Griebenow, Malte; Kifmann, Mathias

Working Paper

Diagnostics and treatment: On the division of labor between primary care physicians and specialists

HCHE Research Paper, No. 25

Provided in Cooperation with:
Hamburg Center for Health Economics (hche), University of Hamburg

Suggested Citation: Griebenow, Malte; Kifmann, Mathias (2021) : Diagnostics and treatment: On the division of labor between primary care physicians and specialists, HCHE Research Paper, No. 25, University of Hamburg, Hamburg Center for Health Economics (HCHE), Hamburg

This Version is available at:
https://hdl.handle.net/10419/247690

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

hche | Hamburg Center for Health Economics

Diagnostics and Treatment: On the Division of Labor between Primary Care Physicians and Specialists

Malte Griebenow, Mathias Kifmann

Research Paper
Year: 2021
No: 25

Diagnostics and Treatment: On the Division of Labor between Primary Care Physicians and Specialists

Malte Griebenow*, Mathias Kifmann

***Corresponding author:** Malte Griebenow, Hamburg Center for Health Economics (HCHE), Universität Hamburg, Esplanade 36, 20354 Hamburg, Germany. Malte.griebenow@uni-hamburg.de, Tel: +49 (0)40-42838-9116.

Mathias Kifmann, Hamburg Center for Health Economics (HCHE), Universität Hamburg, Esplanade 36, 20354 Hamburg, Germany. Mathias.kifmann@uni-hamburg.de.

Abstract

This paper analyzes the referral processes between a gatekeeping primary-care physician (PCP) and a specialist. Specialists provide superior treatment for some patients but are more costly than PCPs. Agency problems arise because diagnostic signals are private information of the physicians. Welfare optimizing contracts can call for a markup either to the PCP for treating patients without referral or to the specialist for referring patients back to the PCP. If the benefit of specialist treatment is uncertain, small markups for the specialist enhance welfare compared to a cost-based fee-for-service contract. Additionally, we consider how waiting costs for referrals affect our main results.

Highlights

- We study referrals between a PCP and a specialist.
- Optimal payment system calls for markup for either immediate PCP treatment or for the specialist when she refers patients back to the PCP.
- Markups for the specialist may be superior because of better rent efficiency.

1 Introduction

An important topic in health care is that patients obtain diagnosis and treatment by the appropriate provider. This often requires that providers refer patients to other providers. For instance, in 2009, 9.3% of patient visits to US ambulatory physicians lead to a referral to another physician (Barnett et al., 2012). The literature on referrals has shown that the extent of appropriate referrals depends on provider incentives. If providers are paid by a capitation scheme, they have an incentive to refer more patients than otherwise, since they can save on their own costs by not treating the patient. Conversely, if providers are paid by fee-for-service (FFS) payments, they are incentivized to (over-)treat patients themselves (Iversen and Lurås, 2000; Allard et al., 2011; Sarma et al., 2018). Not referring patients that would have greatly benefited from a referral deteriorates patient's outcomes (under-referral), whereas referring patients who do not or only marginally benefit from a referral leads to unnecessary costs on the health care system (over-referral). Empirically, there is evidence for both over- and under-referrals (Mehrotra et al., 2011).

Previous literature has focused on the initial referral decision. This paper goes further and also considers possible strategic decisions by the specialist to whom the patient is referred. The specialist diagnoses the patient and decides whether to treat herself or to refer the patient back to primary care. Primary care physicians (PCPs) can not be expected to be proficient enough in every specialty to perfectly diagnose a patient's health status. Therefore, some patients who do not require specialist treatment may be referred anyways. Since treatment costs for specialists are often higher than the costs of PCP treatment (Whittle et al., 1998; Harrold et al., 1999), it can be efficient to refer the patient back to the PCP even if specialist treatment confers some additional benefit over PCP treatment. Thus, both providers need to be incentivized to make appropriate referral decisions.

We consider two information structures for the PCP's diagnostic procedure. In our benchmark case, the PCP is able to identify some low-severity patients while being unable to identify high-severity patients. This is relevant if severe cases always exhibit specific symptoms. If a symptom is not present, the PCP can conclude that the patient is not severely ill. In the second alternative structure, by contrast, the PCP is able to identify some high-severity patients while being unable to identify low-severity patients. This is relevant if the existence of specific symptoms is highly indicative of the patient's severe disease state but patients without those symptoms may still be severely ill.

As PCPs can only imperfectly determine whether a patient benefits from the specialist's treatment they should refer some patients with an unclear diagnosis. The specialist, on the other hand, should refer back patients who would only benefit little from her treatment. This situation is of particular relevance for patients suffering from chronic diseases. For example, older patients with diabetes can be treated either by their PCP or an endocrinologist. Research has shown that treatment by endocrinologists is more costly but does not necessarily lead to better health outcomes (Chin et al., 2000). Therefore, the PCP should not refer all patients to the endocrinologist and the endocrinologist should refer back patients that can be treated by the PCP. Similarly, a patient who suffers from a mild case of asthma can be treated by the PCP,

whereas more severe cases should be referred to a pulmonologist (Government of Western Australia - Department of Health, 2006). After the patient's condition has stabilized he should be referred back to the PCP (Schermer et al., 2003). Patients with chronic kidney disease in stages 3 and 4 can be managed in either primary or secondary care. Among the treatment options are that the patient is referred for diagnosis to a nephrologist and then transferred back to the PCP for care (Wilson et al., 2012).

We develop a theoretical model to analyze the referral processes between a gatekeeping PCP and a specialist. Both providers are assumed to partly internalize patients' benefits. An agency problem arises because the payer can not observe the patients' severities and the physicians' diagnostic results. Even after the treatment is performed it is not possible to verify whether the treatment was appropriate. Therefore, our model deals with a credence good (see Darby and Karni, 1973, further Dulleck and Kerschbamer, 2006 and Kerschbamer and Sutter, 2017 for surveys). Both under- and over-treatment can potentially arise. Over-charging is not an issue since we assume that the treatments provided by physicians can be verified.

The key problem in our setting is to implement cost-effective treatment by the appropriate provider. Physicians who partly internalize the benefit of treatment that accrues to the patient cannot be expected to make the corresponding referral choices if they do not internalize the costs of the other physician or the system as a whole. In particular, this can lead to an over-supply of specialist treatment if specialist treatment is more effective than PCP treatment. The aim of this paper is to find socially efficient contracts to counteract this problem. An important aspect in this context is that different fee schemes may lead to different information rents for the physicians if payments need to be non-negative. We consider this aspect in our analysis.

Payment systems that optimally incentivize both providers' referrals are derived. We find that altruistic physicians tend to over-supply specialist care under a cost-based FFS contract. Hence, under both information structures welfare optimizing contracts can call for markups

(a) to the PCP for treating patients without referral, or

(b) to the specialist for referring patients back plus cost sharing for treatment.

Either option can be efficient, depending on the benefit that patients receive from specialist treatment and the difference in the treatment costs between the physicians. Additionally, under the alternative information structure, employing both markups may be necessary.

Markups for the PCP can generate rents for the PCP if payments can not be negative, whereas the rent from a markup to the specialist for back-referring the patient can be extracted through employing cost sharing when the specialist treats patients. This makes markups to the PCP less attractive for the payer.

We also consider the case that the payer faces uncertainty with regard to the benefit of specialist treatment. Then markups for the specialist are welfare enhancing as long as they are sufficiently small. Furthermore, we examine the impact of waiting costs on our main results. In this case the patient suffers waiting costs whenever he is referred to another physician. If waiting costs are a factor, it is more likely to be optimal to incentivize only the PCP to discriminate based on her diagnostic signal.

The paper proceeds as follows. Section 2 discusses related literature. In Section 3, we present the model. In Section 4, we characterize the first-best division of labor between PCPs and specialists. In Section 5, we derive optimal contracts, given that the payer does not observe diagnostic signals. In Section 6, we examine the model robustness with regard to different assumptions. We consider uncertainty with regard to the benefit of specialist treatment, waiting cost of referrals, and the alternative information structure. Section 7 concludes.

2 Literature Review

So far, the theoretical literature on referrals has focused on the incentives for gatekeeping primary-care physicians (PCPs) and has mostly not considered interactions with other providers of care. Garcia-Mariñoso and Jelovac (2003) and Malcomson (2004) derive optimal payment contracts for gatekeeping PCPs who can vary their diagnosis effort. Effort can be incentivized by imposing cost sharing on the PCP when she refers a patient. However, this may lead to fewer referrals than the efficient amount. Cost-responsibility for PCPs' referrals have been employed in the fundholding scheme of the NHS. This led to the desired effect of lowering elective hospital admissions (Dusheiko, Gravelle, Jacobs, et al., 2006) at the cost of reduced patient care satisfaction (Dusheiko, Gravelle, Yu, et al., 2007).

González (2009) compares gatekeeping with free specialist choice when some patients make informed decisions. Allard et al. (2011) compare the efficiency of common payment systems with regard to optimal referral decisions. They find that both FFS and PCP cost sharing arrangements can reduce unnecessary specialist treatment. Allard et al. (2014) consider PCP self-selection into capitation or FFS payments and show that this is never optimal under endogenous diagnostic effort or competition. Shumsky and Pinker (2003) consider a situation in which the gatekeeper not only has an information advantage with regard to the optimal treatment decision but also his own ability. They find that a bonus for patient volume in addition to bonuses based on referral rates may be necessary for first-best performance.

A limitation of this literature is the analysis of specialist behavior. Specialists are assumed to treat all patients who are referred. They do not act strategically themselves. However, incentives for specialists are important as well. Similarly to PCPs, the specialists' treatment decisions affect the patient benefit and the costs of care. Furthermore, the specialist's behavior may affect the PCP's behavior. This may have an influence on the optimal payment system from the payer's viewpoint.

A few papers have considered incentives for specialists. In Brekke et al. (2007) hospitals can choose their specialization and quality. The authors show under which circumstances gatekeeping is superior to free specialist choice. They do not incorporate strategic interactions between the referring PCP and the hospital. Similarly to our paper, Godager, Iversen, et al. (2015) consider a specialist who can refer patients back to the PCP. However, they do not derive welfare maximizing payment contracts. There are two papers which consider referrals between heterogeneous experts in a more general setting. Experts in Liu, Ma, and Mak (2018)

differ in ability and costs, and work in a partnership which is constrained by a minimum profit constraint. They find that an expert partnership with unknown altruism can be incentivized through this constraint. By contrast, we consider non-cooperative physicians and do not impose a joint profitability constraint. Grassi and Ma (2016) consider profit-maximizing experts with differing cost advantages between projects who can refer clients between each other. However, their focus is on expert organizations and not on payment contracts.[1]

Kerschbamer, Sutter, and Dulleck (2017) consider a credence goods market in which consumers can verify the quality of the good they receive. In this setting, the standard model of selfish utility maximization predicts that providers are willing to provide the efficient quality of the good if and only if they receive equal markups for any product quality. This corresponds to a cost-based FFS contract in our setting. In their experiments, however, the authors find that providers tend to over-treat the consumer in the equal markup case. This confirms the importance of other-regarding preferences. Consequently, it is valuable to analyze optimal contracts for pro-social experts in a credence goods market. In our setting, specialist over-treatment is problematic because societies' resources are not used in an efficient manner.

We contribute to the literature by deriving contracts that efficiently solve the problem of specialist over-treatment resulting from physician altruism. In this context, the efficiency of a contract is determined by the resulting treatment paths of the patients as well as the information rents that accrue to the physicians. In contrast to previous literature, we consider incentives for both PCPs and specialists.

3 The Model

A health care payment authority (the payer) contracts with a PCP and a specialist to treat patients who suffer from a disease which can take the severities $k \in \{L, H\}$. The share $0 < p < 1$ of the patients is severely ill ($k = H$), the remaining share $1-p$ suffers from a mild illness ($k = L$). The disease can be treated by both types of physicians and both physicians are assumed to have sufficient capacity to treat all patients. However, the effectiveness of treatment differs between the physicians. In particular, the specialist can treat the high-severity cases better due to her more sophisticated disease-specific technical and human capital.[2]

We model the differences in treatment abilities by assuming that each physician provides one treatment which has different benefits for the patients depending on severity. Treating a patient of type k with treatment $j \in \{P, S\}$ confers a benefit of b_k^j to the patient. High-severity patients receive a surplus benefit from specialist treatment ($\kappa_H := b_H^S - b_H^P > 0$) which is greater than the surplus benefit for low-severity patients ($\kappa_H > b_L^S - b_L^P =: \kappa_L$). Furthermore,

[1] An alternative to creating optimal payment schemes is to allow kickback payments between physicians. Pauly (1979) finds that this can be welfare enhancing by giving physician an incentive to refer patients who can be treated more cost-efficiently by another physician. Inderst and Ottaviani (2012) find that mandatory disclosure of kickbacks can have ambiguous welfare effects. We do not consider kickbacks in this paper.

[2] For ease of exposition, we adopt the linguistic convention that the physicians are female and the payer and the patient male.

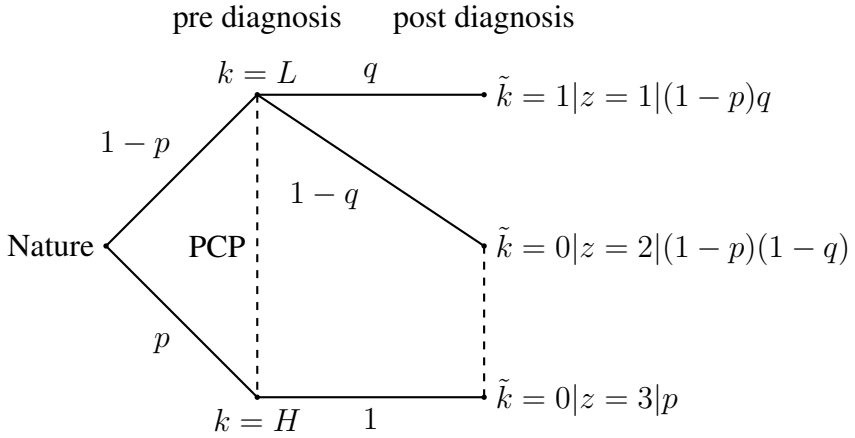


Figure 1: Cases z, dashed lines indicate that the PCP can not differentiate between these cases and therefore cannot identify a high-type patient with certainty

κ_L may be positive or negative. In the first case, a low-severity type patient benefits from the higher sophistication of the specialist treatment. In the second case, he suffers from over-treatment or the PCP is more experienced in treating low-severity diseases. Since we assume that both treatments cure the disease, patients can not receive both treatments. The costs of treatment, c_j, depend on the physician as well. Due to the higher sophistication of the specialist, it is reasonable to assume that she has higher costs of treating a patient than the PCP, i.e. $c_S > c_P$. Any cost parameter presented in this model includes both the direct costs and time-costs of the physicians.

Before a treatment decision can be made, the patient needs to be diagnosed. In our benchmark case, we assume that the PCP can identify a low-type patient with probability q but can not identify a high-type patient with certainty, since she is less experienced with this disease severity. In Subsection 6.3, we explore the alternative assumption that the PCP can identify a share of high-severity patients but is not able to perfectly identify low-severity patients. For simplicity, we assume that this diagnosis is costless. The specialist, by contrast, can identify any patient with certainty at cost $d_S > 0$.[3] We assume that physicians always diagnose their patients.

The state of PCP-knowledge is denoted by $\tilde{k} \in \{0, 1\}$. If the PCP can identify the severity of the patient's disease, she receives $\tilde{k} = 1$, otherwise $\tilde{k} = 0$. This leads to the definition of the following three cases $z \in \{1, 2, 3\}$:

$$z = \begin{cases} 1, & k = L,\ \tilde{k} = 1 \\ 2, & k = L,\ \tilde{k} = 0 \\ 3, & k = H,\ \tilde{k} = 0 \end{cases} \tag{1}$$

[3]The assumption of perfect diagnostic ability is made only to save on notation. Our results also hold when specialists have sufficiently high diagnosis accuracy given their costs.

Figure 1 depicts an overview of the possible cases. In case 1, the PCP diagnoses a low-severity type with certainty (signal $\tilde{k} = 1$). This case arises with probability $(1-p)q$. In cases 2 and 3, the PCP receives signal $\tilde{k} = 0$ and can not distinguish low-severity from high-severity types. Case 2 arises with probability $(1-p)(1-q)$, case 3 with probability p. The PCP uses Bayesian inference to update her beliefs.

The PCP takes on the role of a gatekeeper who receives and diagnoses all patients who seek medical care. There are three possible *treatment paths* $T \in \{P_1, P_2, S\}$ for each case z; the PCP may immediately treat the patient after the diagnosis (P_1), the specialist may treat the patient after the PCP has referred the patient to her (S) or the PCP may treat the patient after the specialist has referred him back (P_2). For now, we assume that the benefits and costs of treatment paths are the same for P_1 and P_2 (excluding the specialist's diagnosis costs) and correspond to b_k^P and c_P. In Subsection 6.2 we consider the impact of waiting time costs.

For each case z, T_z indicates the treatment path of a patient; i.e., a patient of type L with $\tilde{k} = 1$ receives T_1, a patient of type L with $\tilde{k} = 0$ receives T_2, and a patient of type H with $\tilde{k} = 0$ receives T_3. The vector

$$\vec{T} = \begin{pmatrix} T_1 & T_2 & T_3 \end{pmatrix}^T \tag{2}$$

summarizes the treatment paths for all cases z. For example, $\begin{pmatrix} P_1 & S & P_2 \end{pmatrix}^T$ indicates that a patient in case 1 is treated by the PCP after diagnosis, a patient in case 2 is treated by the specialist and a patient in case 3 is referred back by the specialist to the PCP.

Physicians are partially altruistic with $\beta_j \in [0, 1]$ measuring the degree of altruism of a physician of type $j \in \{P, S\}$. For simplicity we assume that the altruism factor is known by the payer. This allows us to derive the optimal type of contract for self-interested and altruistic physicians. Utility is given by a linear combination of the altruistic benefit $\beta_j b$ and the profit from treatment Π_j. These depend on the treatment paths T_z, i.e., on how a patient is treated in case z:

$$U_j^k(T_z) = \beta_j b_k^{T_z} + \Pi_j(T_z),\ j \in \{P, S\},\ k \in \{L, H\} \tag{3}$$

Furthermore, both physicians know their own and the other physicians' degree of altruism β_j. The specialist always knows the state of the patient. Thus, she maximizes (3) with $T_z \in \{S, P_2\}$ for each state in which she receives a referral. The PCP can only choose between treating a patient immediately (P_1) or letting the specialist treat the patient according to the specialist's preferences. If she has identified a low-type patient, she maximizes (3). However, if she can not identify the patient's type, she chooses treatment T^0 so as to maximize her conditional expected utility

$$\mathbf{E}U_P(\tilde{k} = 0) = p_L^0 U_P^L(T^0) + p_H^0 U_P^H(T^0),\ j \in \{P, S\}, \tag{4}$$

where $p_L^0 := \Pr(k = L|\tilde{k} = 0) = \frac{(1-p)(1-q)}{(1-q)(1-p)+p}$, $p_H^0 := (1 - p_L^0)$.

The payer ensures that the physicians are willing to accept their contracts by designing a payment scheme that leads to at least zero (economic) profits Π_j in expectation:

$$\mathbf{E}\Pi_j = q(1-p)\Pi_j(T_1^*) + (1-p)(1-q)\Pi_j(T_2^*) + p\Pi_j(T_3^*) \geq 0,\ j \in \{P, S\}, \tag{5}$$

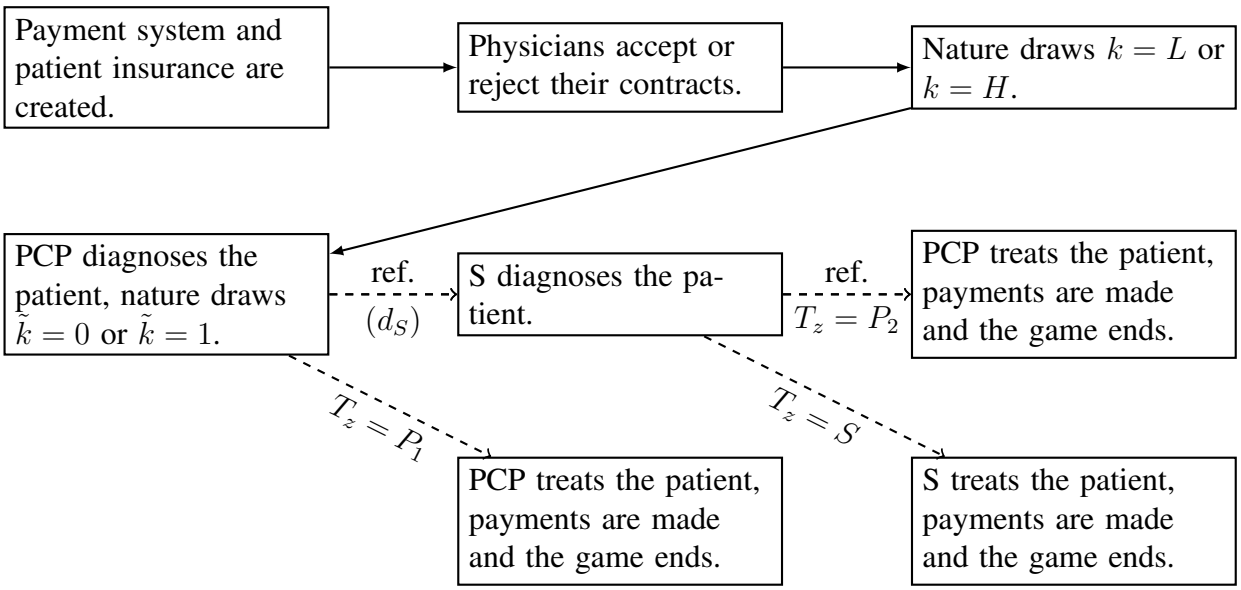


Figure 2: Sequence of Events

where T_z^* is the implemented treatment path in case z.[4]

Patients are fully insured at an actuarially fair premium. They passively follow their physicians' recommendations. Finally, we assume that for all patients at least receiving PCP treatment always confers a greater benefit to the patient than the costs of treatment, i.e. $b_k^P > c_P$ for all k.

The payer is assumed to maximize patient welfare which is given by the expected difference between patient benefits and the sum of the physicians' profits and the costs for treatment and diagnosis,

$$\mathbf{E}W = \mathbf{E}b - \mathbf{E}\Pi - \mathbf{E}c. \tag{6}$$

Figure 2 displays the sequence of events. First, the payer designs the payment scheme. Physicians accept the payment scheme if their zero profit constraint is met in expectation. If the PCP refuses, the game ends. If only the specialist refuses, the PCP will treat all patients as long as her participation constraint is still fulfilled. Afterwards, nature draws the type of patient. The PCP diagnoses the patient and can either treat the patient herself or refer the patient to the specialist. If the patient is referred to the specialist, the patient is diagnosed again regardless of whether the PCP detected the type or not. The specialist decides to treat the patient or to refer the patient back to the PCP. Neither physician has prior information about the type of the patient.

[4]Following Liu and Ma (2013), Liu, Ma, and Mak (2018), and Olivella and Siciliani (2017), we assume that the zero profit constraint is sufficient to guarantee participation.

4 The First-Best Solution

In this section, we derive the first-best treatment paths $\vec{T}^{FB}$ which maximize patient welfare (6) subject to the physicians' participation constraints and assuming that the diagnostic outcome is known to the payer. The first-best solution corresponds to a setting in which the payer can design a contract contingent on diagnostic outcomes. In this case, any treatment path vector which is compatible with the participation constraints can be implemented. No rent accrues to the physicians, i.e. $\mathbf{E}\Pi = 0$.[5]

In Appendix A.1, we prove Theorem 1 which derives the first-best treatment path vectors.

Theorem 1. *The first-best vector of treatment paths $\vec{T}^{FB}$ is given by*

$$\vec{T}^{FB} = \begin{cases} (P_1\ P_1\ P_1)^T \textit{ if} & p_L^0\kappa_L + p_H^0\kappa_H \leq d_S + c_S - c_P, \\ & \kappa_H \leq c_S - c_P + d_S/p_H^0; \\ (P_1\ P_2\ S)^T \textit{ if} & \kappa_H \geq c_S - c_P + d_S/p_H^0, \\ & \kappa_L \leq c_S - c_P; \\ (P_1\ S\ S)^T \textit{ if} & c_S - c_P \leq \kappa_L \leq d_S + c_S - c_P, \\ & p_L^0\kappa_L + p_H^0\kappa_H \geq d_S + c_S - c_P; \\ (S\ S\ S)^T \textit{ if} & \kappa_L \geq d_S + c_S - c_P. \end{cases} \tag{7}$$

Figure 3 shows the optimal treatment path vector $\vec{T}^{FB}$ depending on the specialist surplus benefits. We assume $\kappa_H > \kappa_L$ because it seems reasonable that high-severity patients would benefit more from the specialist's greater sophistication than low-severity patients. Thus, we only consider the contracts of κ_L and κ_H below the $45°$-line. Here we obtain the following results:

- Treatment path $(P_1\ P_1\ P_1)^T$ is optimal if both κ_L and κ_H are sufficiently small. In this case, it is optimal to have the PCP treat both patient types because the benefits from specialist treatment are small.
- Treatment path $(P_1\ P_2\ S)^T$ is optimal for large values of κ_H and small values of κ_L. In this case, it is efficient for the specialist to treat high-type patients while referring back low-type patients who were not detected by the PCP. The cost savings of referring back a low-type patient outweigh the forgone patient benefits.

[5] Since the gain in expected patient welfare from assigning one patient to some treatment path is independent of the other patients, any first-best optimal solution will assign the same treatment path to every patient in the same case z.

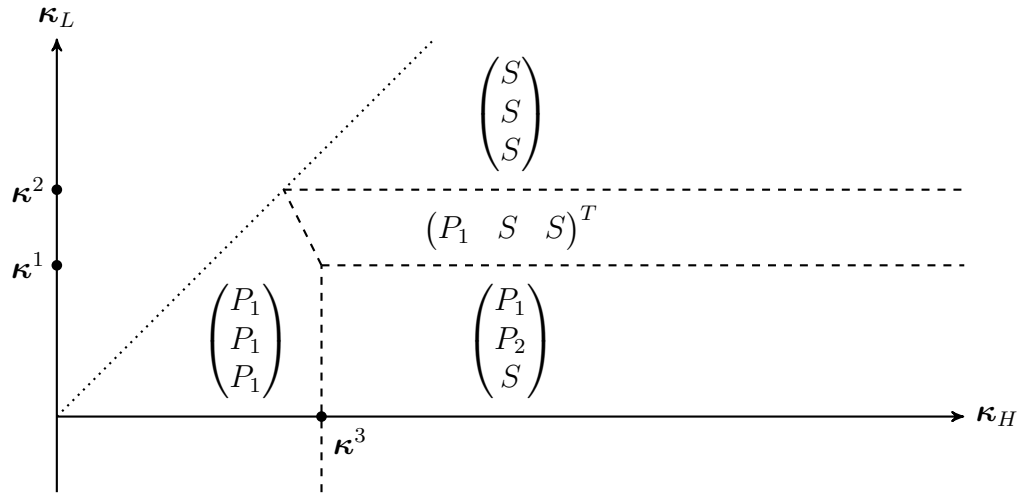


Figure 3: First-best (dashed lines) treatment path vector depending on the specialist surplus benefit. Dotted line indicates $\kappa_H = \kappa_L$. $\kappa^1 = c_S - c_P$, $\kappa^2 = d_S + c_S - c_P$, $\kappa^3 = c_S - c_P + d_S/p_H^0$.

- Treatment path $(P_1\ S\ S)^T$ is optimal if κ_H is sufficiently large and κ_L is larger than $\kappa^1 = c_S - c_P$ but smaller than $\kappa^2 = d_S + c_S - c_P$. In this case it is efficient for the PCP to treat detected low-type patients and for the specialist to treat both undetected types. Detected low-types would incur additional diagnostic costs if they were treated by the specialist whereas the diagnostic costs of the low-types who have only been detected by the specialist are already sunk.
- Treatment path $(S\ S\ S)^T$ is optimal if κ_L, and therefore κ_H, is larger than $\kappa^2 = d_S + c_S - c_P$. It is efficient to have the specialist treat both patient types because the patient benefits for either type are larger than the additionally incurred costs of specialist diagnosis and treatment.

5 Private Diagnostic Signals

We now turn to the second-best problem of incentivizing physicians when the payer can not observe the diagnostic signals of the physicians. However, the payer can verify whether a patient has visited a physician and which treatment was provided. The payments to the physicians γ_j^T can therefore be made contingent on the treatment path $T \in \{P_1, P_2, S\}$. With these payments, capitations and FFS payments, as well as any mix of the two, can be implemented by the payer. For a capitation, the payer needs to set $\gamma_P^{P_1} = \gamma_P^{P_2} = \gamma_P^S$ for the PCP and $\gamma_S^{P_2} = \gamma_S^S$ for the specialist. Under a payment with a FFS component physician activity gets rewarded: $\gamma_P^{P_1/P_2} > \gamma_P^S, \gamma_S^S > \gamma_S^{P_2}$. Furthermore, the payer can let the PCP share a part of the specialist's diagnosis costs by setting $\gamma_P^{P_2} < \gamma_P^{P_1}$.

In the following, contracts are derived which implement the candidate first-best treatment path vectors from Equation (7) at minimal rents for the physicians. We impose the *non-negative payments* condition

$$\text{NNP: } \gamma_j^T \geq 0,\ j \in \{P, S\},\ T \in \{P_1, P_2, S\}. \tag{8}$$

We first consider $(P_1\ P_2\ S)^T$.[6] First, the PCP decides whether to treat or refer the known low-type patients ($z = 1$) and the patients of unknown type ($z = 2, 3$). Afterwards, the specialist learns the real type of the patients and chooses to treat or refer back those low- and high-type patients who she received from the PCP. The PCP correctly anticipates the behavior of the specialist and adjusts her behavior accordingly. Hence, the problem gets solved through backward induction. We assume that physicians choose the patient welfare-maximizing option, whenever they are indifferent between two options. $(P_1\ P_2\ S)^T$ can be implemented by fulfilling System of Inequations (9).

$$\begin{aligned}
IC_P^1 &: \gamma_P^{P_1} \geq \gamma_P^{P_2} \\
IC_P^2 &: \gamma_P^{P_1} - c_P \leq p_H^0(\beta_P \kappa_H + \gamma_P^S) + p_L^0(\gamma_P^{P_2} - c_P) \\
PC_P &: (1-p)q(\gamma_P^{P_1} - c_P) + (1-p)(1-q)(\gamma_P^{P_2} - c_P) \\
&\quad + p\gamma_P^S \geq 0 \\
IC_S^1 &: \beta_S \kappa_H + \gamma_S^S - c_S \geq \gamma_S^{P_2} \\
IC_S^2 &: \beta_S \kappa_L + \gamma_S^S - c_S \leq \gamma_S^{P_2} \\
PC_S &: (1-p)(1-q)(\gamma_S^{P_2} - d_S) + p(\gamma_S^S - c_S - d_S) \geq 0 \\
NNP &: \gamma_j^T \geq 0
\end{aligned} \tag{9}$$

For low-type patients, the PCP needs to treat patients without referral rather than after a back-referral from the specialist. This is ensured by IC_P^1. IC_P^2 guarantees that patients of unknown type are not immediately treated by the PCP. Instead, the PCP needs to treat low-type patients after a back-referral by the specialist and not treat high-type patients. IC_S^1 ensures that the specialist treats high-type patients, IC_S^2 that she refers back low-type patients. PC_S and PC_P represents the physicians' participation constraints.

In the following we will derive a contract that implements $(P_1\ P_2\ S)^T$ without rents accruing for the physicians. Consider first incentives for the specialist. Figure 4 visualizes how proper incentives can be given under the assumption that the PCP refers unknown types only. It shows the incentives constraints of the specialist depending on the remuneration for back-referral and treatment. To meet the incentives constraint, $\gamma_S^{P_2}$ and γ_S^S must be between the dashed lines. Then, the specialist prefers to treat high-types (IC_S^1 fulfilled) and refers back low-types (IC_S^2

[6] The indiscriminate outcomes can easily be implemented without rent payments, $(P_1\ P_1\ P_1)^T$ by not paying the specialist and paying the PCP according to her participation constraint and $(S\ S\ S)^T$ by paying both providers by a FFS payment that just covers their expected costs. In the following, we therefore focus on the implementation of the treatment paths $(P_1\ P_2\ S)^T$ and $(P_1\ S\ S)^T$.

fulfilled). In order to maximize welfare, rent payments should be kept as low as possible. Optimally, the payer wants to leave no rent to the physicians. Hence, the optimal contract lies on the specialist's zero profit line $\mathbf{E}[\Pi_S] = 0$. The solid lines marked by γ_S^* in Figure 4 contain all contracts for the specialist that implement $(P_1\ P_2\ S)^T$ without rent.

$(P_1\ P_2\ S)^T$ can be implemented by a cost-based FFS contract with $\gamma_S^{P_2} = d_S$ and $\gamma_S^S = d_S + c_S$ for the specialist if her surplus benefit $\beta_S \kappa_L$ from treatment for low-type patients is negative. This is displayed in Figure 4a. IC_S^2 is met. The intuition is that under cost-based payment, an altruistic specialist does not need to be incentivized to refer patients back if her treatment is inferior for low-type patients. By contrast, if the specialist is altruistic and $\kappa_L > 0$, it is necessary to pay a markup for referring the patient back to counter the altruistic incentive (Figure 4b). IC_S^2 has shifted downward as the specialist now needs to paid less to treat the patient. In order to extract the rent, $\gamma_S^{P_2}$ must be increased and γ_S^S reduced which calls for a markup to the specialist for referring patients and cost sharing for treatment.

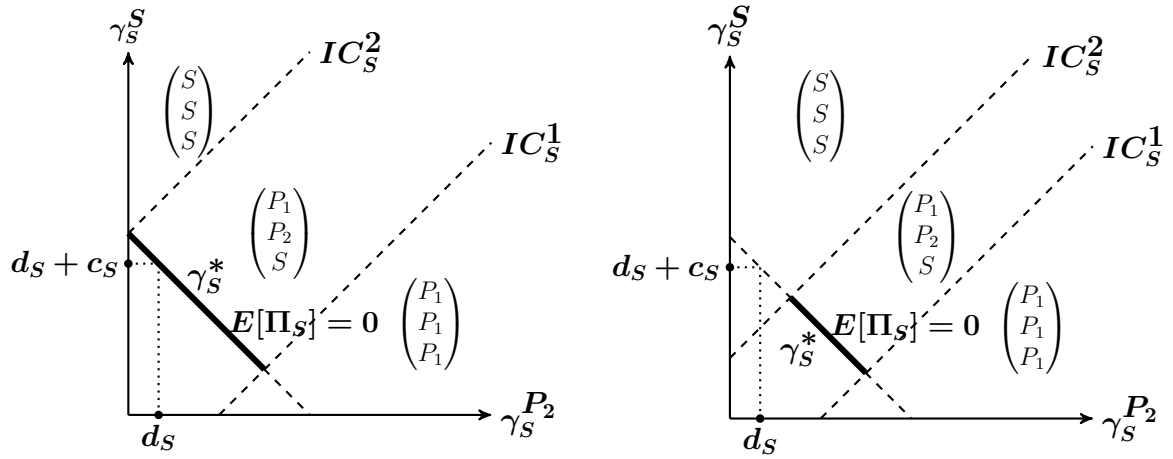
We now turn to the PCP's decision assuming that the specialist treats high-type patients and refers back low-type patients. The PCP expects this and will, therefore, not refer low-type patients who she detected (case $z = 1$) to the specialist if she doesn't benefit from delayed treatment (i.e. $\gamma_P^{P_1} \geq \gamma_P^{P_2}$). Thus, if the specialist is properly incentivized, the PCP can be paid with a simple cost-based FFS payment $(\gamma_P^{P_1}, \gamma_P^{P_2}, \gamma_P^S) = (c_P, c_P, 0)$. Since the PCP only cares about patient benefits under this contract, she will refer all unknown-type patients.

Theorem 2 (proof in Appendix A.2) shows that for $\kappa_L \leq \kappa^2 = d_S + c_S - c_P$ it is always possible to implement $(P_1\ P_2\ S)^T$ without rent and without violating the non-negative payment constraints. It is uninteresting to consider $\kappa_L > \kappa^2$, since then $(S\ S\ S)^T$ is first-best and second-best optimal (see Figure 3).

Theorem 2. *Let $\kappa_L \leq \kappa^2$. The following contract implements $(P_1\ P_2\ S)^T$ without a rent payment for either physician:*

$$
\begin{aligned}
\gamma_P^{P_1 *} &= c_P \\
\gamma_P^{P_2 *} &= c_P \\
\gamma_P^{S *} &= 0 \\
\gamma_S^{P_2 *} &= d_S + p_H^0 \max(\beta_S \kappa_L, 0) \\
\gamma_S^{S *} &= d_S + c_S + p_L^0 \min(-\beta_S \kappa_L, 0).
\end{aligned}
$$

Observe that the contract from Theorem 2 includes the minimum $\gamma_S^{P_2}$ such that the specialists incentive constraints are fulfilled. However, if $\beta_S \kappa_L > 0$, $\gamma_S^{P_2 *}$ is larger than the diagnostic costs of the specialist. The altruistic specialist thus needs to be paid a markup in order to not (inefficiently) treat a low-severity patient, whereas the PCP can be simply paid back her costs. The health economic literature on referrals has focused mostly on incentives for the



(a) $\beta_S \kappa_L < 0$ (b) $\beta_S \kappa_L > 0$

Figure 4: Second-best optimal contracts without rent payments to the specialist that implement $(P_1\ P_2\ S)^T$ given $(\gamma_P^{P_1}, \gamma_P^{P_2}, \gamma_P^S) = (c_P, c_P, 0)$. Dashed lines indicate incentive- and participation constraints, the solid and thick line indicates the set of optimal contracts.

PCP and assumed the specialist to behave independently of economic incentives (see for example Allard et al., 2011 and Garcia-Mariñoso and Jelovac, 2003). This result highlights that agency problems can also exist on the specialist side when it comes to the optimal allocation of treatments.

Finally, we derive an optimal contract that implements $(P_1\ S\ S)^T$. This treatment path can be implemented by fulfilling System of Inequations (10).

$$\begin{aligned}
IC_P^1 &: \gamma_P^{P_1} - c_P \geq \beta_P \kappa_L + \gamma_P^S \\
IC_P^2 &: \gamma_P^{P_1} - c_P \leq \beta_P (p_H^0 \kappa_H + p_L^0 \kappa_L) + \gamma_P^S \\
PC_P &: (1-p)q(\gamma_P^{P_1} - c_P) + [(1-p)(1-q) + p]\gamma_P^S \geq 0 \\
IC_S^1 &: \beta_S \kappa_H + \gamma_S^S - c_S \geq \gamma_S^{P_2} \\
IC_S^2 &: \beta_S \kappa_L + \gamma_S^S - c_S \geq \gamma_S^{P_2} \\
PC_S &: \gamma_S^S - c_S - d_S \geq 0 \\
NNP &: \gamma_j^T \geq 0
\end{aligned} \tag{10}$$

The PCP needs to treat low-type patients, rather than referring them for specialist treatment (IC_P^1) and needs to refer patients of unknown type for specialist treatment, rather than treating them herself (IC_P^2). The specialist needs to treat all patient types who get referred to her. This is represented by IC_S^1 and IC_S^2. Furthermore, participation constraints and non-negative payment constraints need to be fulfilled for both physicians.

If $\kappa_L < 0$, $(P_1\ S\ S)^T$ is not first-best optimal (see Figure 3). Furthermore, $(P_1\ P_1\ P_1)^T$ and $(P_1\ P_2\ S)^T$ can be implemented without rent. Therefore, we assume $\kappa_L \geq 0$ in Theorem 3 (proof is in Appendix A.3).

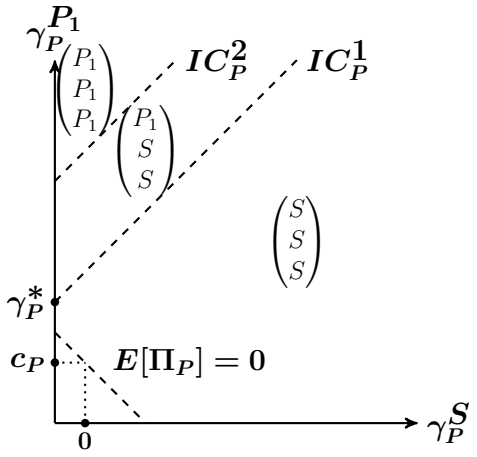


Figure 5: Second-best optimal contract for the PCP that implements $(P_1 \ S \ S)^T$ given $(\gamma_S^{P_2}, \gamma_S^S) = (d_S, d_S + c_S)$ and $\beta_P \kappa_L > 0$. Dashed lines indicate incentive- and participation constraints.

Theorem 3. *Let $\kappa_L \geq 0$. The contract*

$$\begin{aligned} \gamma_P^{P_1 *} &= c_P + \beta_P \kappa_L \\ \gamma_P^{P_2 *} &= 0 \\ \gamma_P^{S *} &= 0 \\ \gamma_S^{P_2 *} &= d_S \\ \gamma_S^{S *} &= d_S + c_S \end{aligned} \tag{11}$$

is the unique rent-minimizing contract (except for variations in $\gamma_P^{P_2}$ and $\gamma_S^{P_2}$) with rents of $(1-p)q\beta_P\kappa_L$ for the PCP.

Theorem 3 shows that the specialist can be incentivized to provide treatment to all patients by a cost-based FFS contract, whenever $\kappa_L \geq 0$. Given her altruistic orientation, she will always treat each patient. By contrast, a markup to the PCP for treating the patient without referral is necessary in order to implement $(P_1 \ S \ S)^T$ because the altruistic PCP needs to be prevented from over-referring the known low-types to the specialist. This markup leads to an information rent for the PCP.

Figure 5 shows the optimal contract. To fulfill the incentive constraints, γ_P^S and $\gamma_P^{P_1}$ must be between the incentive constraints. Then the PCP prefers to treat low-types (IC_P^1 fulfilled) and to refer patients of unknown type (IC_P^2 fulfilled). The rent-minimizing contract sets $\gamma_P^S = \gamma_P^{P_2} = 0$ and $\gamma_P^{P_1}$ to the minimum level that still meets incentive constraint IC_P^1.

The rent payment to the PCP rises in her degree of altruism. If the PCP is not altruistic at all, no incentive problem exists and the first-best can be implemented by a cost-based FFS contract.

According to the previous literature on physician payment a highly altruistic physician should share a large portion of her incurred costs in order to not overtreat the patient (see Ellis and McGuire, 1986; Chalkley and Malcomson, 1998). By contrast, in our setting a PCP should receive a markup on her immediate treatment costs, in order to not over-refer the patient to more expensive specialist care.

As a rent payment is necessary to implement the treatment path $(P_1\ S\ S)^T$, it is not clear that implementing this path is optimal in the second-best. Theorem 4 shows that for highly altruistic PCPs switching to other treatment paths is superior (details and the proof in Appendix A.4).

Theorem 4. *If* $0 < \beta_P \leq \beta_P^* := \frac{(1-q)d_S}{c_S - c_P + qd_S}$, *the second-best region in which implementing* $(P_1\ S\ S)^T$ *is optimal is reduced compared to the first-best.*

If $\beta_P > \beta_P^*$, *implementing* $(P_1\ S\ S)^T$ *is not second-best optimal.*

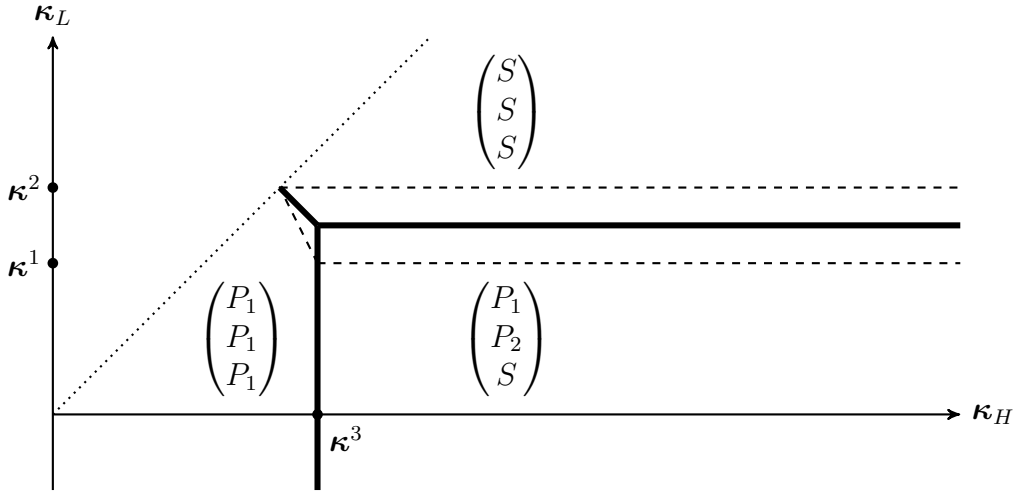


Figure 6: Second-best (solid and thick) vs. first-best (dashed) treatment path vector depending on the specialist surplus benefits, $\beta_P \geq \beta_P^*$.

Figure 6 visualizes the case in which implementing $(P_1\ S\ S)^T$ is not second-best optimal due to an excessive information rent. If both κ_L and κ_H are sufficiently large, letting the specialist treat all cases is second-best and if both are sufficiently small, letting the PCP treat all cases is second-best. Finally, if κ_L is sufficiently small and κ_H is sufficiently large, having the PCP treat the known low-types immediately and the unknown low-types later and the specialist treat the high-types is second-best.

6 Extensions

In this section, we explore the impact of a change in the model's assumptions. In Subsection 6.1 we drop the assumption that the payer has perfect knowledge about the size of the surplus benefits κ_k. In Subsection 6.2, we study the effect of referral costs for the patient. In Subsection 6.3, we explore an alternative assumption on the information structure of the game.

6.1 Uncertain Surplus Benefits

The second-best optimal payment contracts derived in Section 5 require the payer to know the exact size of the surplus benefits that accrue to patients from specialist treatment. More realistically, the payer's knowledge is uncertain because patients of the same type may still differ in the degree they benefit from specialist treatment. For example, physicians are more aware of the patient's medical history than the payer. Therefore, the payer may only have access to some probabilistic distribution function over the space (κ_H, κ_L). This is what we assume in this section and call the *third-best* problem. Diagnostic signals remain private information of the physicians.

We restrict our analysis to the contract designs from Section 5:

(I) The cost-based FFS contract:

$$(\gamma_P^{P_1}, \gamma_P^{P_2}, \gamma_P^S, \gamma_S^{P_2}, \gamma_S^S) = (c_P, c_P, 0, d_S, d_S + c_S)$$

(II) FFS + markup for the specialist's back-referral:
$\gamma_S^{P_2} = d_S + p_H^0 m_S, \gamma_S^S = d_S + c_S - p_L^0 m_S,\ m_S > 0$

(III) FFS + markup for immediate PCP treatment:
$\gamma_P^{P_1} = c_P + m_P, \gamma_P^{P_1} = \gamma_P^{P_2} = 0\ m_P > 0$

Contract (I) can be viewed as a benchmark contract against which to compare the other contracts. Lemma 1 in Appendix A.5 describes the physicians' behavior under FFS payment plus markups. We assume $\beta_P, \beta_S > 0$ in order to have an interesting problem. Otherwise, the cost-based FFS contract implements the first-best without rents for either provider. Figure 7 compares the resulting treatment path vectors for the cost-based FFS contract with the first-best. Shaded areas indicate that the contract's treatment path vector is first-best, $\mathbf{OT^j}$ indicates too much specialist treatment for $j \in \{1, 2, 3\}$ of the cases. Clearly, cost-based FFS contracts incentivize over-treatment for many distributions over the (κ_H, κ_L) space. Markups for specialist referrals can help to alleviate over-treatment. Figure 8 shows the impact of the introduction of a small markup for specialist back-referral. As explained in Section 5, markups for the specialist's referral work both directly and indirectly. The direct effect is that the specialist is more willing to refer back low-type patients to the PCP. Consequently, the region in which $(P_1\ P_2\ S)^T$ is played expands to $\kappa_L \leq \frac{m_S}{\beta_S}$. The indirect effect is that the PCP predicts that the specialist will not treat some low-type patients, which incentivizes the PCP to treat

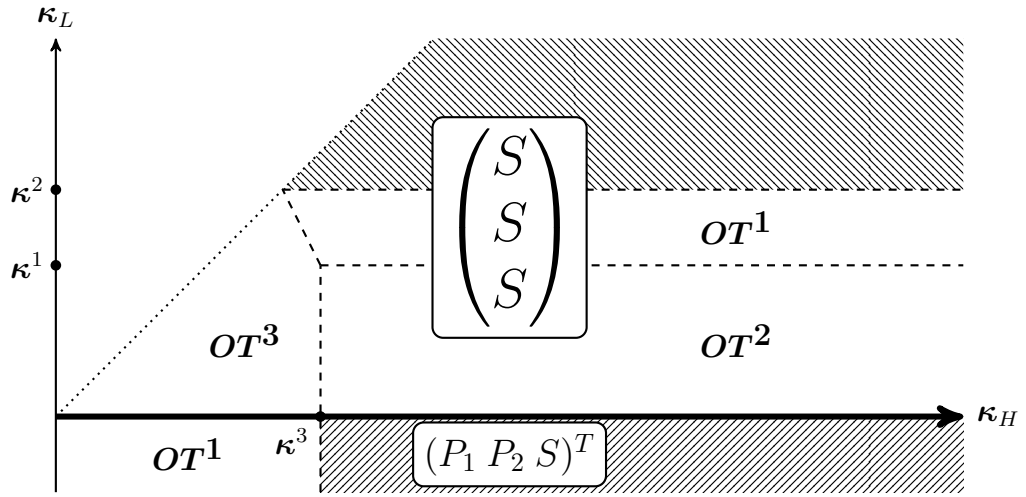


Figure 7: Contract (I) – Cost-based FFS (solid and thick) vs first-best (dashed) treatment path vectors.

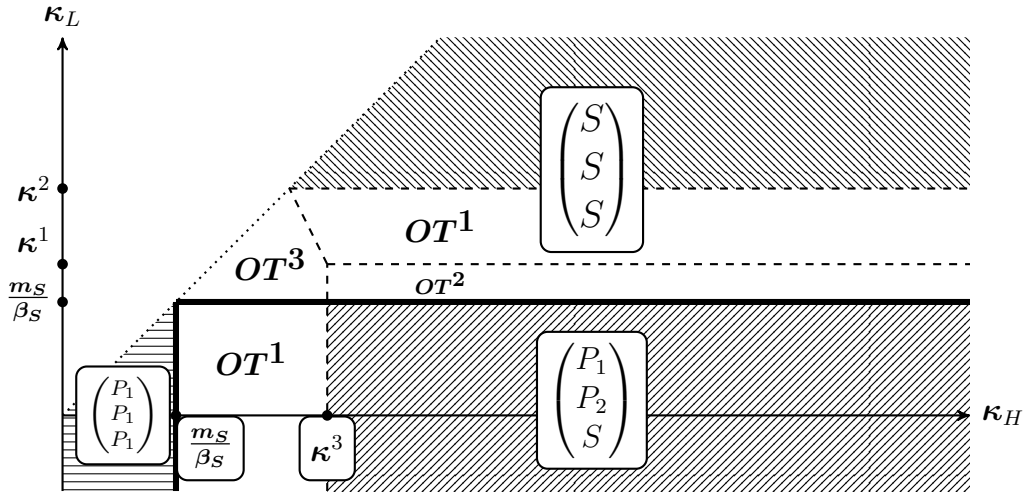


Figure 8: Contract (II) – Treatment path vectors resulting from a markup m_S for specialist referral (solid and thick) vs first-best (dashed) treatment path vectors. Shaded regions indicate congruence between first-best and treatment paths resulting from markup.

these patients immediately. Thus, there now exists a region in which $(P_1\ P_1\ P_1)^T$ is played for $\kappa_H \leq \frac{m_S}{\beta_S}$.

Comparing Figure 8 to Figure 7, it is evident that social welfare is improved over the cost-based FFS contract by reducing over-referral. Theorem 5 shows that this holds true for any sufficiently small markup (proof in Appendix A.6).

Theorem 5. *If $\frac{m_S}{\beta_S} \leq qd_S + c_S - c_P$, contract design (II) is weakly superior to contract (I) with regards to expected patient welfare for any distribution over (κ_H, κ_L).*

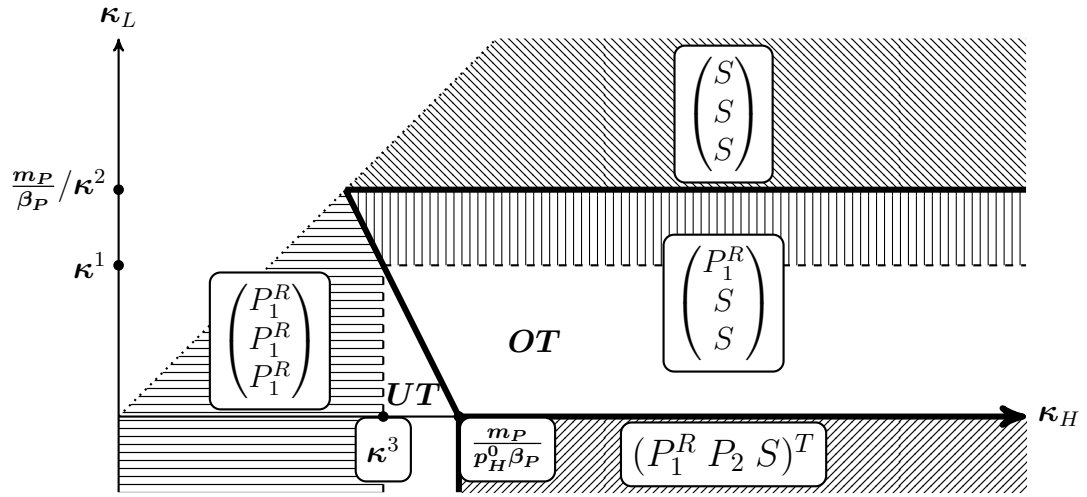


Figure 9: Contract (III) – Markup for PCP treatment without referral (solid and thick) vs first-best (dashed) treatment path vectors.

The intuition behind Theorem 5 is as follows. The first patients who get referred back under contract (II) and who would not have been referred back under contract (I) are those patients who cannot be identified by the PCP and only minimally benefit from specialist treatment. Hence, even for small markups cost savings can be made without significantly reducing patient benefit. Since specialist markups do not incur a rent, patient welfare is weakly larger than under the cost-based FFS contract.

Alternatively, the PCP can be incentivized directly to not refer the patient by paying her a markup for treating the patient immediately (contract design (III)). Figure 9 ($\boldsymbol{UT}$ denotes under-treatment, R denotes rents) depicts the resulting treatment path vectors for a markup m_P. The result differs from contract (I) in two aspects. First, for small κ_L and κ_H, the PCP treats all patient types, second, treatment path $(P_1\ S\ S)^T$ is played for $\kappa_L \leq \frac{m_P}{\beta_P}$ and large κ_H. Intuitively, if she is paid a markup for treatment, the PCP treats patients that would only benefit little from specialist treatment. Thus, comparing Figure 7 to Figure 9 shows that adding markups for the PCP to contract (I) also reduces over-referral.

However, there are two drawbacks with this approach. Since contract design (III) relies only on direct rather than indirect incentivization, rent payments accrue for the PCP when she treats a patient type, whereas in design (II) no rent accrues to the specialist. Hence, for regions (κ_H, κ_L), in which the same path is implemented for both contracts, design (II) is superior to design (III). Another drawback is that $(P_1\ P_2\ S)^T$ can not be implemented. Thus, for the region in which this treatment path is optimal markups for the specialist are superior.

Similarly to small markups m_S, small markups m_P improve the allocation of treatments. However, small markups for the PCP do not always improve welfare if $\kappa_L \leq 0$ due to rents. For $\kappa_L > 0$, expected welfare is weakly improved. This is demonstrated in Theorem 6 (proof in Appendix A.7).

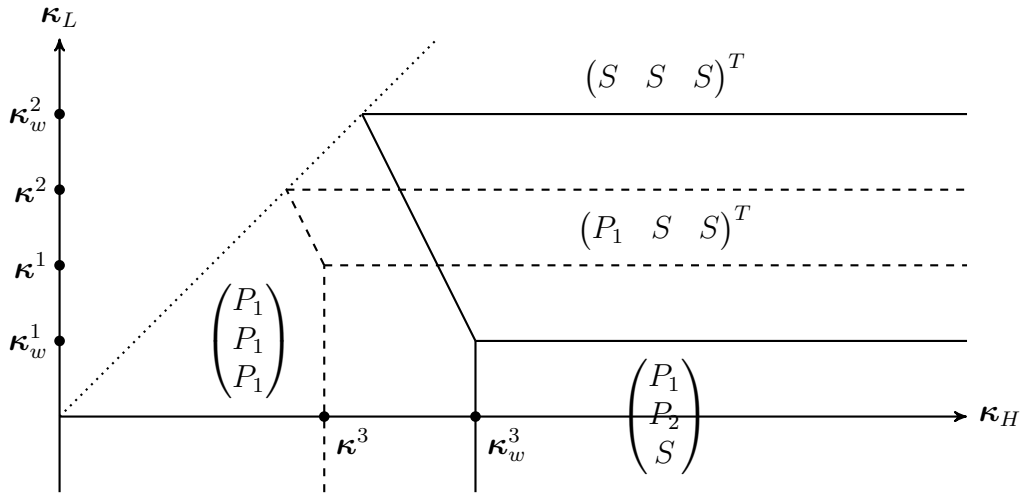


Figure 10: First-best treatment path vector with (regular lines) and without (dashed lines) waiting costs depending on the specialist surplus benefit. Dotted line indicates $\kappa_H = \kappa_L$. $\kappa_w^1 := c_S - c_P - w$, $\kappa_w^2 := d_S + c_S - c_P + w$, and $\kappa_w^3 := d_S/p_H^0 + (\frac{2p_L^0}{p_H^0} + 1)w + c_S - c_P$.

Theorem 6. *a) If $\frac{m_P}{\beta_P} \leq d_S + p_H^0(c_S - c_P)$, the allocation of treatments (ignoring rents) under contract design (III) is weakly superior to contract (I) for any distribution over (κ_H, κ_L).*

b) If $\frac{m_P}{\beta_P} \leq \frac{\kappa^2}{1+\beta_P}$, contract design (III) is weakly superior to contract (I) with regards to expected patient welfare for any distribution over (κ_H, κ_L) with $\kappa_L > 0$.

Concluding, whereas small markups for the specialist always have non-negative welfare effects, markups for the PCP may have negative effects because they lead to information rents for the PCP. Small PCP markups have non-negative welfare effects if the benefit of specialist treatment for L-types is positive.

6.2 Cost of Referrals

So far, we assumed that referring a patient from one physician to the other is costless. More realistically, however, referrals are associated with additional waiting costs for the patients. We model this by subtracting a waiting cost $w > 0$ from the patients utility if he is treated by the specialist and $2w$ if he is treated after a back-referral to the PCP. The result of this modification can be seen in Figure 10. Here, the first-best treatment path vectors are depicted with and without waiting costs. The definition of κ_L and κ_H is the same as before (difference in treatment benefits without waiting costs). For $(P_1 \quad P_1 \quad P_1)^T$ there are no waiting costs of the patient. Therefore the area in which this treatment path is optimal grows vis-à-vis all

other paths. For $\begin{pmatrix} P_1 & S & S \end{pmatrix}^T$ there are less waiting costs compared to both $\begin{pmatrix} S & S & S \end{pmatrix}^T$ and $\begin{pmatrix} P_1 & P_2 & S \end{pmatrix}^T$. Therefore, the area in which $\begin{pmatrix} P_1 & S & S \end{pmatrix}^T$ is optimal grows as long as κ_L and κ_H are large enough. For small κ_L and κ_H the boundary between $\begin{pmatrix} P_1 & P_1 & P_1 \end{pmatrix}^T$ and $\begin{pmatrix} P_1 & S & S \end{pmatrix}^T$ shifts to the right due to the higher waiting costs for unknown patient types.

Implementation in the second-best is the same for the blind treatments. To implement $\begin{pmatrix} P_1 & P_2 & S \end{pmatrix}^T$ with waiting time costs, the following incentive constraints change compared to (9):

$$\begin{aligned} &IC_P^{1,w} : \gamma_P^{P_1} \geq \gamma_P^{P_2} - 2\beta_P w \\ &IC_P^{2,w} : \gamma_P^{P_1} - c_P \leq p_L^0(\gamma_P^{P_2} - 2\beta_P w - c_P) + p_H^0[\gamma_P^S + \beta_P(\kappa_H - w)] \\ &IC_S^{1,w} : \gamma_S^S + \beta_S \kappa_H - c_S \geq \gamma_S^{P_2} - \beta_S w \\ &IC_S^{2,w} : \gamma_S^S + \beta_S \kappa_L - c_S \leq \gamma_S^{P_2} - \beta_S w \end{aligned} \tag{12}$$

Incentivizing the PCP to keep low-type patients is easier due to the waiting costs, whereas referring patients of unknown type is more costly to the altruistic PCP. However, a cost-based FFS contract still implements the appropriate referral behavior from the PCP since the first-best region in which $\begin{pmatrix} P_1 & P_2 & S \end{pmatrix}^T$ is implemented shrinks. Similarly, it is easier to incentivize the specialist to treat high-type patients and it is more costly for specialists to refer back low-type patients. Thus, the contract from Theorem 2 has to be amended to pay larger markups to the specialist for referring the patient back and correspondingly larger cost sharing when treating the patient. This does not lead to a rent-payment for the specialist in the region in which $\begin{pmatrix} P_1 & P_2 & S \end{pmatrix}^T$ is first-best. For details see Theorem 7 (proof in Appendix A.8).

Theorem 7. *Let* $\begin{pmatrix} P_1 & P_2 & S \end{pmatrix}^T$ *be first-best optimal. To implement* $\begin{pmatrix} P_1 & P_2 & S \end{pmatrix}^T$ *the contract from Theorem 2 has to be amended by paying larger markups to the specialist for referring the patient back and correspondingly larger cost sharing when treating the patient:*

$$\begin{aligned} \gamma_P^{P_1 *} &= c_P \\ \gamma_P^{P_2 *} &= c_P \\ \gamma_P^{S *} &= 0 \\ \gamma_S^{S *} &= d_S + c_S + p_L^0 \min(-\beta_S(\kappa_L + w), 0) \\ \gamma_S^{P_2 *} &= d_S + p_H^0 \max(\beta_S(\kappa_L + w), 0) \end{aligned} \tag{13}$$

To implement $\begin{pmatrix} P_1 & S & S \end{pmatrix}^T$ with waiting time costs, the following incentive constraints change

compared to (10):

$$\begin{aligned}
&IC_P^{1,w}: \gamma_P^{P_1} - c_P \geq \gamma_P^S + \beta_P(\kappa_L - w) \\
&IC_P^{2,w}: \gamma_P^{P_1} - c_P \leq \gamma_P^S + \beta_P[p_L^0(\kappa_L - w) + p_H^0(\kappa_H - w)] \\
&IC_S^{1,w}: \gamma_S^S + \beta_S\kappa_H - c_S \geq \gamma_S^{P_2} - \beta_S w \\
&IC_S^{2,w}: \gamma_S^S + \beta_S\kappa_L - c_S \geq \gamma_S^{P_2} - \beta_S w
\end{aligned} \tag{14}$$

Fulfilling $IC_P^{1,w}$ exactly, fulfills $IC_P^{2,w}$ with the smallest possible $\gamma_P^{P_1}$. Since patients suffer from waiting costs, it is now easier to incentivize the PCP to not refer low-type patients for the same surplus benefits. The incentive constraints for the specialist can still be fulfilled by a cost-based FFS payment since additional waiting costs make it even less beneficial for the altruistic specialist to refer patients. Thus, the contracts from Theorem 3 have to be amended to pay smaller markups to the PCP for treating the patient immediately. If waiting costs exceed the surplus benefit of treating low-type patients, no rents need to be paid at all. In this case, $\begin{pmatrix} P_1 & S & S \end{pmatrix}^T$ can be implemented with a cost-based FFS contract since the PCP is not motivated to over-refer L-type patients. Furthermore, if rents are paid, they are smaller for a given κ_L. For details see Theorem 8 (proof in Appendix A.9).

Theorem 8. *Let* $\begin{pmatrix} P_1 & S & S \end{pmatrix}^T$ *be first-best optimal. To implement* $\begin{pmatrix} P_1 & S & S \end{pmatrix}^T$ *with waiting time costs the contracts from Theorem 3 have to be amended by paying smaller markups to the PCP for treating the patient immediately. If* $w \geq \kappa_L$*, the cost-based FFS contract implements* $(P_1\ S\ S)^T$ *without rent payments.*

If $w < \kappa_L$*, the contract*

$$\begin{aligned}
\gamma_P^{P_1 *} &= c_P + \beta_P(\kappa_L - w) \\
\gamma_P^{P_2 *} &= 0 \\
\gamma_P^{S *} &= 0 \\
\gamma_S^{S *} &= d_S \\
\gamma_S^{P_2 *} &= d_S + c_S
\end{aligned} \tag{15}$$

is the unique rent-minimizing contract (except for variations in $\gamma_P^{P_2}$ *and* $\gamma_S^{P_2}$*) with rents of* $(1-p)q\beta_P(\kappa_L - w)$ *for the PCP.*

Thus, in the second-best implementation, only the second-best region in which $\begin{pmatrix} P_1 & S & S \end{pmatrix}^T$ is implemented shrinks compared to the first-best region. Nevertheless, the second-best region in which the treatment path is optimal is larger than in the case without waiting costs.

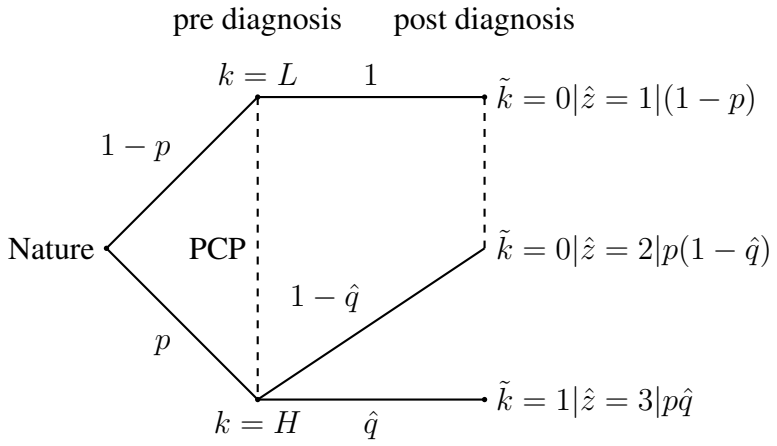


Figure 11: Cases z if the PCP can identify high-type patients, dashed lines indicate that the PCP can not differentiate between these cases.

Concluding, the set of optimal contracts when waiting costs are a factor is the same as the set of optimal contracts without waiting costs (cost-based FFS, markups for immediate PCP treatment, and markups for specialist back-referral). However, as long as κ_H is large enough, it is more likely to be optimal both in the first-best and second-best to only have the PCP discriminate based on her diagnostic signal. Furthermore, if waiting costs w are larger than κ_L, no markup is required for the PCP to implement this treatment path.

6.3 PCP Diagnosis Identifies High-Severity Patients

So far, we assumed that the PCP can identify a fraction of low-severity patients perfectly. Now we consider an alternative information structure in which the PCP can identify a fraction $\hat{q}$ of the high-severity patients but no longer identifies low-severity patients. This case is relevant if the symptoms of a severe case provide strong evidence to the underlying disease of the patient whereas a patient without these symptoms may still be severely ill. Three cases $\hat{z}$ are possible:

1. The patient is of low type and the PCP can not identify him.
2. The patient is of high type and the PCP can not identify him.
3. The patient is of high type and the PCP can identify him.

Figure 11 depicts the new diagnostic process. Two new treatment path vectors emerge that can be first-best optimal, namely $\begin{pmatrix} P_1 & P_1 & S \end{pmatrix}^T$ and $\begin{pmatrix} P_2 & S & S \end{pmatrix}^T$. In the first outcome only patients that were identified as high-types receive treatment from the specialist, whereas the

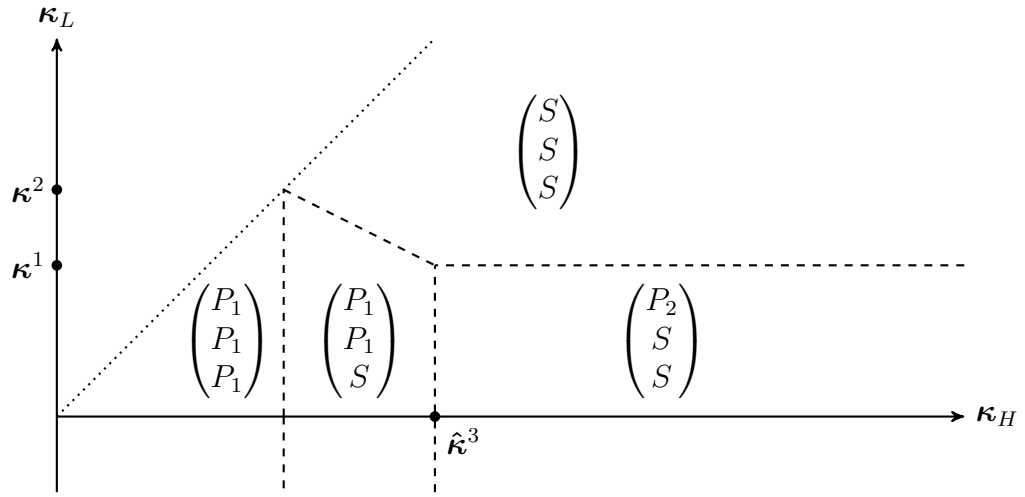


Figure 12: First-best (dashed lines) treatment path vector given the alternative information structure. Dotted line indicates $\kappa_H = \kappa_L$. $\hat{\kappa}^3 := c_S - c_P + \frac{d_S}{\hat{p}_H^0}$.

rest of the patients receive treatment by the PCP. In the second outcome all patients get referred to the specialist and every low-type gets referred back to the PCP. Figure 12 depicts the first-best under the new information structure.

$\begin{pmatrix} P_1 & P_1 & S \end{pmatrix}^T$ is first-best optimal if

$$\begin{aligned} \kappa^2 = c_S - c_P + d_S \leq \kappa_H \leq c_S - c_P + \frac{d_S}{\hat{p}_H^0} =: \hat{\kappa}^3, \\ \kappa^2 \geq \hat{p}_L^0 \kappa_L + \hat{p}_H^0 \kappa_H. \end{aligned} \tag{16}$$

with $\hat{p}_L^0 := \frac{1-p}{1-\hat{q}p}, \hat{p}_H^0 := \frac{(1-\hat{q})p}{1-\hat{q}p}$.

$\begin{pmatrix} P_2 & S & S \end{pmatrix}^T$ is first-best optimal if

$$\begin{aligned} \kappa_H \geq \hat{\kappa}^3, \\ \kappa_L \leq c_S - c_P = \kappa^1. \end{aligned} \tag{17}$$

Note that $\begin{pmatrix} P_2 & S & S \end{pmatrix}^T$ is similar to $\begin{pmatrix} P_1 & P_2 & S \end{pmatrix}^T$. Both call for the specialist to refer back low-severity patients. The implementation of $\begin{pmatrix} P_2 & S & S \end{pmatrix}^T$ is therefore essentially the same

for $\kappa_L > 0$. The constraints that need to be fulfilled are:

$$\begin{aligned}
IC_P^1 &: \hat{p}_L^0(\gamma_P^{P_2} - c_P) + \hat{p}_H^0(\gamma_P^S + \beta_P\kappa_H) \geq \gamma_P^{P_1} - c_P \\
IC_P^2 &: \gamma_P^S + \beta_P\kappa_H \geq \gamma_P^{P_1} - c_P \\
PC_P &: (1-p)(\gamma_P^{P_2} - c_P) + p\gamma_P^S \geq 0 \\
IC_S^1 &: \gamma_S^{P_2} \geq \beta_S\kappa_L + \gamma_S^S - c_S \\
IC_S^2 &: \gamma_S^{P_2} \leq \beta_S\kappa_H + \gamma_S^S - c_S \\
PC_S &: (1-p)(\gamma_S^{P_2} - d_S) + p(\gamma_S^S - c_S - d_S) \geq 0 \\
NNP &: \gamma_j^T \geq 0
\end{aligned} \tag{18}$$

Theorem 9 (proof in Appendix A.10) shows how $\begin{pmatrix} P_2 & S & S \end{pmatrix}^T$ can be implemented by the payer.

Theorem 9. *Let $\kappa_L \leq \kappa^2$. For $\kappa_L > 0$ the contract*

$$\begin{aligned}
\gamma_P^{P_1*} &= c_P \\
\gamma_P^{P_2*} &= c_P \\
\gamma_P^{S*} &= 0 \\
\gamma_S^{P_2*} &= d_S + p\beta_S\kappa_L \\
\gamma_S^{S*} &= d_S + c_S - (1-p)\beta_S\kappa_L
\end{aligned} \tag{19}$$

implements $\begin{pmatrix} P_2 & S & S \end{pmatrix}^T$ without rent payments for the physicians. For $\kappa_L \leq 0$, the cost-based FFS contract implements the treatment path.

For $\kappa_L > 0$, a cost-based FFS contract for the PCP and, for the specialist, a markup for the back-referral to the PCP plus cost sharing for the specialist's treatment implements the outcome without rent for the physicians. For $\kappa_L \leq 0$, a cost-based FFS contract for both physicians implements $\begin{pmatrix} P_2 & S & S \end{pmatrix}^T$ because the specialist will refer back low-severity patients for altruistic reasons.

In Subsection 6.1 we have shown that, if specialist surplus benefits are uncertain, small markups for the specialist are welfare enhancing under the original information structure. In Theorem 11 in Appendix A.12 we show that this is also true under the alternative information structure. In order to implement $\begin{pmatrix} P_1 & P_1 & S \end{pmatrix}^T$ the PCP needs to be incentivized to keep patients of unknown type and refer high-type patients. For the specialist, there are two options:

(1) refer back any L-types she receives,

(2) treat all referred patients.

In the equilibrium path the specialist does not receive any L-types. However, her behavior out of equilibrium is still important since it indirectly influences the PCP's decision making. In the first case the conditions that need to be fulfilled are

$$\begin{aligned}
IC_P^1 &: \gamma_P^{P_1} - c_P \geq \hat{p}_L^0(\gamma_P^{P_2} - c_P) + \hat{p}_H^0(\gamma_P^S + \beta_P \kappa_H) \\
IC_P^2 &: \gamma_P^{P_1} - c_P \leq \gamma_P^S + \beta_P \kappa_H \\
PC_P &: (1 - p\hat{q})(\gamma_P^{P_1} - c_P) + p\hat{q}\gamma_P^S \geq 0 \\
IC_S^1 &: \gamma_S^{P_2} \geq \gamma_S^S + \beta_S \kappa_L - c_S \\
IC_S^2 &: \gamma_S^{P_2} \leq \gamma_S^S + \beta_S \kappa_H - c_S \\
PC_S &: \gamma_S^S - c_S - d_S \geq 0 \\
NNP &: \gamma_j^T \geq 0
\end{aligned} \tag{20}$$

and in the second case

$$\begin{aligned}
IC_P^1 &: \gamma_P^{P_1} - c_P \geq \hat{p}_L^0(\gamma_P^S + \beta_P \kappa_L) + \hat{p}_H^0(\gamma_P^S + \beta_P \kappa_H) \\
IC_P^2 &: \gamma_P^{P_1} - c_P \leq \gamma_P^S + \beta_P \kappa_H \\
PC_P &: (1 - p\hat{q})(\gamma_P^{P_1} - c_P) + p\hat{q}\gamma_P^S \geq 0 \\
IC_S^1 &: \gamma_S^{P_2} \leq \gamma_S^S + \beta_S \kappa_L - c_S \\
IC_S^2 &: \gamma_S^{P_2} \leq \gamma_S^S \beta_S \kappa_H - c_S \\
PC_S &: \gamma_S^S - c_S - d_S \geq 0 \\
NNP &: \gamma_j^T \geq 0.
\end{aligned} \tag{21}$$

It follows Theorem 10 (more details and proof in Appendix A.11).

Theorem 10. *Let* $\begin{pmatrix} P_1 & P_1 & S \end{pmatrix}^T$ *be first-best optimal. The PCP can be (partially) incentivized indirectly by having the specialist back-refer low-types and by setting* $\gamma_P^{P_2} = 0$*. Further, a markup for immediate PCP treatment may be necessary. This may lead to a rent for the PCP.*

If $\kappa_L < 0$*, both physicians could, alternatively, be paid a markup on treatment.*

If $\kappa_L \geq 0$, there is no rent payment necessary for the incentivization of the specialist for either option. For option (1), a markup for the patient's back-referral plus cost sharing for specialist treatment is necessary. Option (2) is implemented by the cost-based FFS contract. The first option is to be preferred since the PCP is indirectly incentivized through the specialist. If the PCP expects that the specialist will refer back L-types, she will be less willing to refer patients of unknown type and, thus, IC_P^1 is easier to fulfill.

The PCP's preference to over-refer patients can be curbed in two ways. First, by setting the payment for returning patients $\gamma_P^{P_2}$ to zero. Second, by paying for the immediate treatment

of the patient combined with cost sharing if the specialist treats the patient. If the PCP is sufficiently altruistic, a markup on treatment costs for the immediate treatment is necessary. This markup can lead to a rent for the PCP if she is very altruistic.

For $\kappa_L < 0$, incentivizing the specialist to treat all referred patients may further save on rents. This would be damaging for low-type patients. However, exactly for this reason the PCP would be less willing to refer patients of unknown type. In treatment path $\begin{pmatrix} P_1 & P_1 & S \end{pmatrix}^T$ the specialist never actually treats any low-type patients; however, the threat of this treatment off the equilibrium path can discipline the PCP. Incentivizing the specialist to treat L-types requires a markup, and therefore possibly a rent payment, for the specialist's treatment of the patient. This is necessary as she prefers not treating L-types due to her altruism. If these additional rents generate larger rent savings for the PCP, a markup for the specialist is optimal. Thus, in contrast to our previous results, using strategic markups for both physicians simultaneously can be optimal.

Summarizing, the set of optimal contracts from the original information structure (cost-based FFS, markups for immediate PCP treatment, and markups for specialist back-referral) appear in the case with the alternative information structure as well. Furthermore, implementing specialist back-referral does not lead to information rents when back-referral is first-best, whereas markups for immediate PCP treatment may. This is also consistent with the results for the original information structure. However, under the alternative information structure specialist incentives may differ for efficient rent extraction. Firstly, we find that using both markup types together can be efficient. Secondly, a new contract type emerges in which the specialist is paid a markup for her treatment in order to prevent the PCP from referring low-type patients that would suffer a health loss from the over-provision of care.

7 Conclusion

In this paper, we analyzed the referral processes between a gatekeeping PCP and a specialist. Both physicians were assumed to have the ability to diagnose and treat patients, though the PCP can only imperfectly diagnose. We consider two information structures of the diagnosis. In the first structure the PCP is able to identify some low-severity patients, whereas in the alternative structure she is able to identify some high-severity patients. Agency problems arise because diagnostic signals are private information of the physicians. The PCP should treat those patients who she can adequately treat and refer those patients who may significantly benefit from specialist treatment. However, since the PCP's diagnostic ability is imperfect, it may be optimal for the specialist to refer back some patients who she has received from the PCP. This is more likely to be optimal if the difference in costs between specialist and PCP treatment is large and the diagnostic costs of the specialist are small. Conversely, if the difference in costs between specialist and PCP treatment is small and the diagnostic costs of the specialist are large, only the PCP should discriminate based on her diagnostic signal.

The following results are true for both information structures. If physicians are altruistic, too many patients will receive specialist treatment under a cost-based FFS contract for both physicians. This can be prevented by paying a markup

(a) to the PCP for treating patients without referral, or

(b) to the specialist for referring patients back plus cost sharing for treatment.

Option (a) directly rewards the PCP for treating more patients, however it does not incentivize the specialist to refer back more patients. This markup can lead to an information rent for the PCP. For this reason, if the PCP's altruism is sufficiently large, it may be suboptimal to utilize it. This result is qualitatively robust when considering waiting costs for the patient when referred. However, information rents are reduced in this case.

Option (b) directly rewards the specialist for diagnosing and referring the patient. It also indirectly incentivizes the PCP not to refer low-severity patients, since she can predict that these patients will get referred back anyways. Furthermore, markups to the specialist do not lead to an information rent for the specialist. Therefore, it can be more attractive to pay a markup to the specialist rather than to the PCP.

If the payer additionally faces uncertainty with regard to the benefit of specialist treatment, small markups for the specialist enhance welfare compared to the cost-based FFS contract. Under the first information structure even small PCP markups may deteriorate welfare due to the accruing rents.

Under the alternative information structure we derive the following additional results. Implementing both markups (a) and (b) at the same time may allow for more rent-efficient contracts. Furthermore, under a specific set of assumptions, markups for specialist treatment can be optimal if the benefit for specialist treatment of low-severity patients is negative. This counter-intuitive result is a consequence of the indirect incentivization of the PCP. If the PCP expects

specialist over-treatment that is detrimental to the patients health, it can incentivize her to not refer some patients in the first place. This, in turn, reduces information rents of the PCP.

In the health care system of some countries, e.g. Germany (Kassenärztliche Bundesvereinigung, 2018), Austria and Poland (Paris et al., 2010), PCPs are paid mostly with a capitated payment, whereas specialists are mostly paid by FFS payments. In our setting this payment system generally leads to an over-supply of specialist treatment. Instead, our model suggest that PCPs should receive a markup on their treatment without referral or the specialist's payment should contain a capitated component.

The optimal contracts outlined above depend on the level of the physicians' altruism which is difficult to estimate in practice. Nevertheless, we contribute by demonstrating which types of contracts should be used. Furthermore, some papers (see e.g. Godager and Wiesen, 2013) attempt to estimate a distribution of altruism coefficients which may be used to estimate second-best contracts.

We did not consider physician capacity constraints. In this case budgeted payment may become necessary to implement first-best behavior as Emons (2013) shows. Our model assumed symmetric information between the physicians with regard to their altruism. Furthermore, we did not examine side contracting or repeated interactions between the physicians. Further research could examine to what extent these limitations affect the ability of the payment system to improve the allocation of treatments.

Appendix: Mathematical Proofs

A.1 Theorem 1

Proof.

Definition 1. *If a treatment vector $\vec{T}^1$ yields a higher expected patient welfare than another vector $\vec{T}^2$, i.e. $EW(\vec{T}^1) > (\geq) EW(\vec{T}^2)$, this is denoted by*

$$\vec{T}^1 \succ (\succeq) \vec{T}^2. \tag{22}$$

To determine the optimal vector of treatment paths, it can first be noted that it can never be optimal to refer a patient in case 1 to the specialist and then refer him back, because it only incurs costs d_S without adding any benefit. Likewise, it can not be optimal that the specialist refers back patients in both cases 2 and 3 simultaneously. Furthermore, it can not be optimal for the specialist to treat patients of the same type differently. Consequently, the set of candidates for the first-best optimum is

$$\left\{ \begin{pmatrix} P_1 \\ P_1 \\ P_1 \end{pmatrix}, \begin{pmatrix} P_1 \\ S \\ P_2 \end{pmatrix}, \begin{pmatrix} P_1 \\ P_2 \\ S \end{pmatrix}, \begin{pmatrix} P_1 \\ S \\ S \end{pmatrix}, \begin{pmatrix} S \\ P_1 \\ P_1 \end{pmatrix}, \begin{pmatrix} S \\ S \\ P_2 \end{pmatrix}, \begin{pmatrix} S \\ S \\ S \end{pmatrix} \right\}. \tag{23}$$

If $(T_1\ S\ P_2)^T \succeq (T_1\ P_1\ P_1)^T$ then $(T_1\ S\ S)^T \succ (T_1\ S\ P_2)^T$ for any treatment path T_1 since

$$\begin{aligned} (T_1\ S\ P_2)^T \succeq (T_1\ P_1\ P_1)^T \iff \kappa_L \geq c_S - c_P + d_S[1 + \frac{p}{(1-p)(1-q)}], \\ \text{and } \kappa_H > \kappa_L \implies (T_1\ S\ S)^T \succ (T_1\ S\ P_2)^T. \end{aligned} \tag{24}$$

Therefore, any treatment path $(T_1\ S\ P_2)^T$ is dominated.

Furthermore, if $(S\ P_1\ P_1)^T \succeq (P_1\ P_1\ P_1)^T$, then $(S\ S\ S)^T \succ (S\ P_1\ P_1)^T$ since

$$\kappa_H > \kappa_L \geq d_S + c_S - c_P. \tag{25}$$

The remaining, non-dominated treatment path vectors are

$(P_1\ P_1\ P_1)^T, (P_1\ P_2\ S)^T, (P_1\ S\ S)^T$ and $(S\ S\ S)^T$. The first-best treatment path vector is determined by calculating the constraints on the model constants under which each candidate treatment path vector is optimal, i.e. delivers an equal or greater expected patient welfare than all the other candidates:

$$\vec{T}^{FB} = \begin{cases} (P_1\ P_1\ P_1)^T \text{ if} & (P_1\ P_1\ P_1)^T \succeq (P_1\ S\ S)^T, \\ & (P_1\ P_1\ P_1)^T \succeq (P_1\ P_2\ S)^T; \\ (P_1\ P_2\ S)^T \text{ if} & (P_1\ P_2\ S)^T \succeq (P_1\ P_1\ P_1)^T, \\ & (P_1\ P_2\ S)^T \succeq (P_1\ S\ S)^T; \\ (P_1\ S\ S)^T \text{ if} & (P_1\ S\ S)^T \succeq (P_1\ P_2\ S)^T \\ & (P_1\ S\ S)^T \succeq (S\ S\ S)^T, \\ & (P_1\ S\ S)^T \succeq (P_1\ P_1\ P_1)^T; \\ (S\ S\ S)^T \text{ if} & (S\ S\ S)^T \succeq (P_1\ S\ S)^T. \end{cases} \tag{26}$$

It is simple to verify that the missing inequations are already implied by (26):

$$\begin{aligned} (P_1\ P_1\ P_1)^T \succeq (P_1\ S\ S)^T &\implies \kappa_L < d_S + c_S - c_P \\ &\implies (P_1\ S\ S)^T \succ (S\ S\ S)^T, \\ (P_1\ P_2\ S)^T \succeq (P_1\ S\ S)^T &\iff \kappa_L < c_S - c_P \\ &\implies (P_1\ S\ S)^T \succ (S\ S\ S)^T, \\ (S\ S\ S)^T \succeq (P_1\ S\ S)^T &\iff \kappa_L \geq d_S + c_S - c_P \\ &\implies (S\ S\ S)^T \succeq (P_1\ P_2\ S)^T, (P_1\ P_1\ P_1)^T. \end{aligned} \tag{27}$$

□

A.2 Theorem 2

Proof. The PCP's incentive constrained are fulfilled and no rent payment occurs.

Specialist:

The specialist's participation constraint is fulfilled:

$$\begin{aligned}\mathbf{E}\Pi_S &= p(\gamma_S^{S^*} - c_S - d_S) + (1-p)(1-q)(\gamma_S^{P_2{}^*} - d_S) \\ &= pp_L^0 \min(-\beta_S\kappa_L, 0) \\ &- (1-p)(1-q)p_H^0 \min(-\beta_S\kappa_L, 0) \\ &= 0.\end{aligned} \tag{28}$$

Her incentive constraints are fulfilled as well, since $\gamma_S^{S^*} - \gamma_S^{P_2{}^*} = c_S + \min(-\beta_S\kappa_L, 0)$. IC_S^2 is obviously fulfilled, IC_S^1 is fulfilled since $[\kappa_H - \kappa_L]\beta_S \geq 0$ and $\beta_S\kappa_H \geq 0$. The non-negative payment constraints are fulfilled as well since $p_L^0\beta_S\kappa_L < \kappa_L \leq d_S + c_S - c_P < d_S + c_S$ and therefore $\gamma_S^{S^*} > 0$.

□

A.3 Theorem 3

Proof. Stage 2 (Specialist): The ICs are fulfilled by a cost-based FFS contract.

Stage 1 (PCP): Let $0 \geq (1-p)q\beta_P\kappa_L$: $\gamma_P^{P_2}$ exactly fulfills their non-negative payment constraint. Fulfilling IC_P^1 with equality already fulfills IC_P^2, hence $\gamma_P^{P_1} = c_P + \gamma_P^S + \beta_P\kappa_L$. Inserting this into PC_P yields $(1-p)q\beta_P\kappa_L + \gamma_P^S = 0$. This implies $\gamma_P^S = 0 - (1-p)q\beta_P\kappa_L \geq 0$ as per assumption. Inserting this into IC_P^1 yields $\gamma_P^{P_1} = c_P + [(1-p)(1-q) + p]\beta_P\kappa_L > 0$.

Let $0 < (1-p)q\beta_P\kappa_L$: $\gamma_P^{P_2}$ and γ_P^S exactly fulfill their non-negative payment constraints. $\gamma_P^{P_1}$ exactly fulfills IC_P^1, such that choosing a smaller $\gamma_P^{P_1}$ is not possible. Therefore the contract minimizes the rent paid to the PCP. IC_P^2 and $NNP_P^{P_1}$ are fulfilled as well.

□

A.4 Theorem 4

Theorem. *If* $\beta_P \leq \beta_P^* := \frac{(1-q)d_S}{c_S - c_P + qd_S}$*, the second-best vector of treatment paths* $\vec{T}^{SB}$ *is given by*

$$\vec{T}^{SB} = \begin{cases} (P_1\ P_1\ P_1)^T \textit{ if} & (1-p)[(1-q)-q\beta_P]\kappa_L + p\kappa_H \leq \\ & [(1-p)(1-q)+p](d_S + c_S - c_P), \\ & \kappa_H \leq c_S - c_P + d_S/p_H^0, \\ & (1-p)\kappa_L + p\kappa_H \leq d_S + c_S - c_P; \\ (P_1\ P_2\ S)^T \textit{ if} & \kappa_H \geq c_S - c_P + d_S/p_H^0, \\ & \kappa_L \leq \frac{(1-q)(c_S-c_P)/(1-p)}{1-q-\beta_P q}; \\ (P_1\ S\ S)^T \textit{ if} & \kappa_L \geq \frac{(1-q)(c_S-c_P)/(1-p)}{1-q-\beta_P q}, \\ & \kappa_L \leq (d_S + c_S - c_P)/(1+\beta_P), \\ & (1-p)[(1-q)-q\beta_P]\kappa_L + p\kappa_H \geq \\ & [(1-p)(1-q)+p](d_S + c_S - c_P); \\ (S\ S\ S)^T \textit{ if} & \kappa_L \geq (d_S + c_S - c_P)/(1+\beta_P), \\ & (1-p)\kappa_L + p\kappa_H \geq d_S + c_S - c_P, \end{cases} \tag{29}$$

else if $\beta_P > \beta_P^*$*,* $\vec{T}^{SB}$ *is given by*

$$\vec{T}^{SB} = \begin{cases} (P_1\ P_1\ P_1)^T \textit{ if} & (1-p)\kappa_L + p\kappa_H \leq d_S + c_S - c_P, \\ & \kappa_H \leq c_S - c_P + d_S/p_H^0; \\ (P_1\ P_2\ S)^T \textit{ if} & \kappa_H \geq c_S - c_P + d_S/p_H^0, \\ & \kappa_L \leq (qd_S + c_S - c_P); \\ (S\ S\ S)^T \textit{ if} & \kappa_L \geq (qd_S + c_S - c_P), \\ & (1-p)\kappa_L + p\kappa_H \geq d_S + c_S - c_P. \end{cases} \tag{30}$$

Proof. Every feasible treatment outcome except $(P_1\ P_1\ P_1)^T$, $(P_1\ P_2\ S)^T$, $(P_1\ S\ S)^T$ and $(S\ S\ S)^T$ is dominated in the second-best.

Since the indiscriminate treatment path vectors $(P_1\ P_1\ P_1)^T$ and $(S\ S\ S)^T$ can be implemented without rent, the proof for this from Theorem 1 still holds.

Calculate the boundaries between the optimal regions:

$$\begin{aligned}
&(P_1\ S\ S)^T \succeq (S\ S\ S)^T \iff \kappa_L \leq (d_S + c_S - c_P)/(1+\beta_P)\\
&(P_1\ S\ S)^T \succeq (P_1\ P_2\ S)^T \iff\\
&\qquad q(-\beta_P \kappa_L) + (1-q)[\kappa_L - c_S + c_P] \geq 0 \iff\\
&\qquad \kappa_L \geq \frac{(1-q)(c_S - c_P)/(1-p)}{1-q-\beta_P q}
\end{aligned} \tag{31}$$

There exists a κ_L such that $(P_1\ S\ S)^T \succeq (S\ S\ S)^T$ and $(P_1\ S\ S)^T \succeq (P_1\ P_2\ S)^T$ if and only if

$$\begin{aligned}
&(d_S + c_S - c_P)/(1+\beta_P) \geq \frac{(1-q)(c_S - c_P)/(1-p)}{1-q-\beta_P q} \iff\\
&(1+\beta_P)[(1-q)(c_S - c_P)/(1-p)] \leq\\
&\qquad (d_S + c_S - c_P)[1-q-\beta_P q] \iff\\
&\qquad\qquad \beta_P[c_S - c_P + q d_S] \leq (1-q) d_S \iff\\
&\qquad\qquad\qquad \beta_P \leq \frac{(1-q) d_S}{c_S - c_P + q d_S}
\end{aligned} \tag{32}$$

If $\beta_P > \beta_P^*$, $(P_1\ S\ S)^T$ is never second-best optimal. The rest of the Theorem follows from

$$\begin{aligned}
&(P_1\ P_1\ P_1)^T \succeq (S\ S\ S)^T \iff (1-p)(d_S + c_S - c_P - \kappa_L)\\
&\qquad + p(d_S + c_S - c_P - \kappa_H) \geq 0 \iff\\
&\qquad (1-p)\kappa_L + p\kappa_H \leq d_S + c_S - c_P,\\
&(P_1\ P_1\ P_1)^T \succeq (P_1\ P_2\ S)^T \iff p(\kappa_H - d_S - c_S + c_P)\\
&\qquad + (1-p)(1-q)(-d_S + c_P) \leq 0 \iff\\
&\qquad \kappa_H \leq c_S - c_P + d_S/p_H^0,\\
&(P_1\ P_2\ S)^T \succeq (S\ S\ S)^T \iff q(-\kappa_L + c_S - c_P + d_S)\\
&\qquad + (1-q)[-\kappa_L + c_S - c_P] \leq 0 \iff\\
&\qquad \kappa_L \leq q d_S + c_S - c_P.
\end{aligned} \tag{33}$$

□

A.5 Lemma 1

Lemma 1. *The patients receive the following treatment paths, whenever the condition on the right side is fulfilled.*

$$
\begin{aligned}
(P_1\ P_1\ P_1)^T, \textit{ if } &\begin{cases} m_S \geq \beta_S \kappa_H & \textit{or} \\ \beta_S \kappa_H \geq m_S \geq \beta_S \kappa_L \textit{and } m_P \geq p_H^0 \beta_P \kappa_H & \textit{or} \\ \beta_S \kappa_L \geq m_S, \textit{and } m_P \geq p_H^0 \beta_P \kappa_H + p_L^0 \beta_P \kappa_L. \end{cases} \\
(P_1\ P_2\ S)^T, &\textit{ if } \beta_S \kappa_H \geq m_S \geq \beta_S \kappa_L \textit{ and } m_P \leq p_H^0 \beta_P \kappa_H \\
(P_1\ S\ S)^T, &\textit{ if } \beta_S \kappa_L \geq m_S \textit{ and } \beta_P \kappa_L \leq m_P \leq p_H^0 \beta_P \kappa_H + p_L^0 \beta_P \kappa_L \\
(S\ S\ S)^T, &\textit{ if } \beta_S \kappa_L \geq m_S \textit{ and } m_P \leq \beta_P \kappa_L.
\end{aligned}
\tag{34}
$$

Proof. The PCP always weakly prefers to treat a patient without referral rather than after a back-referral, since she is paid at least as much in the first case as she is in the second case. Therefore, conditions regarding this have been omitted.

The three conditions that implement $(P_1\ P_1\ P_1)^T$ imply that the PCP prefers to not refer any patient type when the specialist 1.) refers back all patients, 2.) refers back low-type patients and treats high-type patients, and 3.) treats all patients.

The condition that implements $(P_1\ P_2\ S)^T$ implies that the PCP prefers not referring low-type patients and referring patients of unknown type when the specialist treats high-type patients and refers back low-type patients.

The condition that implements $(P_1\ S\ S)^T$ implies that the PCP prefers not referring low-type patients and referring patients of unknown type when the specialist treats both patient types.

The condition that implements $(S\ S\ S)^T$ implies that the PCP prefers referring all patient types when the specialist treats both patient types.

□

A.6 Theorem 5

Proof. According to Lemma 1, paying a markup m_S to the specialist has two effects on the allocation of treatments. One, the boundary between $(P_1\ P_2\ S)^T$ and $(S\ S\ S)^T$ gets shifted upwards, two, a region emerges, in which $(P_1\ P_1\ P_1)^T$ gets played instead of $(P_1\ P_2\ S)^T$ and $(S\ S\ S)^T$ (see Figure 8). We will show that this improves the allocation of treatments if $\frac{m_S}{\beta_S} \leq qd_S + c_S - c_P$, where $\frac{m_S}{\beta_S} = \kappa_L$ and $\frac{m_S}{\beta_S} = \kappa_H$ define the behavioral boundaries between '$(P_1\ P_2\ S)^T$ and $(S\ S\ S)^T$' and '$(P_1\ P_1\ P_1)^T$ and $(P_1\ P_2\ S)^T$' respectively.

Let $\vec{T}^{m_S}$ be the played treatment path vector given m_S.

$$EW^{FB}[(P_1\ P_2\ S)^T] \geq EW^{FB}[(S\ S\ S)^T] \iff \kappa_L \leq qd_S + c_S - c_P, \tag{35}$$

hence expected welfare is improved for $\vec{T}^{m_S} = (P_1\ P_2\ S)^T$.

Turning to $\vec{T}^{m_S} = (P_1\ P_1\ P_1)^T$:

$$\begin{aligned} &EW^{FB}[(P_1\ P_1\ P_1)^T] \geq EW^{FB}[(S\ S\ S)^T] \iff (1-p)\kappa_L + p\kappa_H \leq \kappa^2 \text{ and} \\ &EW^{FB}[(P_1\ P_1\ P_1)^T] \geq EW^{FB}[(P_1\ P_2\ S)^T] \iff \kappa_L \leq \kappa^3. \end{aligned} \tag{36}$$

Since $\frac{m_S}{\beta_S} \leq qd_S + c_S - c_P \leq \kappa^2 \leq \kappa^3$, expected welfare is improved for $\vec{T}^{m_S} = (P_1\ P_1\ P_1)^T$ as well. Furthermore, note that $\gamma_S^S \geq 0$, thus the non-negative payment constraints are not violated.

□

A.7 Theorem 6

Proof. We proceed in the same manner as in the proof of Theorem 5. Paying a markup of m_P to the PCP has three effects on the allocation of treatments. One, for $\kappa_H \leq 0$ a region emerges in which $(P_1\ P_1\ P_1)^T$ gets played instead of $(P_1\ P_2\ S)^T$; two, for $\kappa_H \geq 0$ a region emerges in which $(P_1\ P_1\ P_1)^T$ is played instead of $(S\ S\ S)^T$; three, a region emerges in which $(P_1\ S\ S)^T$

is played instead of $(S\ S\ S)^T$ (see Figure 9).

$$\begin{aligned} \frac{m_P}{\beta_P} &= p_H^0 \kappa_H, \\ \frac{m_P}{\beta_P} &= p_H^0 \kappa_H + p_L^0 \kappa_L, \text{ and} \\ \frac{m_P}{\beta_P} &= \kappa_L \end{aligned} \tag{37}$$

define the boundaries between '$(P_1\ P_1\ P_1)^T$ and $(P_1\ P_2\ S)^T$', '$(P_1\ P_1\ P_1)^T$ and $(P_1\ S\ S)^T$' and '$(P_1\ S\ S)^T$ and $(S\ S\ S)^T$' respectively.

a) First we deal with $\kappa_L \leq 0$: The allocation of treatments is improved

$$\iff EW^{FB}[(P_1\ P_1\ P_1)^T] \geq EW^{FB}[(P_1\ P_2\ S)^T] \iff \kappa_H \leq \kappa^3. \tag{38}$$

Now m_P can be raised until $\frac{m_P}{p_H^0 \beta_P} \leq \kappa^3 \iff \frac{m_P}{\beta_P} \leq d_S + p_H^0(c_S - c_P)$.

Now we deal with $\kappa_L > 0$: For $\frac{m_P}{\beta_P} = \kappa^2 > d_S + p_H^0(c_S - c_P)$ the boundary between $(P_1\ S\ S)^T$ and $(S\ S\ S)^T$ is exactly the first-best boundary. Hence, for $\frac{m_P}{\beta_P} = d_S + p_H^0(c_S - c_P)$, the allocation of treatments must be improved.

The first-best boundary between $(P_1\ P_1\ P_1)^T$ and $(S\ S\ S)^T$ is given by

$$EW^{FB}[(P_1\ P_1\ P_1)^T] = EW^{FB}[(S\ S\ S)^T] \iff \kappa_L = (\kappa^2 - p\kappa_H)/(1-p). \tag{39}$$

It is less steep in κ_H than the boundary between $(P_1\ P_1\ P_1)^T$ and $(P_1\ S\ S)^T$ for contract design (III) (see Figure 9), which is given by $\kappa_L = \frac{m_P}{p_L^0 \beta_P} - \frac{p\kappa_H}{(1-p)(1-q)}$. For $\kappa_L = \kappa_H = \kappa^2$ it holds that $EW^{FB}[(P_1\ P_1\ P_1)^T] = EW^{FB}[(S\ S\ S)^T]$. Hence, the allocation of treatments for $\vec{T}^{m_P} = (P_1\ P_1\ P_1)^T$ is improved for all (κ_L, κ_H) given $\frac{m_P}{\beta_P} \leq d_S + p_H^0(c_S - c_P) < \kappa^2$.

b) For $\kappa_L > 0$, the only time rents need to be paid is when the behavior of the physicians changes. Hence, it is sufficient to prove that the improved allocation of treatments has more positive welfare effects than the negative effect of the rent payments.

Considering rent payments:

$$EW[(P_1\ P_1\ P_1)^T] \geq EW[(S\ S\ S)^T] \iff \kappa_L \leq (\kappa^2 - p\kappa_H - m_P)/(1-p). \tag{40}$$

We insert $\kappa_H = \kappa_L$: $\kappa_L \leq \kappa^2 - m_P$. Now, m_P can be raised until $\frac{m_P}{\beta_P} \leq \kappa^2 - m_P \iff \frac{m_P}{\beta_P} \leq \frac{\kappa^2}{1+\beta_P}$.

Turning to the boundary between $(P_1\ S\ S)^T$ and $(S\ S\ S)^T$:

$$EW[(P_1\ S\ S)^T] \geq EW[(S\ S\ S)^T] \iff \kappa_L \leq \kappa^2 - m_P. \tag{41}$$

Thus, expected patient welfare improves for the rest of the patients as well.

□

A.8 Theorem 7

Proof. PCP:

$$\begin{aligned} \gamma_P^{P_1} - \gamma_P^{P_2} + 2\beta_P w &\overset{!}{\geq} 0 \\ p_L^0(\gamma_P^{P_2} - \beta_P w) + p_H^0(\gamma_P^S + \beta_P \kappa_H + c_P) - \gamma_P^{P_1} - \beta_P w &\overset{!}{\geq} 0 \end{aligned} \tag{42}$$

Inserting the cost-based FFS contract yields

$$\begin{aligned} \beta_P w &\overset{!}{\geq} 0 \\ \beta_P[p_H^0 \kappa_H - (1 + p_l^0)w] &\overset{!}{\geq} 0. \end{aligned} \tag{43}$$

This is fulfilled if $\begin{pmatrix} P_1 & P_2 & S \end{pmatrix}^T$ is first-best optimal since $\kappa_H \geq \kappa_w^3$.

Specialist:

$$\begin{aligned} \beta_S(\kappa_H + w) - c_S + \gamma_S^S - \gamma_S^{P_2} &\overset{!}{\geq} 0 \\ \beta_S(\kappa_L + w) - c_S + \gamma_S^S - \gamma_S^{P_2} &\overset{!}{\leq} 0 \end{aligned} \tag{44}$$

Inserting the proposed contract yields (for $\kappa_L + w > 0$)

$$\gamma_S^S - \gamma_S^{P2} = c_S - \beta_S(\kappa_L + w). \tag{45}$$

For $\kappa_L + w \leq 0$:

$$\gamma_S^S - \gamma_S^{P2} = c_S. \tag{46}$$

Profits are 0: $\mathbf{E}\Pi_S = (1-p)(1-q)(\gamma_S^{P_2} - d_S) + p(\gamma_S^S - c_S - d_S) = 0$. Furthermore, $\gamma_S^S > 0$ if $\begin{pmatrix} P_1 & P_2 & S \end{pmatrix}^T$ is first-best since $\kappa_L \leq c_S - c_P - w$. □

A.9 Theorem 8

Proof. PCP:

$$\begin{aligned}(I) :&\gamma_P^{P_1} - \gamma_P^S - c_P - \beta_P(\kappa_L - w) \overset{!}{\geq} 0 \\ (II) :&\gamma_P^{P_1} - \gamma_P^S - c_P - \beta_P[p_L^0(\kappa_L - w) + p_H^0(\kappa_H - w)] \overset{!}{\leq} 0\end{aligned} \tag{47}$$

Fulfilling (I) exactly (minimizing $\gamma_P^{P_1}$), fulfills (II) as well. Thus, for $w < \kappa_L$, the proposed contract fulfills all PCP conditions with information rent $(1-p)q\beta_P(\kappa_L - w)$.

For $w \geq \kappa_L$, (I) is fulfilled with the cost-based FFS contract. The first-best boundary between $(P_1\ P_1\ P_1)^T$ and $(P_1\ S\ S)^T$ is given by

$$d_S + c_S - c_P = p_L^0(\kappa_L - w) + p_H^0(\kappa_H - w). \tag{48}$$

Thus, (II) holds as well whenever $(P_1\ S\ S)^T$ is first-best.

Specialist:

$$\gamma_S^S - \gamma_S^{P_2} + \beta_S(\kappa_L + w) - c_S \overset{!}{\geq} 0 \tag{49}$$

needs to hold. This is fulfilled by the cost-based FFS contract since $\kappa_L + w > 0$ whenever $\begin{pmatrix} P_1 & S & S \end{pmatrix}^T$ is first-best (and second-best) optimal. □

A.10 Theorem 9

Proof. The ICs that need to be fulfilled for S are

$$\begin{aligned}(I) :&\gamma_S^{P_2} - \gamma_S^S \geq \beta_S\kappa_L - c_S \\ (II) :&\gamma_S^{P_2} - \gamma_S^S \leq \beta_S\kappa_H - c_S\end{aligned} \tag{50}$$

Fulfilling (I) exactly, implies (II). Setting the rent equal to 0 delivers the result for the specialist for $\kappa_L \geq 0$. This contract fulfills the non-negativity constraints since $\kappa_L \leq \kappa^2 = d_S + c_S - c_P$. For $\kappa_L \leq 0$ the specialist will only treat high-type patients under cost-based FFS. The PCP always refers patients under cost-based FFS since the specialist will only treat high-types. □

A.11 Theorem 10

Theorem. *Consider the alternative information structure. Let* $\begin{pmatrix} P_1 & P_1 & S \end{pmatrix}^T$ *be first-best optimal. The treatment path can be implemented in the following way:*

1. For $\kappa_L \geq 0$*:*

If $0 \geq (1 - p\hat{q})[\hat{p}_H^0 \beta_P \kappa_H - \hat{p}_L^0 c_P]$*, the contract*

$$\begin{aligned} {\gamma_P^{P_1}}^* &= \hat{p}_H^0(\beta_P \kappa_H + \gamma_P^{S*} + c_P) \\ {\gamma_P^{P_2}}^* &= 0 \\ \gamma_P^{S*} &= \frac{0 - (1 - p\hat{q})[\hat{p}_H^0 \beta_P \kappa_H - \hat{p}_L^0 c_P]}{p} \\ {\gamma_S^{P_2}}^* &= d_S + p\beta_S \kappa_L \\ \gamma_S^{S*} &= d_S + c_S - (1 - p)\beta_S \kappa_L \end{aligned} \tag{51}$$

implements $\begin{pmatrix} P_1 & P_1 & S \end{pmatrix}^T$ *without rent payments for the physicians.*

If $0 \leq (1 - p\hat{q})[\hat{p}_H^0 \beta_P \kappa_H - \hat{p}_L^0 c_P]$*, the unique rent-minimizing contract is*

$$\begin{aligned} {\gamma_P^{P_1}}^* &= \hat{p}_H^0(\beta_P \kappa_H + c_P) \\ {\gamma_P^{P_2}}^* &= 0 \\ \gamma_P^{S*} &= 0 \\ {\gamma_S^{P_2}}^* &= d_S + p\beta_S \kappa_L \\ \gamma_S^{S*} &= d_S + c_S - (1 - p)\beta_S \kappa_L \end{aligned} \tag{52}$$

2. For $\kappa_L < 0$*: If*

$$\begin{aligned} \max(-p\hat{q}(\beta_S \kappa_L + d_S), 0) + \max((1 - p\hat{q})\beta_P[\hat{p}_L^0 \kappa_L + \hat{p}_H^0 \kappa_H], 0) \geq \\ \max((1 - p\hat{q})[\beta_P \hat{p}_H^0 \kappa_H - \hat{p}_L^0 c_P], 0), \end{aligned} \tag{53}$$

the contracts from 1. with cost-based FFS payment for the specialist implement $\begin{pmatrix} P_1 & P_1 & S \end{pmatrix}^T$ *with minimal rents.*

Otherwise, both physicians needs to receive a markup on immediate treatment. Then the PCP earns a rent if $(1-p\hat{q})\beta_P[\hat{p}_L^0\kappa_L+\hat{p}_H^0\kappa_H] \geq 0$ *and the specialist earns a rent if* $\beta_S\kappa_L+d_S \geq 0$*. If rents accrue to both physicians, the unique rent minimizing contract is*

$$\begin{aligned}
\gamma_P^{P_1*} &= \beta_P(\hat{p}_L^0\kappa_L + \hat{p}_H^0\kappa_H) + c_P \\
\gamma_P^{P_2*} &= 0 \\
\gamma_P^{S*} &= 0 \\
\gamma_S^{P_2*} &= 0 \\
\gamma_S^{S*} &= c_S - \beta_S\kappa_L.
\end{aligned} \tag{54}$$

Proof. 1. The PCP is incentivized to refer only high-type patients and the specialist is incentivized to treat them and refer any back any low-type patients if they were referred (see Appendix A.10). The ICs that need to be fulfilled for P are

$$\begin{aligned}
(I) &: \hat{p}_L^0(\gamma_P^{P_1} - \gamma_P^{P_2}) + \hat{p}_H^0(\gamma_P^{P_1} - \gamma_P^S - \beta_P\kappa_H - c_P) \geq 0 \\
(II) &: \gamma_P^S - \gamma_P^{P_1} \geq -\beta_P\kappa_H - c_P
\end{aligned} \tag{55}$$

$\gamma_P^{P_2}$ can be set to 0 to minimize rents. This changes (I) to

$$(I^*) : \gamma_P^{P_1} \geq \hat{p}_H^0(\beta_P\kappa_H + \gamma_P^S + c_P) \tag{56}$$

If (I^*) is binding, (II) is fulfilled. Let $\gamma_P^S = 0$. The PC_P is fulfilled and a positive rent accrues to P if and only if

$$\mathbf{E}\Pi_P = (1-p\hat{q})(\hat{p}_H^0\beta_P\kappa_H - \hat{p}_L^0 c_P) \overset{!}{\geq} 0 \tag{57}$$

and the rent-minimizing contract is

$$\begin{aligned}
\gamma_P^{P_1*} &= \hat{p}_H^0(\beta_P\kappa_H + c_P) \\
\gamma_P^{P_2*} &= 0 \\
\gamma_P^{S*} &= 0 \\
\gamma_S^{P_2*} &= d_S + p\beta_S\kappa_L \\
\gamma_S^{S*} &= d_S + c_S - (1-p)\beta_S\kappa_L
\end{aligned} \tag{58}$$

$\gamma_S^{S*} \geq 0$ because $\kappa_L \leq \kappa^2$, thus NNP is fulfilled.

Otherwise, the outcome can be implemented by setting the rent to 0:

$$\begin{aligned} \mathbf{E}\Pi_P = (1-p\hat{q})[\hat{p}_H^0(\beta_P\kappa_H + \gamma_P^S + c_P) - c_P] + p\hat{q}\gamma_P^S \stackrel{!}{=} 0 \iff \\ \gamma_P^S = \frac{-(1-p\hat{q})[\hat{p}_H^0\beta_P\kappa_H - \hat{p}_L^0 c_P]}{p}. \end{aligned} \tag{59}$$

2. The specialist can be incentivized to treat all referred patients by paying her a markup on treatment of $-\beta_S\kappa_L$ compared to back-referral. This leads to a rent of $\Pi_S = \max(-p\hat{q}(\beta_S\kappa_L + d_S), 0)$. The ICs for the PCP are now

$$\begin{aligned} (I) : \gamma_P^{P_1} - \gamma_P^S \geq \beta_P(\hat{p}_L^0\kappa_L + \hat{p}_H^0\kappa_H) + c_P \\ (II) : \gamma_P^S - \gamma_P^{P_1} \geq -\beta_P\kappa_H - c_P \end{aligned} \tag{60}$$

Fulfilling (I) exactly, implies (II). If rents accrue, they are minimized by setting $\gamma_P^S = 0$. The PCP's profits are now $\Pi_P = (1-p\hat{q})\beta_P[\hat{p}_L^0\kappa_L + \hat{p}_H^0\kappa_H]$. If the sum of PCP's and specialists profits are smaller than the PCP's profits from the contract in 2

$$\begin{aligned} \iff \max(-p\hat{q}(\beta_S\kappa_L - d_S), 0) + \max((1-p\hat{q})\beta_P[\hat{p}_L^0\kappa_L + \hat{p}_H^0\kappa_H], 0) \leq \\ \max((1-p\hat{q})[\beta_P\hat{p}_H^0\kappa_H - \hat{p}_L^0 c_P], 0), \end{aligned} \tag{61}$$

the specialist should be incentivized to treat all referred patients. □

A.12 Theorem 11

Theorem 11. *Consider the alternative information structure. Consider the contract*

$$\begin{aligned} \gamma_P^{P_1*} &= c_P \\ \gamma_P^{P_2*} &= c_P \\ \gamma_P^{S*} &= 0 \\ \gamma_S^{P_2*} &= d_S + pm_S \\ \gamma_S^{S*} &= d_S + c_S - (1-p)m_S \end{aligned} \tag{62}$$

with $m_S \leq \beta_S(c_S - c_P)$. This contract is always welfare enhancing over the cost-based FFS contract.

Proof. Under the cost-based FFS contract, the resulting treatment path vector is

$$\begin{cases} \begin{pmatrix} S & S & S \end{pmatrix}^T & \text{for } \kappa_L > 0 \\ \begin{pmatrix} P_2 & S & S \end{pmatrix}^T & \text{for } \kappa_L \leq 0. \end{cases} \tag{63}$$

Under contract (62) it is

$$\begin{cases} \begin{pmatrix} S & S & S \end{pmatrix}^T & \text{for } \kappa_L > \frac{m_S}{\beta_S} \\ \begin{pmatrix} P_2 & S & S \end{pmatrix}^T & \text{for } \kappa_L \leq \frac{m_S}{\beta_S}. \end{cases} \tag{64}$$

In the first-best $\begin{pmatrix} P_2 & S & S \end{pmatrix}^T$ yields larger expected welfare than $\begin{pmatrix} S & S & S \end{pmatrix}^T$ if and only if

$$\kappa_L \leq c_S - c_P. \tag{65}$$

Thus, the proposed contract is welfare enhancing if and only if $m_S \leq \beta_S(c_S - c_P)$. □

References

Allard, Marie, Jelovac, Izabela, and Léger, Pierre Thomas (2011). "Treatment and referral decisions under different physician payment mechanisms". In: *Journal of Health Economics* 30.5, pp. 880–893. DOI: 10.1016/j.jhealeco.2011.05.016.

— (2014). "Payment mechanism and GP self-selection: capitation versus fee for service". In: *International Journal of Health Care Finance and Economics* 14.2, pp. 143–160. DOI: 10.1007/s10754-014-9143-z.

Barnett, Michael L., Song, Zirui, and Landon, Bruce E. (2012). "Trends in Physician Referrals in the United States, 1999-2009". In: *Archives of Internal Medicine* 172.2, p. 163. DOI: 10.1001/archinternmed.2011.722.

Brekke, Kurt R., Nuscheler, Robert, and Straume, Odd Rune (2007). "Gatekeeping in health care". In: *Journal of Health Economics* 26.1, pp. 149–170. DOI: 10.1016/j.jhealeco.2006.04.004.

Chalkley, Martin and Malcomson, James M. (1998). "Contracting for Health Services when Patient Demand Does Not Reflect Quality". In: *Journal of Health Economics* 17, pp. 1–19.

Chin, M H, Zhang, J X, and Merrell, K (2000). "Specialty differences in the care of older patients with diabetes." In: *Medical Care* 38 (2), pp. 131–140. ISSN: 0025-7079.

Darby, M.R. and Karni, E. (1973). "Free Competition and the Optimal Amount of Fraud". In: *Journal of Law & Economics* 16, pp. 67–88.

Dulleck, Uwe and Kerschbamer, Rudolf (2006). "On doctors, mechanics, and computer specialists: The economics of credence goods". In: *Journal of Economic literature* 44.1, pp. 5–42. DOI: 10.1257/002205106776162717.

Dusheiko, Mark, Gravelle, Hugh, Jacobs, Rowena, and Smith, Peter (2006). "The effect of financial incentives on gatekeeping doctors: evidence from a natural experiment." In: *Journal of health economics* 25 (3), pp. 449–478. ISSN: 0167-6296. DOI: 10.1016/j.jhealeco.2005.08.001. ppublish.

Dusheiko, Mark, Gravelle, Hugh, Yu, Ning, and Campbell, Stephen (2007). "The impact of budgets for gatekeeping physicians on patient satisfaction: Evidence from fundholding".

In: *Journal of Health Economics* 26.4, pp. 742–762. DOI: 10.1016/j.jhealeco.2006.12.003.

Ellis, Randall P. and McGuire, Thomas G. (1986). "Provider Behavior Under Prospective Reimbursement: Cost Sharing and Supply". In: *Journal of Health Economics* 5, pp. 129–151.

Emons, W. (2013). "Incentive compatible reimbursement schemes for physicians". In: *Journal of Institutional and Theoretical Economics* 159, pp. 605–620. DOI: 10.1628/093245613X671869.

Garcia-Mariñoso, Begoña and Jelovac, Izabela (2003). "GPs' payment contracts and their referral practice". In: *Journal of Health Economics* 22.4, pp. 617–635. ISSN: 0167-6296. DOI: 10.1016/S0167-6296(03)00008-0.

Godager, Geir, Iversen, Tor, and Ma, Ching-to A. (2015). "Competition, gatekeeping, and health care access". In: *Journal of Health Economics* 39, pp. 159–170. DOI: 10.1016/j.jhealeco.2014.11.005.

Godager, Geir and Wiesen, Daniel (2013). "Profit or patients' health benefit? Exploring the heterogeneity in physician altruism". In: *Journal of Health Economics* 32.6, pp. 1105–1116. DOI: 10.1016/j.jhealeco.2013.08.008.

González, Paula (2009). "Gatekeeping versus direct-access when patient information matters". In: *Health Economics* 19.6, pp. 730–754. DOI: 10.1002/hec.1506.

Government of Western Australia - Department of Health (2006). *Respiratory Referral Recommendations*. URL: http://www.gp.health.wa.gov.au/CPAC/speciality/docs/REFREC022.pdf (visited on 09/30/2021).

Grassi, Simona and Ma, Ching-to A. (2016). "Information acquisition, referral, and organization". In: *The RAND Journal of Economics* 47.4, pp. 935–960. DOI: 10.1111/1756-2171.12160.

Harrold, Leslie R., Field, Terry S., and Gurwitz, Jerry H. (1999). "Knowledge, patterns of care, and outcomes of care for generalists and specialists". In: *Journal of General Internal Medicine* 14.8, pp. 499–511. DOI: 10.1046/j.1525-1497.1999.08168.x.

Inderst, Roman and Ottaviani, Marco (2012). "Competition through Commissions and Kickbacks". In: *American Economic Review* 102.2, pp. 780–809. DOI: 10.1257/aer.102.2.780.

Iversen, Tor and Lurås, Hilde (2000). "The effect of capitation on GPs' referral decisions". In: *Health Economics* 9.3, pp. 199–210. ISSN: 1099-1050. DOI: 10.1002/(SICI)1099-1050(200004)9:3<199::AID-HEC514>3.0.CO;2-2.

Kassenärztliche Bundesvereinigung (2018). *Honorarbericht 2/2016*. URL: http://www.kbv.de/media/sp/Honorarbericht_Quartal_2_2016.pdf (visited on 09/30/2021).

Kerschbamer, Rudolf and Sutter, Matthias (2017). "The Economics of Credence Goods – a Survey of Recent Lab and Field Experiments". In: *CESifo Economic Studies* 63.1, pp. 1–23. DOI: 10.1093/cesifo/ifx001.

Kerschbamer, Rudolf, Sutter, Matthias, and Dulleck, Uwe (2017). "How Social Preferences Shape Incentives in (Experimental) Markets for Credence Goods". In: *The Economic Journal* 127.600, pp. 393–416. DOI: 10.1111/ecoj.12284.

Liu, Ting and Ma, Ching-to A. (2013). "Health insurance, treatment plan, and delegation to altruistic physician". In: *Journal of Economic Behavior & Organization* 85, pp. 79–96. DOI: 10.1016/j.jebo.2012.11.002.

Liu, Ting, Ma, Ching-to A., and Mak, Henry Y. (2018). "Incentives for motivated experts in a partnership". In: *Journal of Economic Behavior & Organization* 152, pp. 296–313. DOI: 10.1016/j.jebo.2018.05.003.

Malcomson, James M. (2004). "Health Service Gatekeepers". In: *The RAND Journal of Economics* 35.2, pp. 401–421. ISSN: 0741-6261. URL: http://www.jstor.org/stable/1593698.

Mehrotra, Ateev, Forrest, Christopher B., and Lin, CAROLINE Y. (2011). "Dropping the Baton: Specialty Referrals in the United States". In: *Milbank Quarterly* 89.1, pp. 39–68. DOI: 10.1111/j.1468-0009.2011.00619.x.

Olivella, Pau and Siciliani, Luigi (2017). "Reputational concerns with altruistic providers". In: *Journal of Health Economics* 55, pp. 1–13. DOI: 10.1016/j.jhealeco.2017.05.003.

Paris, V., Devaux, M., and Wei, L. (2010). "Health Systems Institutional Characteristics: A Survey of 29 OECD Countries". In: *OECD Health Working Papers* 50. DOI: 10.1787/5kmfxfq9qbnr-en.

Pauly, Mark V. (1979). "The Ethics and Economics of Kickbacks and Fee Splitting". In: *The Bell Journal of Economics* 10.1, p. 344. DOI: 10.2307/3003336.

Sarma, Sisira, Mehta, Nirav, Devlin, Rose Anne, Kpelitse, Koffi Ahoto, and Li, Lihua (2018). "Family physician remuneration schemes and specialist referrals: Quasi-experimental evidence from Ontario, Canada". In: *Health Economics* 27.10, pp. 1533–1549. DOI: 10.1002/hec.3783.

Schermer, T, Smeenk, F, and Weel, C van (2003). "Referral and consultation in asthma and COPD: an exploration of pulmonologists' views." In: *The Netherlands journal of medicine* 61 (3), pp. 71–81. ISSN: 0300-2977.

Shumsky, Robert A. and Pinker, Edieal J. (2003). "Gatekeepers and Referrals in Services". In: *Management Science* 49.7, pp. 839–856. DOI: 10.1287/mnsc.49.7.839.16387.

Whittle, J., Lin, C. J., Lave, J. R., Fine, M. J., Delaney, K. M., Joyce, D. Z., Young, W. W., and Kapoor, W. N. (1998). "Relationship of provider characteristics to outcomes, process, and costs of care for community-acquired pneumonia." eng. In: *Medical care* 36 (7), pp. 977–87.

Wilson, Charlotte, Campbell, Stephen M., Luker, Karen A., and Caress, Ann-Louise (2012). "Referral and management options for patients with chronic kidney disease: perspectives of patients, generalists and specialists". In: *Health Expectations* 18.3, pp. 325–334. DOI: 10.1111/hex.12025.

hche | Hamburg Center for Health Economics

hche Research Paper Series, ISSN 2191-6233 (Print), ISSN 2192-2519 (Internet)

2011/1 Mathias Kifmann and Kerstin Roeder, Premium Subsidies and Social Insurance: Substitutes or Complements? March 2011

2011/2 Oliver Tiemann and Jonas Schreyögg, Changes in Hospital Efficiency after Privatization, June 2011

2011/3 Kathrin Roll, Tom Stargardt and Jonas Schreyögg, Effect of Type of Insurance and Income on Waiting Time for Outpatient Care, July 2011

2012/4 Tom Stargardt, Jonas Schreyögg and Ivan Kondofersky, Measuring the Relationship between Costs and Outcomes: the Example of Acute Myocardial Infarction in German Hospitals, August 2012

2012/5 Vera Hinz, Florian Drevs, Jürgen Wehner, Electronic Word of Mouth about Medical Services, September 2012

2013/6 Mathias Kifmann, Martin Nell, Fairer Systemwettbewerb zwischen gesetzlicher und privater Krankenversicherung, July 2013

2013/7 Mareike Heimeshoff, Jonas Schreyögg, Estimation of a physician practise cost function, August 2013

2014/8 Mathias Kifmann, Luigi Siciliani, Average-cost Pricing and Dynamic Selection Incentives in the Hospital Sector, October 2014

2015/9 Ricarda Milstein, Jonas Schreyögg, A review of pay-for-performance programs in the inpatient sector in OECD countries, December 2015

2016/10 Florian Bleibler, Hans-Helmut König, Cost-effectiveness of intravenous 5 mg zoledronic acid to prevent subsequent clinical fractures in postmenopausal women after hip fracture: a model-based analysis, January 2016

2016/11 Yauheniya Varabyova, Rudolf Blankart, Jonas Schreyögg, Using Nonparametric Conditional Approach to Integrate Quality into Efficiency Analysis: Empirical Evidence from Cardiology Departments, May 2016

2016/12 Christine Blome Ph.D., Prof. Dr. Matthias Augustin, Measuring change in subjective well-being: Methods to quantify recall bias and recalibration response shift, 2016

2016/13 Michael Bahrs, Mathias Schumann, Unlucky to be Young? The Long-Term Effects of School Starting Age on Smoking Behaviour and Health, August 2016

2017/14 Konrad Himmel, Udo Schneider, Ambulatory Care at the End of a Billing Period, March 2017

2017/15 Philipp Bach, Helmut Farbmacher, Martin Spindler, Semiparametric Count Data Modeling with an Application to Health Service Demand, September 2017

2018/16 Michael Kvasnicka, Thomas Siedler, Nicolas R. Ziebarth, The Health Effects of Smoking Bans: Evidence from German Hospitalization Data, June 2018

2019/17 Jakob Everding, Jan Marcus, The Effect of Unemployment on the Smoking Behavior of Couples, May 2019

2019/18 Jakob Everding, Heterogeneous Spillover Effects of Children's Education on Parental Mental Health, July 2019

2019/19 Esra Eren Bayindir, Hospital Ownership Type and Service Provision, a Structural Approach, November 2019

2019/20 Shushanik Margaryan, Low Emission Zones and Population Health, December 2019

2020/21 Barbara Boggiano, Long-ter effects of the Paraguayan War (1864-1870): from male scarcity to intimate partner violence, May 2020

2020/22 Matthias Bäuml, Christian Kümpel, Hospital Responses to the Introduction of Reimbursements by Treatment Intensity in a (Presumably Lump Sum) DRG System, May 2020

2020/23 Philipp Bach, Victor Schernozhukov, Martin Spindler, Insights from optimal pandemic shielding in a multi-group SEIR framework, November 2020

2021/24 Florian Hofer, Benjamin Birkner, Martin Spindler, Power of machine learning algorithms for predicting dropouts from a German telemonitoring program using standardized claims data, June 2021

2021/25 Malte Griebenow, Mathias Kifmann, Diagnostics and Treatment: On the Division of Labor between Primary Care Physicians and Specialists, November 2021

The Hamburg Center for Health Economics is a joint center of Universität Hamburg and the University Medical Center Hamburg-Eppendorf (UKE).

hche | Hamburg Center for Health Economics

Esplanade 36
20354 Hamburg
Germany
Tel: +49 (0) 42838-9515/9516
Fax: +49 (0) 42838-8043
Email: info@hche.de
http://www.hche.de

HCHE Research Papers are indexed in RePEc and SSRN.
Papers can be downloaded free of charge from http://www.hche.de.