in health care. We showed that improving the contracting environment tends to raise welfare. Instruments to do this include developing provider quality indicators (reducing asymmetric information), increasing co-payments, introducing medical guidelines and protocols and teaching medical students to think more in terms of costs and benefits instead of focusing on treatment effectiveness (i.e. pushing toward a VBD environment). If travel costs are not significant, such reforms tend to favor CfM.

[21] Mougeot and Naegelen (2005) also find –in a different model– that a global budget cap is not optimal.

References

Arrow, K. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review 53*(5), 941–973.

Baicker, K., K. S. Buckles, and A. Chandra (2006). Geographic variation in the appropriate use of cesarean delivery. *Health Affairs 25*(5), w355–w367.

Baron, D. P. and R. B. Myerson (1982). Regulating a monopolist with unknown costs. *Econometrica*, 911–930.

Boone, J. and C. Schottmueller. Health provider networks, quality and costs. Mimeo.

Chalkley, M. and J. Malcomson (1998). Contracting for health services with unmonitored quality. *The Economic Journal 108*, 1093–1110.

Chalkley, M. and J. Malcomson (2000). *Handbook of Health Economics (Volume 1A)*, Chapter Government Purchasing of Health Services, pp. 847–890. Elsevier.

Chandra, A., D. Cutler, and Z. Song (2011). Chapter six - who ordered that? the economics of treatment choices in medical care. In T. G. M. Mark V. Pauly and P. P. Barros (Eds.), *Handbook of Health Economics (Volume 2)*, Volume 2 of *Handbook of Health Economics*, pp. 397 – 432. Elsevier.

Chandra, A. and J. Skinner (2012). Technology growth and expenditure growth in health care. *Journal of Economic Literature 50(3)*, 645–680.

Chandra, A. and Staiger (2007). Productivity spillovers in healthcare:evidence from the treatment in heart attacks. *Journal of Political Economy 115(1)*, 103–140.

Clemens, J. and J. D. Gottlieb (2014). Do physicians' financial incentives affect medical treatment and patient health? *American Economic Review 104*(4), 1320–49.

Cutler, D. (2006). The economics of health system payment. *De Economist 154*, 1–18.

Dranove, D., M. Shanley, and W. White (1993). Price and concentration in hospital markets: the switch from patient-driven to payer-driven competition. *Journal of Law and Economics 36*(1), 179–204.

Evans, R. (1974). Supplier-induced demand: Some empirical evidence and implications. In M. Perlman (Ed.), *The Economics of Health and Medical Care*, The Economics of Health and Medical Care, pp. 162–173. MacMillan.

Gaynor, M., D. Haas-Wilson, and W. Vogt (2000). Are invisible hands good hands? moral hazard, competition, and the second-best in health care markets. *Journal of Political Economy 108*(5), 992–1005.

Getzen, T. (2007). *Health Economics and Financing*. Wiley.

Giuffrida, A. and H. Gravelle (2001). Inducing or restraining demand: the market for night visits in primary care. *Journal of Health Economics 20*(5), 755 – 779.

Gowrisankaran, G. (2008). *Incentives and Choice in Health Care*, Chapter Competition, Information Provision, and Hospital Quality, pp. 319–352. The MIT press.

Laffont, J. and J. Tirole (1993). *A theory of incentives in procurement and regulation*. MIT Press.

Maskin, E. (1999). Nash equilibrium and welfare optimality. *The Review of Economic Studies 66*(1), pp. 23–38.

McClellan, M. (2011). Reforming payments in healthcare:providers. *Journal of Economic Perspectives 25(2)*, 69–92.

McClellan, M., A. N. McKethan, J. L. Lewis, J. Roski, and E. S. Fisher (2010, May). A national strategy to put accountable care into practice. *Health Affairs 29*(5), 982–990.

McGuire, T. (2000). *Physician Agency*, Chapter 9, pp. 461–160. Elsevier.

McGuire, T. (2011a). Chapter 25 - physician agency and payment for primary medical care. In S. Glied and P. Smith (Eds.), *Oxford Handbook of Health Economics*, Oxford Handbook of Health Economics, pp. 602 – 623. Oxford University Press.

McGuire, T. G. (2011b). Chapter five - demand for health insurance. In T. G. M. Mark V. Pauly and P. P. Barros (Eds.), *Handbook of Health Economics*, Volume 2 of *Handbook of Health Economics*, pp. 317 – 396. Elsevier.

Moore, J. (1999). *Advances in economic theory: sixth world congress of the econometric society*, Volume 1, Chapter Implementation in environments with complete information, pp. 182–282. Cambridge University Press.

Mougeot, M. and F. Naegelen (2005). Hospital price regulation and expenditure cap policy. *Journal of Health Economics 24*, 55–72.

Palfrey, T. R. and S. Srivastava (1991). Nash implementation using undominated strategies. *Econometrica 59*(2), pp. 479–501.

Pope, D. G. . (2009.). Reacting to rankings: Evidence from "america's best hospitals". *Journal of Health Economics 28(6).*, 1154–1165.

Propper, C. and G. Leckie (2011). Chapter 28 - increasing competition between providers in health care markets: the economic evidence. In S. Glied and P. Smith (Eds.), *Oxford Handbook of Health Economics*, Oxford Handbook of Health Economics, pp. 671 – 687. Oxford University Press.

Schut, E., S. Sorbe, and J. Hoj (2013). Health care reform and long-term care in the netherlands. Economic Department Working paper ECO/WKP(2013)2, OECD.

Schut, F. T. and M. Varkevisser (2013). Tackling hospital waiting times: The impact of past and current policies in the netherlands. *Health Policy forthcoming*.

Skinner, J. (2012). *Handbook of Health Economics (Volume 2)*, Chapter Causes and Consequences of Regional Variations in Health Care, pp. 45–93. Elsevier.

Song, Z., D. G. Safran, B. E. Landon, M. B. Landrum, Y. He, R. E. Mechanic, M. P. Day, and M. E. Chernew (2012). The 'alternative quality contract,' based on a global budget, lowered medical spending and improved quality. *Health Affairs 31*(8), 1885–1894.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization 1*, 39–60.

Vijsel, A.R. van de, P. E. and G. Westert (2011). Rendering hospital budgets volume based and open ended to reduce waiting lists: does it work? *Health Policy 100(1)*, 60–70.

Wennberg, J. (2010). *Tracking Medicine*. Oxford University Press.

Werner, R., J. Kolstad, E. Stuart, and D. Polsky (2011). The effect of pay-for-performance in hospitals: Lessons for quality improvement. *Health Affairs 30(4)*, 690–698.

A. Proof of results

Proof of lemma 1 First, consider an outcome where both IR constraints are binding. If one of the IC constraints, say equation (IC_h), would be binding as well, we can be in either of two cases: (i)

$$R_l - c(1-\omega_l) - D_h(\omega_l) = 0$$

and we can ignore (IC_h) or (ii)

$$R_l - c(1-\omega_l) - D_h(\omega_l) > 0$$

but this contradicts that IR_h is binding in this outcome.

Second, always at least one IR is binding. Suppose not, then it is possible to reduce both R_h and R_l while satisfying the IC constraints and increasing the planner's objective function (as $\beta < 1$).

Third, it cannot be the case that only (IR_l) and (IC_l) (or equivalently, only (IR_h) and (IC_h)) are binding. If that would be the case R_h (R_l) could be reduced while still satisfying all constraints and increasing the planner's objective function.

Finally, it is not possible that both (IC_l) and (IC_h) are binding. Suppose by contradiction both would be binding. Then adding these two equations yields

$$D_l(\omega_h) - D_l(\omega_l) = D_h(\omega_h) - D_h(\omega_l) \tag{18}$$

This contradicts equation (2). *Q.E.D.*

Proof of proposition 1 Let λ_i denote the lagrange multiplier on (IR_i) and μ_i on (IC_i). First consider case 1: $\lambda_l, \lambda_h > 0$ and $\mu_l = \mu_h = 0$. The first order conditions (for an interior solution) for R_h, R_l yield that

$$\lambda_l = F(1-\beta) \tag{19}$$

$$\lambda_h = (1-F)(1-\beta) \tag{20}$$

Using these equations, it is routine to verify that the first order conditions for ω can be written as equations (4) and (5). In words, if only the IR constraints are binding, the planner implements the first best outcome. In the VBD case, the IC constraints are then satisfied as well.

Consider case 2 in lemma 1: $\lambda_h, \mu_l > 0$ and $\lambda_l = \mu_h = 0$. The first order conditions for R_h and R_l imply that $\mu_l = F(1-\beta)$ and $\lambda_h = 1-\beta$. The first order conditions for $\hat{\omega}_i^2$ can be written as

$$v_l(\hat{\omega}_l^2) = c - D_l'(\hat{\omega}_l^2) \tag{21}$$

$$v_h(\hat{\omega}_h^2) = c - D_h'(\hat{\omega}_h^2) - (1-\beta)\frac{F}{1-F}(D_h'(\hat{\omega}_h^2) - D_l'(\hat{\omega}_h^2)) \tag{22}$$

This is the second best outcome of the PFD case: $\hat{\omega}_l^2$, is equal to its first best outcome $\hat{\omega}_l^1$ while $\hat{\omega}_h^2$ is lower compared to its first best outcome $\hat{\omega}_h^1$. The high quality physician over-treats compared to first best (and thus compared to treatment efficiency) because $D_l'(\omega) < D_h'(\omega)$ (see equation (2)). It is routine to verify that IR_l and IC_h are satisfied as well in this case. *Q.E.D.*

Proof of lemma 2 Consider the CfM outcome with PFD where the payer contracts h-physician in the mixed case. Hence the l-type in the lh-case does not treat anyone. However, to prevent her from claiming to be h, she still needs to be paid! In particular, if the l-physician in case lh claims to be h, the payer randomly selects one physician who then serves the whole market. Thus we find

$$R_m^l \geq \tfrac{1}{2}(R_h - c(1-\omega_h) - D_l(\omega_h)) \tag{23}$$

If h-type claims to be l in hh-case, she gets R_m^l. Hence IC requires

$$\tfrac{1}{2}(R_h - c(1-\omega_h) - D_h(\omega_h)) \geq R_m^l \tag{24}$$

Adding these two inequalities, we find that $D_h(\omega_h) \leq D_l(\omega_h)$ which contradicts (2) because $D_h(0) = D_l(0) = 0$. Hence there is no IC way in which the h-physician can be contracted in the lh-case. *Q.E.D.*

Proof of proposition 2 Consider the case where the planner decides to contract the l-physician in the lh-case. This reduces the information rent in the ll-case to zero. Indeed, the l-type who claims to be h in case of ll loses for sure. Hence IR_l is binding. We assume that IR_h, IC_m^l and IR_m^h are binding and later check that the other constraints are satisfied. IC_m^l can be written as

$$R_m^l - 2(c(1-\omega_m) + D_l(\omega_m)) \geq \tfrac{1}{2}(R_h - 2(c(1-\omega_h) + D_l(\omega_h)))$$

Hence we find

$$R_m^l = 2(c(1-\omega_m) + D_l(\omega_m)) + D_h(\omega_h) - D_l(\omega_h)$$

Hence the payer maximizes

$$2F^2V_l(\omega_l) + 2(1-F)^2V_h(\omega_h) + 2F(1-F)(2V_l(\omega_m) - (1-\beta)(D_h(\omega_h) - D_l(\omega_h)))$$

It follows that $\omega_l = \omega_m = \hat{\omega}_l^1$ and $\omega_h = \hat{\omega}_h^2$. Hence, the expression for welfare under CfM is given by equation (11). We conclude with checking the other constraints. As explained above, IC_l is not binding. It is routine to verify that IC_m^h and IR_m^l are not binding either. Consider IC_h. If h-physician in case hh claims to be l, she wins for sure. She then gets

$$R_m^l - 2(c(1-\omega_m) + D_h(\omega_m))$$

This expression is negative if and only if

$$-2(D_h(\hat{\omega}_l^1) - D_l(\hat{\omega}_l^1)) + D_h(\hat{\omega}_h^2) - D_l(\hat{\omega}_h^2) < 0$$

which holds because $\hat{\omega}_l^1 > \hat{\omega}_h^2$ and $D_h'(\omega) - D_l'(\omega) > 0$ by equation (2). IR_l then implies that h truthfully reveals her type. *Q.E.D.*

Proposition 4 *With complete information, CfM leads to highest welfare.*
If $x = 1$ then $W^{NC} > W^{CoM}$. If $x = 0$, $W^{CoM} > W^{NC}$ if and only if $V_h(\hat{\omega}_h^1) > V_l(\hat{\omega}_l^1)$. In fact, if $V_h(\hat{\omega}_h^1) > V_l(\hat{\omega}_l^1)$, then $W^{CoM} = W^{CfM}$ with $x = 0$.

Proof of proposition 4 Consider CoM. In the hh and ll cases, the payer implements first best: ω_h^1, ω_l^1 resp. In the mixed case, ω_m solves

$$\max_{\omega_m} x\left(\int_{\omega_m}^1 v_l(\omega)d\omega - (c(1-\omega_m) + D_l(\omega_m))\right) + (2-x)\left(\int_{\omega_m}^1 v_h(\omega)d\omega - (c(1-\omega_m) + D_h(\omega_m))\right)$$

First order condition can be written as

$$x(v_l(\hat{\omega}_m^1) - (c - D_l'(\hat{\omega}_m^1))) + (2-x)(v_h(\hat{\omega}_m^1) - (c - D_h'(\hat{\omega}_m^1))) = 0 \tag{25}$$

Hence $\hat{\omega}_m^1$ lies in between $\hat{\omega}_h^1$ and $\hat{\omega}_l^1$. Further, $x = 0$ implies $\hat{\omega}_m^1 = \hat{\omega}_h^1$: if all patients are willing to travel to the h-provider, the payer can implement first best in the mixed case with CoM.

We can write welfare as

$$W^{CoM} = F^2 2V_l(\hat{\omega}_l^1) + (1-F)^2 2V_h(\hat{\omega}_h^1) + 2F(1-F)((2-x)V_h(\hat{\omega}_m^1) + xV_l(\hat{\omega}_m^1)) \tag{26}$$

Compare welfare under CoM and NC. Using equation (8), we find that

$$W^{CoM} - W^{NC} = (1-x)(V_h(\hat{\omega}_m^1) - V_l(\hat{\omega}_l^1)) - \left[x(V_l(\hat{\omega}_l^1) - V_l(\hat{\omega}_m^1)) + V_h(\hat{\omega}_h^1) - V_h(\hat{\omega}_m^1)\right] \tag{27}$$

Note that the expression in square brackets is positive. By definition $\hat{\omega}_i^1$ maximizes $V_i(\omega)$ while this is not the case for $\hat{\omega}_m^1$. Hence, it follows that $W^{NC} > W^{CoM}$ for $x = 1$. Further, with $x = 0$ we know from equation (25) that $\hat{\omega}_m^1 = \hat{\omega}_h^1$. Hence with $x = 0$ we get $W^{CoM} > W^{NC}$ if and only if

$$V_h(\hat{\omega}_h^1) > V_l(\hat{\omega}_l^1) \tag{28}$$

In words, CoM with $x = 0$ yields higher welfare than NC if and only if patients visit provider with highest social surplus. If h-providers yield highest social surplus, it is optimal to let patients choose because they can use decentralized information. If not, NC yields higher welfare either because patients choose the "wrong" provider or because they do not react to decentralized information ($x = 1$).

Equation (28) is satisfied under VBD, but not necessarily under PFD. Indeed, if high quality PFD-physicians have a stronger intrinsic motivation than low quality physicians to treat more patients ($D_h' >> D_l'$) while the additional value of these treatments is rather small then CoM would direct patients to the low surplus h-physician.

Compare welfare under CoM and CfM. Using equation (10) we write

$$W^{CfM} - W^{CoM} = 2F(1-F)[2\max\{V_l(\hat{\omega}_l^1), V_h(\hat{\omega}_h^1)\} - \{(2-x)V_h(\hat{\omega}_m^1) + xV_l(\hat{\omega}_m^1)\}] \geq 0 \qquad (29)$$

The inequality follows because $\hat{\omega}_m^1$ does not maximize $V_i(\omega)$. In fact, the inequality is strict unless $x = 0$ and equation (28) holds. *Q.E.D.*

Proof of proposition 3

First, we show that $x = 1$ implies that ω_m lies in between ω_h and ω_l.

Lemma 4 *With $x = 1$, it is the case that $\hat{\omega}_h^2 \leq \hat{\omega}_m^2 \leq \hat{\omega}_l^2$.*

Proof of lemma 4 We proof this by contradiction. In particular, we show that either $\hat{\omega}_m^2 < \hat{\omega}_h^2$ or $\hat{\omega}_m^2 > \hat{\omega}_l^2$ leads to a violation of IC.

First, assume that $\hat{\omega}_m^2 < \hat{\omega}_h^2$. Then adding (IC_h) and (IC_m^l) in CoM yields

$$D_l(\hat{\omega}_h^2) - D_l(\hat{\omega}_m^2) \geq D_h(\hat{\omega}_h^2) - D_h(\hat{\omega}_m^2)$$

which can be written as

$$\int_{\hat{\omega}_m^2}^{\hat{\omega}_h^2} D_l'(\omega)d\omega \geq \int_{\hat{\omega}_m^2}^{\hat{\omega}_h^2} D_h'(\omega)d\omega$$

However, this contradicts (2). Similarly, adding equations (IC_l) and (IC_m^h) and assuming $\hat{\omega}_m^2 > \hat{\omega}_l^2$, we get

$$\int_{\hat{\omega}_l^2}^{\hat{\omega}_m^2} D_l'(\omega)d\omega \geq \int_{\hat{\omega}_l^2}^{\hat{\omega}_m^2} D_h'(\omega)d\omega$$

which contradicts (2).

Hence, an incentive compatible outcome features $\hat{\omega}_h^2 \leq \hat{\omega}_m^2 \leq \hat{\omega}_l^2$. *Q.E.D.*

We know from proposition 2 that $W^{NC} \geq W^{CfM}$ if equation (12) does not hold (preferences are aligned). Hence the relevant comparison is between W^{NC} and W^{CoM}. The following lemma shows that with $x = 1$, W^{NC} is highest.

Lemma 5 *With $x = 1$, we find that $W^{NC} \geq W^{CoM}$.*

Proof of lemma 5 We show that if $\omega_l, \omega_m, \omega_h, R_l, R_m^h, R_m^l, R_h$ can be implemented under CoM (with $x = 1$) then $\omega_l, \omega_h, R_l, R_h$ can be implemented with NC. The result on welfare then follows because rents are (weakly) lower under NC than under CoM and CoM features the outcome ω_m which is inefficient for either the l or the h-physician or both (lemma 4).

With CoM, combining (IC_l) and (IR_m^h) yields

$$R_l \geq C_l(\omega_l) + C_h(\omega_h) - C_l(\omega_m) \qquad (30)$$

where the function $C_i(\omega)$ is defined in equation (13). Similarly, combining (IC_h) and (IR_m^l) yields

$$R_h \geq C_h(\omega_h) + C_l(\omega_m) - C_h(\omega_m) \tag{31}$$

Now define the following transfers for each monopolist under NC

$$\bar{R}_l = C_l(\omega_l) + \max\{0, C_h(\omega_m) - C_l(\omega_m)\} \tag{32}$$

$$\bar{R}_h = C_h(\omega_h) + \max\{0, C_l(\omega_m) - C_h(\omega_m)\} \tag{33}$$

Clearly, for each monopolist $\bar{R}_i, \omega_i$ satisfies IR_i. Now we show that $\omega_l, \bar{R}_l, \omega_h, \bar{R}_h$ satisfies the two IC constraints under monopoly as well. That is, we need to show that the following inequalities hold

$$C_l(\omega_h) - C_l(\omega_l) \geq \bar{R}_h - \bar{R}_l \geq C_h(\omega_h) - C_h(\omega_l) \tag{34}$$

Using that $\max\{0, x\} - \max\{0, -x\} = x$, we find that

$$\bar{R}_h - \bar{R}_l = C_h(\omega_h) - C_l(\omega_l) + (C_l(\omega_m) - C_h(\omega_m)) \tag{35}$$

Checking the first inequality in equation (34), boils down to verifying that

$$C_l(\omega_h) - C_l(\omega_m) \geq C_h(\omega_h) - C_h(\omega_m)$$

which can be written as

$$\int_{\omega_h}^{\omega_m} D_h'(\omega)d\omega \geq \int_{\omega_h}^{\omega_m} D_l'(\omega)d\omega$$

This inequality holds because of (2) and the result in lemma 4 that $\omega_m \geq \omega_h$. Second inequality in (34) boils down to

$$\int_{\omega_m}^{\omega_l} D_h'(\omega)d\omega \geq \int_{\omega_m}^{\omega_l} D_l'(\omega)d\omega$$

which holds because $\omega_l \geq \omega_m$ by lemma 4. *Q.E.D.*

Hence we have shown that $W^{NC} \geq W^{CoM}$ if $x = 1$. If equation (12) does not hold, we know from proposition 2 that $W^{NC} \geq W^{CfM}$ (this result does not depend on x). Hence with $x = 1$, we know that NC yields highest welfare in this case.

Now consider $x = 0$.

Lemma 6 *Assume $D_h(\omega_h^1) \leq 2D_l(\omega_h^1)$ and (12) does not hold, then $W^{CoM} \geq W^{NC}$ with $x = 0$.*

Proof of lemma 6 We claim that (IR_h), (IR_m^h), (IC_l) and (IR_m^l) are binding under CoM. Later we verify that the other constraints are satisfied as well. As $x = 0$, the l-physician in the mixed case

does not treat anyone and is closed down: $R_m^l = 0$. Given these binding constraints and $x = 0$, we can write the payer's problem as

$$\begin{aligned}
&\max F^2 \left(\int_{\omega_l}^{1} v_l(\omega) d\omega - R_l + \beta(R_l - C_l(\omega_l)) \right) \\
&+ (1-F)^2 \left(\int_{\omega_h}^{1} v_h(\omega) d\omega - C_h(\omega_h) \right) \\
&+ F(1-F) \left(2 \int_{\omega_m}^{1} v_h(\omega) d\omega - C_m^h(\omega_m) \right) \\
&+ \mu_l (R_l - C_l(\omega_l) - C_m^h(\omega_m) + \tilde{C}_{ll}(\omega_m))
\end{aligned}$$

First order condition with respect to R_l yields $\mu_l = F^2(1-\beta)$. The first order conditions for $\omega_{l,h}$ can be written as

$$\begin{aligned}
v_l(\omega_l) &= c - D_l'(\omega_l) \\
v_h(\omega_h) &= c - D_h'(\omega_h)
\end{aligned}$$

Hence, we find that $\omega_l^{CoM} = \hat{\omega}_l^1$ and $\omega_h^{CoM} = \hat{\omega}_h^1$. Using $\mu_l = F^2(1-\beta)$, we write the first order condition for ω_m as

$$(1-F)(-2v_h(\omega_m) - C_m^{h'}(\omega_m)) + F(1-\beta)(-C_m^{h'}(\omega_m) + \tilde{C}_{ll}'(\omega_m)) = 0$$

or equivalently

$$v_h(\omega_m) = (c - D_h'(\omega_m)) + \frac{F}{1-F}(1-\beta)(D_l'(\omega_m) - D_h'(\omega_m))$$

Comparing this to equation (22), we find that $\omega_m^{CoM} = \omega_h^{NC} = \hat{\omega}_h^2$.

We verify that the four constraints that we ignored above are indeed satisfied. As the l-physician in the mixed case is closed down with $R_m^l = 0$, (IR_h) and (IC_h) are the same. Next, $D_h(\hat{\omega}_h^1) \leq 2D_l(\hat{\omega}_h^1)$ together with (IR_h) implies (IC_m^l). Third, (IC_l) together with $D_h(\omega) \geq D_l(\omega)$ implies that (IR_l) is satisfied as well. Finally, (IC_m^h) follows from (IR_m^h) in case

$$R_l - 2D_h(\frac{1+\omega_l}{2}) - c(1-\omega_l) \leq 0 \tag{36}$$

where the deviation threshold $\frac{1+\omega_l}{2}$ is derived in lemma 8 in appendix B. We can solve for R_l from (IC_l) and (IR_m^h). This allows us to write equation (36) as

$$2D_h(\frac{1+\omega_l}{2}) - D_l(\omega_l) \geq 2(D_h(\omega_m) - D_l(\omega_m)) \tag{37}$$

This inequality is satisfied, because of the following chain of inequalities:

$$2D_h(\frac{1+\omega_l}{2}) - D_l(\omega_l) > 2(D_h(\frac{1+\omega_l}{2}) - D_l(\omega_l)) > 2(D_h(\omega_l) - D_l(\omega_l)) > 2(D_h(\omega_m) - D_l(\omega_m)) \tag{38}$$

First inequality follows from $D_l(\omega_l) > 0$. Second inequality follows from $(1+\omega_l)/2 > \omega_l$ and $D'_h(\omega) > 0$. Third inequality follows from $\omega_m < \omega_l$ and equation (2).

Comparing welfare under NC and CoM, we can write

$$W^{CoM} = 2F^2(V_l(\hat{\omega}_l^1) - 2(1-\beta)(D_h(\hat{\omega}_h^2) - D_l(\hat{\omega}_h^2))) + 2(1-F)^2 V_h(\hat{\omega}_h^1) + 4F(1-F)V_h(\hat{\omega}_h^2) \quad (39)$$

and

$$\begin{aligned} W^{NC} =& 2F^2(V_l(\hat{\omega}_l^1) - (1-\beta)(D_h(\hat{\omega}_h^2) - D_l(\hat{\omega}_h^2))) + 2(1-F)^2 V_h(\hat{\omega}_h^2) + \\ & 2F(1-F)(V_h(\hat{\omega}_h^2) + V_l(\hat{\omega}_l^1) - (1-\beta)(D_h(\hat{\omega}_h^2) - D_l(\hat{\omega}_h^2))) \end{aligned} \quad (40)$$

Therefore, $W^{CoM} \geq W^{NC}$ if and only

$$\begin{aligned} & (1-F)^2 \left(V_h(\hat{\omega}_h^1) - [V_h(\omega_h^2) - (1-\beta)\frac{F}{1-F}(D_h(\hat{\omega}_h^2) - D_l(\hat{\omega}_h^2))] \right) \geq \\ & F(1-F) \left(V_l(\hat{\omega}_l^1) - [V_h(\omega_h^2) - (1-\beta)\frac{F}{1-F}(D_h(\hat{\omega}_h^2) - D_l(\hat{\omega}_h^2))] \right) \end{aligned} \quad (41)$$

This inequality holds as the left hand side is positive and the right hand side is negative because equation (12) does not hold. *Q.E.D.*

If (12) does not hold, proposition 2 implies $W^{NC} \geq W^{CfM}$. Hence we find that CoM yields highest welfare in this case.

Now consider the case where inequality (12) does hold. As lemma 5 does not depend on (12), we find that $W^{NC} \geq W^{CoM}$ with $x = 1$. Proposition 2 then implies that W^{CfM} is highest.

Proposition 2 does not depend on x, hence also with $x = 0$ we have $W^{CfM} \geq W^{NC}$ if (12) holds and the relevant comparison is between equations (39) and (11). It is routine to verify that $W^{CoM} \geq W^{CfM}$ if and only if (15) holds. *Q.E.D.*

B. Physician intrinsic motivation

In this appendix, we derive the dis-utility functions $D_i(\omega)$ in definition 2 from first principles. Lemma 8 derives the deviation cost functions $\tilde{C}_{ij}$ that we use in section 5.

Let y denote the value that a physician "feels" she can create for a patient. We make this "feeling" more concrete below. Assume a patient with $y > 0$ walks into a physician's office. We assume that the physician is trained to help patients where she can add value. Hence, if the physician is allowed to treat this patient, she does not experience any dis-utility. Her intrinsic motivation to help patients where she adds value ensures she helps the patient without dis-utility. If, on the other hand, she is not allowed to treat a patient with $y > 0$, she experiences a dis-utility $d_u(y) > 0$. She has to send away a patient where she feels that she could help. This leads to dis-utility.

Now, consider a patient where $y < 0$. If the physician does not treat such a patient, she has no dis-utility. She feels that she cannot help the patient and hence sending him away is the best thing to do. If, on the other hand, she has to treat this patient, her dis-utility is given by $d_t(y) > 0$.

We define the shape of the two disutility funtions $d_u(y)$ and $d_t(y)$ as follows:

Assumption 1 *Physician dis-utility functions satisfy*

$$d_u(y) \begin{cases} = 0 & \text{if } y \leq 0 \\ > 0 & \text{if } y > 0 \end{cases} \text{ and } d_t(y) \begin{cases} = 0 & \text{if } y \geq 0 \\ > 0 & \text{if } y < 0 \end{cases} \tag{42}$$

where $d_u'(y) > 0$ *for* $y > 0$ *and* $d_t'(y) < 0$ *for* $y < 0$.

Function $d_u(y)$ is increasing in the value that could be created by the physician. A physician's dis-utility is higher if she cannot treat a patient where she could create more value. Similarly, having to treat a patient where she creates bigger damage (more negative y) leads to higher dis-utility $d_t(y)$. These two assumptions imply that physicians treat the most severe (highest y) patients first. And if they have to treat patients with $y < 0$, they treat the patients where the damage is lowest. This implies that physicians use a cut off rule in deciding which patients to treat.

Note, that we assume non-negative dis-utility. There are two ways to motivate this assumption. First, a negative disutility would imply that physicians are willing to perform treatments even if the (monetary) costs of treatments are not covered by the (monetary) revenue. Although this does happen, for many physicians this is not sustainable as a long run business strategy. Second, the utility of being a physician and being able to help people is part of the job and here modeled as an outside option (this utility would be lost if one stops being a physician). Similarly, one could assume that dis-utility is always positive. For ease of exposition, we assume this is part of the outside option as well. Below, we normalize the outside option to zero but this does not affect the results.

The value that a physician of quality $i \in \{l, h\}$ "feels" she can create for a patient is given by $y = v_i(\omega) - \alpha$ where we distinguish three physician types α.

Assumption 2 *We distinguish three physician types*

- **PFD-physician** *Patient Focused Disutility:* $\alpha = v_l(0)$
- **VBD-physician** *Value Based Disutility:* $\alpha = c$
- **ECD-physician** *Effort Cost Disutiliy:* $\alpha = v_h(1)$

The PFD physician wants to help any patient since $v_h(\omega) \geq v_l(\omega) \geq v_l(0)$. Assuming $v_l(0) \geq 0$ this implies that the physician is patient focused: a fully insured patient wants to be treated if $v_i(\omega) \geq 0$.

The dis-utility of a PFD physician is line with patient's utility. However, treatments costing $c > v_l(0)$, this is not optimal from society's point of view. A VBD physician compares cost and benefits of treatment and only feels she can contribute if the gain $v_i(\omega)$ of treatment exceeds the cost c. As the insured end up paying for the treatments, a VBD physician can also be characterized as having insured focused dis-utility.

Finally, for an ECD physician $y = v_i(\omega) - v_h(1) \leq 0$. For this physician every treatment involves an effort cost. This is the "homo economicus" usually considered in regulation models.

Now we can derive the dis-utility functions $D_i(\omega)$ used in the main text. As noted above, physicians use a cut off rule: treat everyone with $\omega > \hat{\omega}$ and do not treat patients with $\omega < \hat{\omega}$. Given $\hat{\omega}$, dis-utility D_i is obtained by integrating the dis-utility d_j over all patients that walk into the physicians office.

Definition 4

$$D_h(\hat{\omega}) = \int_0^{\hat{\omega}} d_u(v_h(\omega) - \alpha)d\omega + \int_{\hat{\omega}}^1 d_t(v_h(\omega) - \alpha)d\omega \tag{43}$$

$$D_l(\hat{\omega}) = \int_0^{\hat{\omega}} d_u(v_l(\omega) - \alpha)d\omega + \int_{\hat{\omega}}^1 d_t(v_l(\omega) - \alpha)d\omega \tag{44}$$

The following lemma derives the properties in definition 2.

Lemma 7 *The functions D_h, D_l defined in equations (43) and (44) have the following properties:*

- $D_i(\omega) \geq 0$ *for each* $\omega \in [0, 1]$,
- $D_i(\omega)$ *is convex in* ω,
- $D'_h(\omega) > D'_l(\omega)$ *for each* $\omega \in [0, 1]$,
- $D_i(0) = 0$ *for a PFD physician,*
- $D_i(\omega_i^*) = 0$ *for a VBD physician and*
- $D_i(1) = 0$ *for an ECD physician.*

Proof The first property follows from $d_u(y), d_t(y) \geq 0$. The convexity of D_i follows from $D''_i(\omega) = d'_u - d'_t > 0$. The third property follows from

$$D'_h(\omega) = d_u(v_h(\omega) - \alpha) - d_t(v_h(\omega) - \alpha) > d_u(v_l(\omega) - \alpha) - d_t(v_l(\omega) - \alpha) = D'_l(\omega) \tag{45}$$

because $v_h(\omega) > v_l(\omega)$ and $d'_u(y) > 0$ for $y > 0$, $d'_t(y) < 0$ for $y < 0$. For a PFD physician $d_t(v_i(\omega) - v_l(0)) = 0$ for each $\omega \in [0, 1]$. Hence treating everyone yields zero dis-utility: $D_i(0) = 0$. For a VBD physician $d_t(v_i(\omega) - c) = 0$ for each $\omega \geq \omega_i^*$ and $d_u(v_i(\omega) - c) = 0$ for each $\omega < \omega_i^*$. Hence $D_i(\omega_i^*) = 0$. Finally, for a ECD physician, we find that $d_u(v_i(\omega) - v_h(1)) = 0$ for each $\omega \in [0, 1]$ and hence $D_i(1) = 0$.

Q.E.D.

Lemma 8 *Functions $\tilde{C}$ are given by the following expressions:*

$$\tilde{C}_{ll}(\omega_m) = (1 + I(x<1)) \int_0^{\omega_m} d_u^l(\omega)d\omega + \int_{\omega_m}^1 d_t^l(\omega)d\omega + \int_{\omega_m}^{\tilde{\omega}_{ll}} d_t^l(\omega)d\omega + c(2-x)(1-\omega_m) \tag{46}$$
where $\tilde{\omega}_{ll} = 1 - x(1-\omega_m)$

and $I(x<1) = 1$ *if* $x < 1$ *and zero if* $x = 1$.

$$\tilde{C}_{hl}(\omega_l) = (2-x)\left(\int_0^{\tilde{\omega}_{hl}} d_u^h(\omega)d\omega + \int_{\tilde{\omega}_{hl}}^1 d_t^h(\omega)d\omega\right) + c(1-\omega_l) \tag{47}$$
$$\text{where } \tilde{\omega}_{hl} = \frac{1-x+\omega_l}{2-x}$$

$$\tilde{C}_{lh}(\omega_h) = (1 + I(x<1)) \int_0^{\omega_h} d_u^l(\omega)d\omega + x\int_{\omega_h}^1 d_t^l(\omega)d\omega + (2-x)\int_{\omega_h}^{\tilde{\omega}_{lh}} d_t^l(\omega)d\omega + c(1-\omega_h) \tag{48}$$
$$\text{where } \tilde{\omega}_{lh} = \frac{1-x+\omega_h}{2-x}$$

$$\tilde{C}_{hh}(\omega_m) = \int_0^{\tilde{\omega}_{hh}} d_u^h(\omega)d\omega + \int_{\tilde{\omega}_{hh}}^1 d_t^h(\omega)d\omega + xc(1-\omega_m) \tag{49}$$
$$\text{where } \tilde{\omega}_{hh} = 1 - x(1-\omega_m)$$

Proof The key to understanding the functions $\tilde{C}$ is the allocation of patients over the providers in case of CoM. A deviation by provider P_1 leads to patient allocations over providers that are illustrated in figure 4. Note that if both providers use the same threshold (as they will in equilibrium) a patient who is not treated by physician 1 has no incentive to travel to physician 2. Physicians perfectly observe ω and hence the patient will be turned down by the other physcian as well. But out-of-equilibrium (after a deviation) thresholds can differ and we need to take into account that the patient travels in this case. There are two reasons why this is important. First, travelling patients can crowd out patients (with lower ω) living close to the physician. Second, a physician can experience dis-utility from turning down a patient. As patients travel, potentially more patients need to be turned down.

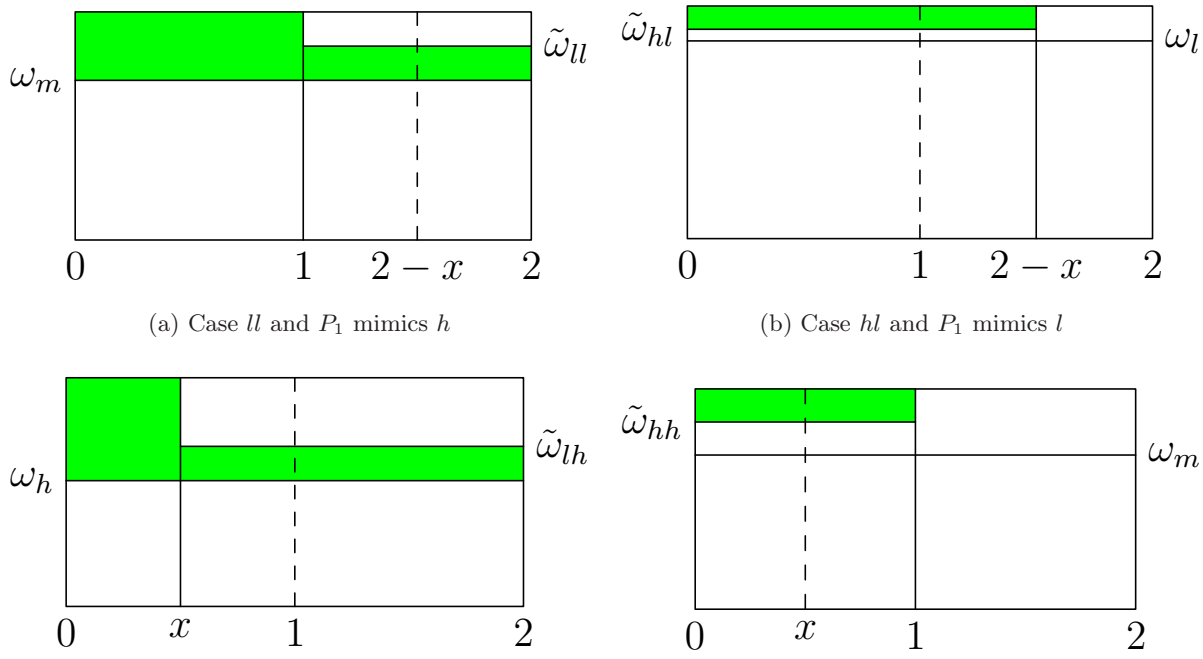
We go over the four cases one by one, where in each case we assume that P_1 deviates. The x patients on the left (right) of each graph tend to go to $P_1(P_2)$ as preferred provider.

First, consider the *ll*-case. As consumers know the quality of the providers, the market initially splits 1:1. If P_1 decides to mimic a *h*-type, the planner gets signals *hl* and implements ω_m where P_1 is supposed to perform $(2-x)(1-\omega_m)$ treatments and P_2 has to do $x(1-\omega_m)$ treatments. However, because the market splits $1:1$ and physicians ration efficiently, P_2 has to use threshold $\tilde{\omega}_{ll}$ where

$$(1-\tilde{\omega}_{ll}) = x(1-\omega_m)$$

That is, P_2 has to adjust the threshold to get the number of treatments specified in the contract. Further, because P_1 and P_2 use different thresholds for $x < 1$, every patient rejected by P_2 goes to P_1

to see whether he can get treatment there. This implies that P_1's disutility for not treating patients equals $2\int_0^{\omega_m} d_u^l(\omega)d\omega$. Hence the costs for P_1 when mimicking an h-type are given by equation (46) where the indicator function is necessary because $x < 1$ implies that $\tilde{\omega}_{ll} > \omega_m$: providers use different cut off values for ω. For $x = 1$, we find that $\tilde{\omega}_{ll} = \omega_m$ and hence a patient who is not treated by P_2 will not visit P_1 as he knows that at P_1 he will not be treated either.



(a) Case ll and P_1 mimics h

(b) Case hl and P_1 mimics l

(c) Case lh and P_1 mimics h

(d) Case hh and P_1 mimics l

Figure 4: Patients treated by P_1 when P_1 deviates.

Consider the hl-case where P_1 pretends to be a l-type. The planner then implements ω_l and P_1 has to perform $1 - \omega_l$ treatments. However, the market splits $2 - x : x$ and P_1 rations efficiently. Hence the threshold $\tilde{\omega}_{hl}$ used by P_1 is given by

$$(2 - x)(1 - \tilde{\omega}_{hl}) = 1 - \omega_l$$

Because P_1 and P_2 use different thresholds, patients rejected by P_1 will go to P_2. However, this effect does not show up in P_1's costs, which are given by (47).

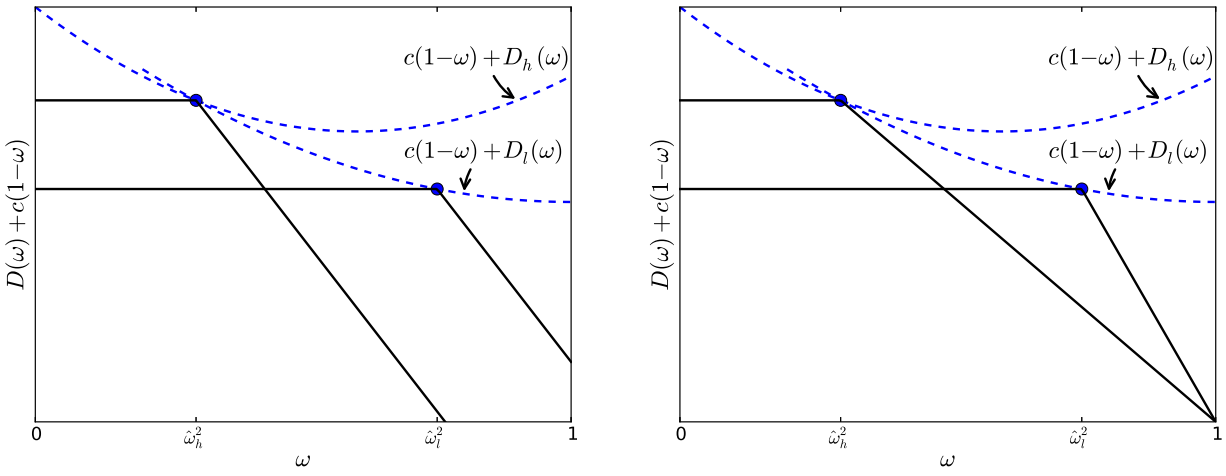
Next, consider the lh-case where P_1 mimics a h-type. Hence the planner implements ω_h and following a similar reasoning as above, we find the costs given by (48).

Finally, in the hh-case where P_1 pretends to be the l-type, we get equation (49). *Q.E.D.*

C. Various optimal payment contracts

First, consider the equilibrium with VBD (see figure 3a). In this particular case the contract can also be characterized by a fee for service contract with the fee equal to the cost of treatment c. If both providers are offered this (linear) contract, each provider will choose the optimal cut off point ω_i^*. Hence, for a VBD physician it is the case that if the payer exactly compensates her treatment costs c then this will result in treatment efficiency.

However, such a simple contract works only for a VBD-physician. Offering this contract to a PFD-physician, leads to an over-provision of treatments. This payment contract results in the intrinsic optimum $\tilde{\omega}_i = 0$ and each patient $\omega \geq 0$ is treated in this case.



(a) two capitation fees

(b) no capitation fee

Figure 5: Capitation, fee for service and budget cap in the PFD case.

Figure 5 illustrates two ways in which the second best optimum can be implemented in the PFD case. First, consider figure 5a. The financial contracts are depicted in this figure as solid (black) lines and the indifference curves for the providers l, h are dashed curves. These are the indifference curves corresponding to the second best outcome in proposition 1. The optimal contracts need to have two properties: (i) the indifference curve through the optimal points $(\tilde{\omega}_l^2, \tilde{R}_l^2)$ and $(\tilde{\omega}_h^2, \tilde{R}_h^2)$ should have these points in common with the respective contracts and (ii) the solid lines should never lie above any dashed curve.[22]

The two solid lines in figure 5a satisfy both properties. The contracts are three-part tariffs. First,

[22] If the solid line lies above the indifference curve at some point (ω, R), the provider strictly prefers this point above the contract that the provider is meant to chose.

there is a capitation fee. This fee is positive for the l-physician and negative for the h-physician. Then there is a fee-for-service part with the fee exceeding c. In the figure, the fee is the same for both types. Finally, there is a spending cap or budget. The h-physician has a higher budget than the l-budget and hence treats more patients.

Figure 5b shows an alternative menu of contracts that can also be used. In this case, there is no capitation fee. The l-physician gets a higher fee-for-service than the h-physician, but she faces a lower budget. Both these menus of contracts yield the same second best outcome. Clearly, there are many more menus of contracts that can implement the same result. Therefore, we focus on the optimal outcome, not on the way this can be implemented.

However, we can point out that some contracts used in practice do not implement the optimal outcome. Payers often pay providers a fixed budget.[23] However, this is not optimal. First, without a requirement on how many patients need to be treated (or without a fee-for-service component), the budget has to be the same for all hospitals. To illustrate, if the planner would offer a higher budget for h-providers, each provider would choose the option with the higher budget. Second, given a fixed budget, a provider chooses the cut off ω to minimize total costs $c(1-\omega)+D(\omega)$. As shown in figure 3, this outcome does not correspond to the second nor the first best; neither for the VBD nor the PFD case.

D. ECD-physician

It is routine to verify that in the ECD case we have $\lambda_l, \mu_h > 0$ and $\lambda_h = \mu_l = 0$. The first order conditions for R_h and R_l imply that $\lambda_l = (\gamma - \beta)$ and $\mu_h = (1-F)(\gamma-\beta)$. The first order conditions for ω can then be written as follows.

$$v_h(\omega_h) = (c - D_h'(\omega_h)) \tag{50}$$

$$v_l(\omega_l) = (c - D_l'(\omega_l)) + (1-\beta)\frac{1-F}{F}(D_h'(\omega_l) - D_l'(\omega_l)) \tag{51}$$

Hence we find the following result in the ECD case.

[23] For example in the Netherlands (see Schut and Varkevisser, 2013), the government paid in the 1980s a fixed budget for each hospital. In 1985 the budget was determined by using "agreed upon level of expected output". In 2000 the government abolished the budgets, to introduce activity based payments which resemble a fee-for-treatment payment. In 2012, again budgets were introduced but now at the macro level for the whole hospital market. In all these payment systems the parameters of the contracts did not vary with the number of patients nor with quality.

Proposition 5

$$1 = \tilde{\omega_l} > \hat{\omega}_l^2 > \hat{\omega}_l^1 > \omega_l^*$$
$$1 = \tilde{\omega_h} > \hat{\omega}_h^2 = \hat{\omega}_h^1 > \omega_h^*$$

In this case, the h-type uses the first best threshold $\hat{\omega}_h^1$. As treatment requires effort cost, the h-type would like to mimic the l-type who performs fewer treatments. In order to avoid this, the payer distorts ω_l upwards (l type performs fewer treatments than in first best). Since our concern in this paper is over-treatment, we do not analyze the ECD case in this paper.