ESSAYS ON HEALTH ECONOMICS:

Cooperation of Physicians

UNIVERSITÄT HAMBURG
Fakultät für Wirtschafts- und Sozialwissenschaften

DISSERTATION
Zur Erlangung der Würde eines Doktors der
Wirtschafts- und Sozialwissenschaften
„Dr. rer. pol." (gemäß der PromO vom 18. Januar 2017)

vorgelegt von

Malte Sven Griebenow
aus Reinbek

Hamburg, den 19.10.2021

Erstbetreuer	Prof. Dr. Mathias Kifmann
Zweitbetreuer	Prof. Dr. Gerd Mühlheußer
Erstgutachter	Prof. Dr. Mathias Kifmann
Zweitgutachter	Prof. Dr. Gerd Mühlheußer
Drittgutachter	Prof. Dr. Burkhard Hehenkamp
Vorsitzende	Prof. Dr. Johanna Kokot
Disputation am	11.07.2022

Contents

1 Introduction

Collaborative, interdisciplinary care is increasingly important for the provision of high-quality medical care. One reason for this is the increasing prevalence of multi-morbid and chronic diseases (Navickas et al., 2016; K. Barnett et al., 2012). Figure 1.1 depicts an increase in the prevalence of multi-morbidity and Figure 1.2 depicts an increase in the prevalence of chronic diseases for US adults between the years 1988 and 2014.

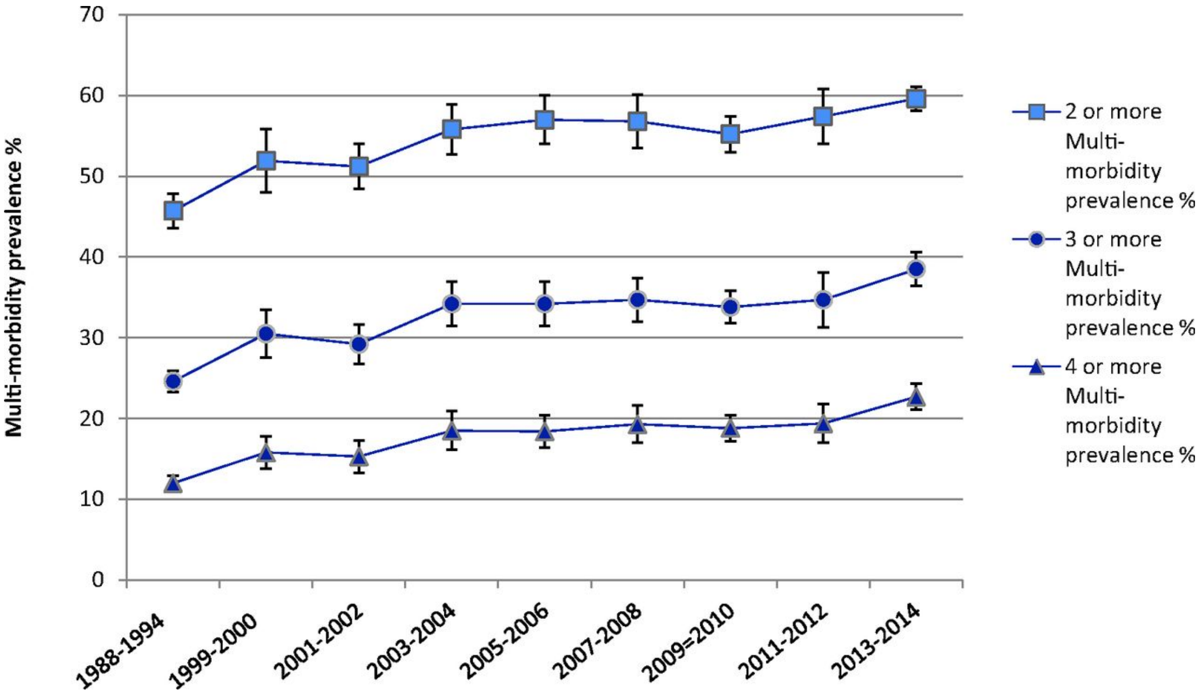


Figure 1.1: Age-standardized trends in multi-morbidity prevalence for participants 20 years or older from NHANES 1988–2014 by number of comorbidities. Source: King et al. (2018)

Demographic change and the resulting complex health care needs of the elderly also increase the need for interdisciplinary care. Though an average US patient sees 18.7 different physicians per year, for elderly patients this number increases to 28.4 (Practice Fusion, 2010). Consequently, the coordination of care between physicians is getting increasingly important for the delivery of high-quality and cost-efficient care. Policy makers have recognized the need for improved care coordination. As a result, numerous interventions to improve coordination have been carried out. To name a few examples, Medicare has started paying physicians for care coordination services of chronically ill

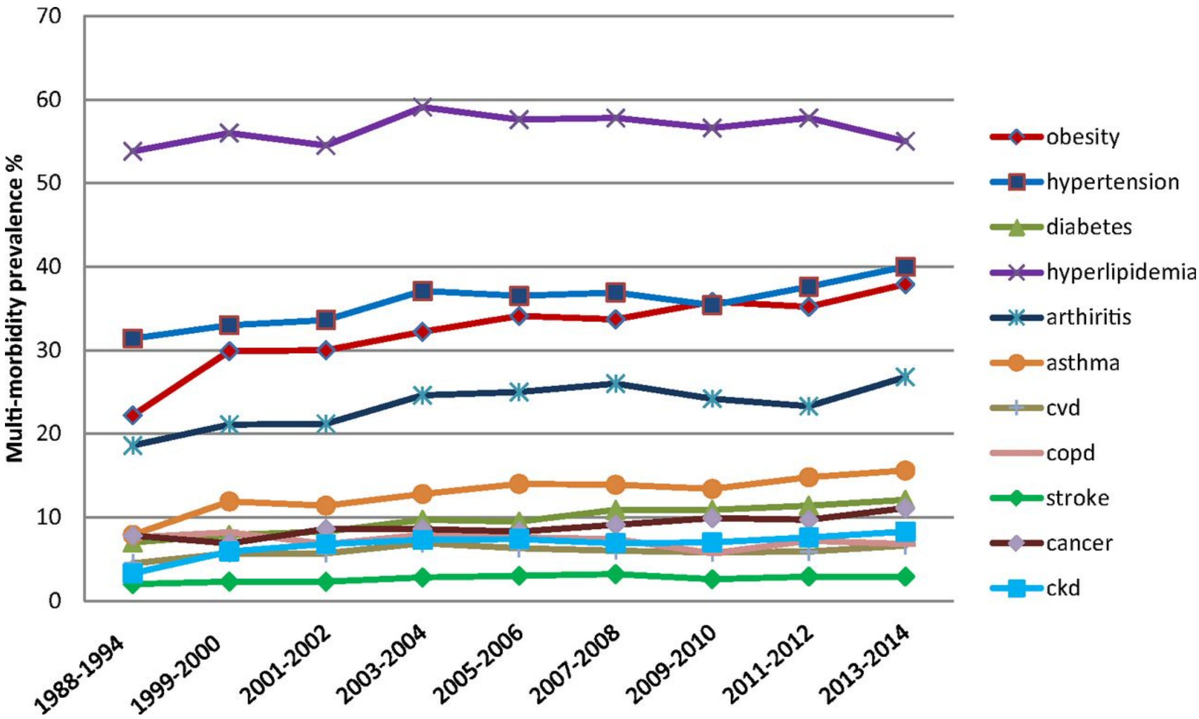


Figure 1.2: Age-standardized prevalence of various chronic conditions in participants 20 years or older from NHANES 1988–2014. Source: King et al. (2018)

patients in 2015 (Centers for Medicare & Medicaid Services, 2014), the Netherlands have experimented with a bundled payment scheme (Bakker et al., 2012), and a German province used a population based approach to improve quality of care (Schubert et al., 2016).

An important aspect of care coordination is the referral process between physicians. When referring a patient, the physician is transferring part or all of the responsibility for the diagnosis and treatment of the patient to another physician. This transfer of responsibility may be temporal or permanent. It is the main channel through which independent ambulatory physicians interact. Due to the increased need for multi-specialty treatment, referral rates have increased in recent years. For example, in the US the referral rate of physicians has almost doubled to about 10% from 1999 to 2009 (M. L. Barnett et al., 2012).

A typical example of a patient referral is the referral between a primary care physician

(PCP) and a specialist. The analysis of referrals between theses types of physicians is the primary concern of this thesis. Whether a patient should be referred in order to receive high-quality and cost-efficient care depends mainly on the severity of his disease and the physicians' respective abilities in diagnosing and treating the patient. If and only if the additional expected value of diagnostic and treatment services that the specialist provides exceed the additional expected costs, the referral should be made from a social welfare perspective. Interestingly, there is a great variation in referral rates between physicians even after accounting for case-mix (Franks et al., 1999). Furthermore, empirical evidence suggests that financial incentives generated by the payment system for physician services affect referral rates. Iversen and Lurås (2000) and Sarma et al. (2018) find that an increase in the capitated component of PCPs' pay and a corresponding decrease in service fees lead to a greater number of patient referrals. This finding can be explained if physicians' decisions do not only depend on the well-being of the patient but also on their own profits. Under a capitated system, a PCP has the incentive to treat as many patients in the most economical way possible. Conversely, under a fee-for-service (FFS) system maximizing the number of services per patient maximizes revenue. Thus, the study of payment systems is important to incentivize desired physician behavior.

Mehrotra et al. (2011) summarize the various ways in which specialty referrals in the US fall short of expectations. They find evidence for both under- and over-referral to specialist care. Under-referral of patients is problematic because patients do not receive treatment which may have greatly improve their health. Over-referral is problematic because patients receive care that yields little benefit for their health at an increased cost for either them or the health insurance system (Starfield and Shi, 2002; Chin et al., 2000; Harrold et al., 1999; Whittle et al., 1998). As potential solutions to these problems Mehrotra et al. (2011) consider the use of financial incentives as well as gatekeeping arrangements.

The problems of over- and under-referral are examined by means of theoretical models in Chapters 3 and 4. They examine referral processes between a PCP and a specialist. For severely ill patients the specialist provides superior treatment. However, the specialist's treatment is more expensive, making it inefficient for mildly ill patients. Chapter 3 focuses on the different diagnostic abilities of PCP and specialist. The PCP is able to identify the disease severity of some patients. However, for other patients the expertise of a specialist is required. Furthermore, some patients require specialist diagnosis but not specialist care. Optimally, the specialist should refer these patients back for PCP treatment.

The main finding of Chapter 3 is that markups to the PCP for treating patients immediately, or to the specialist for referring back patients, are necessary in some cases to induce optimal referrals. However, in contrast to markups for the specialist, markups to the PCP lead to a rent payment. Thus, they may not be optimal. Furthermore, if the payer faces uncertainty regarding the benefits from specialist treatment, sufficiently small markups for the specialist always enhance social welfare. The model is extended by considering waiting costs when the patient is referred. The same contract types as before are optimal, however, it is more beneficial to only have the PCP discriminate based on her diagnostic signal.

Whereas Chapter 3 deals with referrals in a static context, Chapter 4 considers a dynamic setting with chronically ill patients. The main innovation of this chapter is to consider patients whose health states change probabilistically over time according to a Markovian process. Chronic diseases cause large health related and economic costs. High-quality physician care can help to alleviate these costs by reducing the number of complications, related health problems, and specialist visits in the future (Bodenheimer, Wagner, et al., 2002). Bodenheimer, Chen, et al. (2009) suggest that chronic diseases should be treated in multi-disciplinary physician teams in order to facilitate cooperation between the physicians. Chapter 4 presents a theoretical model concerning chronically ill patients who may require treatment from both a PCP and a specialist. The aim of Chapter 4 is to determine whether, and under which conditions, it is desirable to organize the treatment of chronic patients in a physician team or in solo practices. Important aspects are whether the adequate physician treats the patient and whether this physician exerts sufficient effort.

The main findings are as follows. First-best efficient effort levels and referrals can only be achieved in a physician team. Furthermore, the difference in expected treatment costs between patient types needs to be relatively larger for the PCP. In this case it is always optimal for the physicians to form a team since the first-best can never be reached in the solo practices. Treatment fees for the PCP should include a markup whereas the specialist should be paid below costs. However, if the difference in expected treatment costs is relatively larger for the specialist, it can be welfare enhancing to deliver care through physicians who work in solo practices. The reason for this is that in solo practices adverse patient selection incentives only affect the specialist, whereas in the team they also affect the decisions of the PCP. Thus, in the solo practices, the payer can implement a gatekeeping rule that allows the PCP to treat all patients in mild condition until their condition deteriorates. This gatekeeping rule cannot be implemented in the physician team.

In addition to the problem of over- and under-provision of care, Mehrotra et al. (2011) find that information exchange between physicians is often insufficient. In 28-68% of referrals PCPs did not communicate with the specialist, whereas in 4-45% of cases the specialists did not communicate with the PCP. Even if information is provided, it is often of low quality. These problems seem to be more pronounced when a PCP is referring the patient. Resulting from these deficiencies in the referral process, treatment quality is decreased and duplicate testing is increased.

These problems are examined in Chapter 5 by means of a theoretical model and an experimental test. The model is based on a partially altruistic PCP who is referring a patient to specialist care. She can either transfer no information or information of low or high quality to the specialist with high-quality information being more efficient than low-quality information. Either the patient or the specialist benefits from information provision. The degree to which information provision is beneficial is varied between individual beneficiaries.

Two models for describing PCP behavior are considered, a model in which profit enters a utility function linearly and a model in which the PCP evaluates profits with a Prospect Theory type value function including loss aversion (see Kahneman and Tversky, 1979 and Tversky and Kahneman, 1991 for an overview of Prospect Theory in risky and risk-less contexts). Both models predict that a partially altruistic PCP under-provides high-quality information. Bonus payments for information provision are considered as a remedy for the under-provision of information. These bonus payments do not depend on the quality of provided information as it is prohibitively difficult for the payer to verify provided qualities. Thus, the bonus payments impact the PCP's choice in two main ways. Firstly, it increases the profit earned for providing information of any quality compared to the profit earned for not providing any information. This is captured in both the linear and the Prospect Theory model. Secondly, the bonus payments impact the PCP's *perception* of her profits from providing low- and high-quality information in terms of gains and losses. More specifically, if bonus payments are low, the PCP loses profit when providing either information quality, thus perceiving the provision of both qualities as a loss. If bonus payments are medium-sized, the PCP gains profit for providing low-quality information (gain perception) and loses profit when providing high-quality information (loss perception). Lastly, if bonus payments are large, the PCP gains profits for providing either information quality (gain perception). These different perceptions may impact PCP behavior. This aspect is only captured in the Prospect Theory model.

As predicted by the models, PCPs in the experiment pass on more low- and high-quality information as the bonus payment increases. If the bonus payment is at least as high as the costs for the provision of high-quality information, PCPs provide less low-quality information and more high-quality information than in decision tasks with lower bonus payments. This behavioral pattern is in line with the model considering loss aversion in addition to altruism. Moreover, PCPs' reactions to increases in the bonus payment are similar regardless of whether the bonus payment is introduced cost-neutrally or not. If specialists benefit from information provision and earn larger profits than the PCPs, PCPs mainly focus on their own profit and provide less high-quality information than if either patients benefit from information provision or if the PCP earns larger profits.

The remainder of this thesis is organized as follows. Chapter 2 reviews the relevant theoretical and experimental literature on physician behavior and describes how this thesis contributes to it. Chapters 3, 4, and 5 present the individual works of this thesis.

2 Literature Review and Contribution

2.1 Literature Review

This thesis is connected to several strands of the theoretical and experimental health economic literature. The most important strands are discussed in turn. The majority of the theoretical health economic literature is concerned with agency problems concerning physicians and hospitals. Agency problems such as over- and under-treatment, over- and under-referral, and insufficient treatment quality arise due to asymmetric information between physicians and the payer. The aim of many papers, as well as this thesis, is the analysis and derivation of optimal payment systems or forms of organizational design in order to achieve the most socially efficient outcome given the informational and practical restraints. The purpose of the payment system and the organizational design is to internalize the social benefits and costs of physicians' choices[1]. This incentivizes physicians to balance costs and benefits of their diagnostic and therapeutic decisions in such a manner as to maximize social welfare.

Most closely related to this thesis is the literature on physician-initiated patient referrals. Allard, Jelovac, et al. (2011) consider gatekeeping PCPs who may differ in both their altruism and their ability to diagnose patients. More able physicians should put greater trust in their diagnostic result, whereas less able physicians should choose whether to treat or refer blindly. Allard et al. study the referral incentives of three different payment systems: fee-for-service (FFS), capitation, and fund holding. They find that FFS and fund holding tend to result in fewer blind referrals. Thus, rather than incentivizing the economic use of resources, capitations can also lead to increased costs by incentivizing the PCP to increase referrals to specialist care. Furthermore, both fund holding and FFS dominate capitation if physicians are both sufficiently able and altruistic. These results run counter to the usual result in a setting without referrals (see e.g. Chalkley and Malcomson, 1998; Ellis and T. G. McGuire, 1986) that prospective payment systems

[1]See P. Bolton and Dewatripont (2005) for an introduction to contract theory and Heider and Mang (2020) as well as Ma and Mak (2019) for a summary on the incentive effects of different health care payment systems.

such as capitation payments are a cost-containment mechanism, especially for altruistic physicians.

Garcia-Mariñoso and Jelovac (2003) consider endogenous diagnostic effort in addition to referral decisions. Optimal incentives call for a punishment whenever patients are referred to specialist care. Furthermore, gatekeeping performs better than no gatekeeping whenever incentives for the PCP matter. Malcomson (2004) also investigates gatekeeping PCPs with endogenous diagnostic effort. He finds that gatekeeping PCPs who are responsible for the entire cost of care face too strong incentives for cost-containment. Furthermore, if patients can choose between gatekeepers who do and do not receive incentive contracts for cost-containment, the incentive contracts are not useful as patients tend to prefer physicians without an incentive contract.

Allard, Jelovac, et al. (2014) consider PCP self-selection into capitation or FFS contracts. PCPs differ in both their altruism and their ability to diagnose patients. PCP choice of payment contracts is optimal only under specific conditions: the PCPs' abilities needs to be exogenous, the expected loss of failed referrals and the proportion of high-ability PCPs needs to be high. Giving PCPs the ability to self-select into contracts is never optimal if diagnostic effort is endogenous or if PCPs compete for patients.

An extensive strand of the literature is concerned with physician altruism. In their seminal paper, Ellis and T. G. McGuire (1986) add a term to physicians' utility function capturing physicians' altruistic preferences toward the patients in addition to their own profits. They find that if a physician weights the patients' benefits to a lesser degree than her own profits, prospective payment will lead to an under-provision of care. However, if she is paid by retrospective cost-based reimbursement, she will over-provide care. As a solution they propose mixed reimbursements that have a prospective and a retrospective component. Chalkley and Malcomson (1998) extend the analysis of Ellis and T. G. McGuire (1986) to include effort for cost-reduction. As a result they find that both optimal cost-reduction incentives as well as optimal quality provision can only be achieved if physicians are perfectly altruistic, i.e. if they internalize patients' benefits to the same degree as their own profits. In this case, physicians need to be reimbursed prospectively. For partially altruistic physicians, mixed contracts may improve on purely prospective payments by increasing the incentives for quality provision.

Eggleston (2005) considers a physician who can provide two different types of quality, one that can be verified by the payer and another than can not. Too strong incentives for cost control as well as strong incentives for the verifiable quality leads to an under-supply of the non-verifiable quality. Again, mixed payments can help to alleviate this problem. Kaarbøe and Siciliani (2011) consider two quality types that can be either complements or substitutes. They find that stronger performance incentives are optimal if qualities are complements rather than substitutes.

An issue with the optimal contracts from Ellis and T. G. McGuire (1986) and Chalkley and Malcomson (1998) is that they require knowledge of the altruism parameter of the physician. As physician altruism is highly heterogeneous (Godager and Wiesen, 2013), this is unlikely to be the case. Jack (2005) extends Chalkley and Malcomson (1998) by modeling altruism as private information of the physician. He proposes a menu of contracts in which physicians may choose between contracts with a large budget but strong cost sharing or a small budget with weak cost sharing. Similarly, Choné and Ma (2011) also call for a menu of contracts. However, they show that if the payer does not know the patient's valuation of health care, optimal contract menus contain strong pooling. Barham and Milliken (2015) consider the physician's choice of how many patients to treat in addition to the treatment quality per patient. Contrary to Ellis and T. G. McGuire (1986) and Chalkley and Malcomson (1998), they find that altruistic physicians who are not capacity constrained do not have an incentive to over-supply quality. Physicians rather increase the amount of patients treated. Given a constant capitation fee, mixed payments can not alleviate this issue but improve access to care. Furthermore, similarly to Jack (2005) and Choné and Ma (2011), Barham and Milliken (2015) propose menus of contracts to improve the quality of treatment.

There do not exist many instances of menus of contracts being used as a basis for real world health care systems[2]. One reason might be that offering a large menu of contracts is too complex. Liu and Ma (2013) propose an alternative solution to the problem of unverifiable altruism parameters. If physicians are able to commit to a treatment plan before learning the patient's illness severity, a single payment contract can be used to implement socially efficient treatment choices. This contract leaves the physicians no profit whenever they provide the efficient treatments. Furthermore, cost sharing is set to such a level that the least altruistic provider prefers to provide the efficient treatment. Liu, Ma, and Mak (2018) show that a similar contract can implement efficient gatekeeping protocols and time-investment in a partnership with motivated experts.

Similarly to Liu and Ma (2013), Malcomson (2005) considers a physician's choice from a discrete set of treatments. However, he does not consider physician altruism. Patients differ in their disease severity and treatment costs. The payer is able to verify which of two treatments has been provided but is unable to verify whether it was the appropriate treatment for the patient. The implementation of the optimal treatment choice is not always possible. It requires that treatment costs are rising in severity and that the costs of the treatment which is appropriate for less severely ill patients rise faster in patient severity than the costs of the treatment which is appropriate for the more severely ill patients.

Another strand of the literature analyzes the organizational design of agents (Grassi and Ma, 2016; Jelovac and Macho-Stadler, 2002). The focus of these papers is less on the derivation of optimal payment schemes and more on the question of whether organizing agents in a team or partnership is efficiency enhancing. This is also the main focus of Chapter 4 of this thesis. Grassi and Ma (2016) study a referral market between experts who each provide cost-efficient treatment for one of two types of clients. An adverse selection problem can exist because experts do not internalize the other expert's costs. Forming an organization has both positive and negative effects. It is beneficial for referral efficiency as the team members internalize the cost-savings from providing

[2] An exception is the UK fund-holding scheme in the 90s.

appropriate treatment. However, forming an organization reduces incentives for cost control for the referred expert. Jelovac and Macho-Stadler (2002) studies a hospital that may subcontract with its physician. They find that hospital subcontracting may be superior to contracting with both parties simultaneously. This is the case if the hospital's investment decisions matter sufficiently much for the quality of care.

A small strand of the literature uses dynamic models to analyze physician behavior from a theoretical perspective. Hey and Patel (1983) and Hennessy (2008) analyze prevention and cure investments of an individual by means of Markov models. Patients can be in one of two health states —healthy or sick. While they are healthy they can invest into prevention in order to decrease the probability of their health state deteriorating. While they are sick they can invest into curative measures in order to increase the probability of their health state improving. The main finding of Hennessy (2008) is that prevention and cure can be both complements and substitutes. Strong investments into prevention improve the effectiveness of curative investments as returning to the healthy state is more profitable. However, strong investments into curative measures lowers the effectiveness of prevention as sick individuals will be quickly returned to a healthy state anyways. Prevention and curative efforts of physicians are studied in Chapter 4 of this thesis. Another paper using a dynamic setup is Allard, Léger, et al. (2009). They study repeated provider competition and find that competition can create effort incentives. However, switching costs between physicians lowers the effectiveness of competition.

The predictions of the theoretical literature have been tested in laboratory experiments. A large part of this literature is focused on the impact of physician altruism on the provision of health care. Hennig-Schmidt, Selten, et al. (2011) consider treatment quantity choices of students in the role of physicians and find that participants react to changes in the payment system. Under FFS, they provide more treatment quantity than under capitation payments. However, they also find that physicians take into account the patients well-being to a significant degree. Brosig-Koch et al. (2016b) confirm these results for both medical students and medical doctors. Brosig-Koch et al. (2017) experimentally confirms the theoretical result that mixed payments outperform pure FFS or capitation in terms of social welfare. Waibel and Wiesen (2021) study the impact of a referral fee on PCP referral behavior. They find that referral fees increase the number of referred patients for barely altruistic physicians, which can enhance welfare. However, contrary to their theoretical prediction, they do not find an effect on the physicians' diagnostic efforts. Kesternich et al. (2015) consider the impact of professional norms on medical decision making. They find that not only financial incentives from payment systems affect physician behavior. In particular, confronting participants with professional norms derived from the Hippocratic Oath decreases their self-interest and efficiency concerns.

Kerschbamer, Sutter, and Dulleck (2017) consider social preferences of providers in a general credence goods market (see Darby and Karni, 1973, further Dulleck and Kerschbamer, 2006 and Kerschbamer and Sutter, 2017 for surveys). In their model, consumers can verify the quality they receive. In this setting, the standard model of selfish agents who exhibit a weak taste for efficiency 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. However, the authors find that only about a fourth of participants behaved in this manner. Instead, participants exhibited both inequality aversion as well as a strong taste for efficiency. Consequently, institutions designed under the assumption of providers being weakly efficient profit-maximizers may yield sub-optimal results.

2.2 Contribution

2.2.1 Chapter 3: "Diagnostics and Treatment: On the Division of Labor between Primary Care Physicians and Specialists"

Similarly to Kerschbamer, Sutter, and Dulleck (2017), Chapter 3 considers a situation in which a selfish agent with a weak taste for efficiency acts in a socially efficient manner. Specifically, if they receive equal markups regardless of which physician treats the patient, selfish physicians would always choose the efficient treatment path. Altruistic PCPs, on the other hand, over-refer patients to specialists, whereas altruistic specialists do not back-refer patients who only benefit marginally from their treatment. Physician altruism can be a desirable quality as it can simplify the incentivization of effort provision (see Chalkley and Malcomson, 1998). In contrast, Chapter 3 highlights the two-sided nature of altruism. It further contributes by deriving payment contracts that solve this problem. These optimal contracts call for markups for either the immediate treatment of the patient by the PCP or for the back-referral to the PCP by the specialist. Furthermore, Chapter 3 shows that these markups also generate welfare improvements if the payer has limited knowledge on the effectiveness of specialist treatment.

Chapter 3 contributes to the literature on physician-initiated referrals. This strand of the literature has focused on the incentives of PCPs and has mostly not considered incentives affecting specialists' referral decisions[3]. In particular, Chapter 3 contributes by analyzing the strategic interactions between PCP and specialist. This analysis yields the new result that the PCP can be indirectly incentivized by markups for the specialist. This may be desirable for the payer because these markups can be more rent-efficient than markups for the PCP.

[3] Godager, Iversen, et al. (2015) consider referral incentives for a specialist. However, they do not derive optimal payment schemes.

2.2.2 Chapter 4: "Should Physicians Team Up to Treat Chronic Diseases?"

Similarly to Chapter 3, Chapter 4 contributes to the literature on physician-initiated referrals. Both treatment- and effort decisions have been studied simultaneously by Garcia-Mariñoso and Jelovac (2003) before. However, there are three significant differences between the models. First, Chapter 4 considers chronic patients with the help of a dynamic model. To the best of my knowledge, there does not exist another theoretical analysis on this topic that uses a dynamic approach. Second, in addition to deriving optimal treatment fees, Chapter 4 compares two forms of organizational design —solo practices and a physician team. Third, Chapter 4 does not assume that the patient's state is revealed at any point in the game. This allows for the derivation of more realistic payment contracts that do not rely on information about the patient's state. Grassi and Ma (2016) also consider different forms of organizational design. However, they do not consider a dynamic model and do not derive optimal payments.

The main new results are as follows. First, it is possible for the solo practices to outperform the team, even though there is no problem of asymmetric information between team members. This is a consequence of the fact that, contrary to the setting in Grassi and Ma (2016), the cheapest treatment may not always be the socially efficient treatment. Instead, patient health losses from improper treatment are also considered. Thus, the strong cost-efficiency of the team may turn into a disadvantage for the payer. Second, given that team treatment is optimal, optimal payments call for markups for PCP care and cost sharing for specialist care. This makes it more profitable for the team to exert effort in order to keep patients in mild condition.

2.2.3 Chapter 5: "Rewards for Information Provision in Patient Referrals: A Theoretical Model and an Experimental Test"

Chapter 5 contributes both to the theoretical and the experimental literature. In the theoretical part, the effect of a bonus payment for a PCP's provision of medical information is analyzed. Though financial incentives affecting referral behavior of physicians have been considered before, incentives affecting information exchange between physicians has, to the best of my knowledge, not been examined explicitly. The experiment reveals, consistently with empirical evidence, that too little information is provided and that this information is of low quality. Introducing the bonus payment induces the provision of both low- and high quality information. Increasing the bonus payment from just covering the cost of low-quality information to at least covering the costs of high-quality information, results in an increase in high-quality information and a decrease in low-quality information. The theoretical model based on Prospect Theory delivers an explanation for this behavior. However, a welfare analysis indicates that large bonus payments are only efficient from the payer's perspective if the bonus payment is introduced cost-neutrally and if the benefits from information provision are large. This result is useful for the practical implementation of these bonus payments.

In addition to the patient, the specialist is considered as a beneficiary from information provision as well. This is new in the experimental literature. The experiment shows that, compared to all other conditions, PCPs are acting in a significantly less altruistic manner if specialists both benefit from information provision and earn larger profits than the PCP.

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

Malte Griebenow and Mathias Kifmann

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.

3.1 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 (M. L. 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, Jelovac, 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 3.2 discusses related literature. In Section 3.3, we present the model. In Section 3.4, we characterize the first-best division of labor between PCPs and specialists. In Section 3.5, we derive optimal contracts, given that the payer does not observe diagnostic signals. In Section 3.6, 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 3.7 concludes.

3.2 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, Jelovac, 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, Jelovac, 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.

[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.

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.

3.3 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, 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.

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

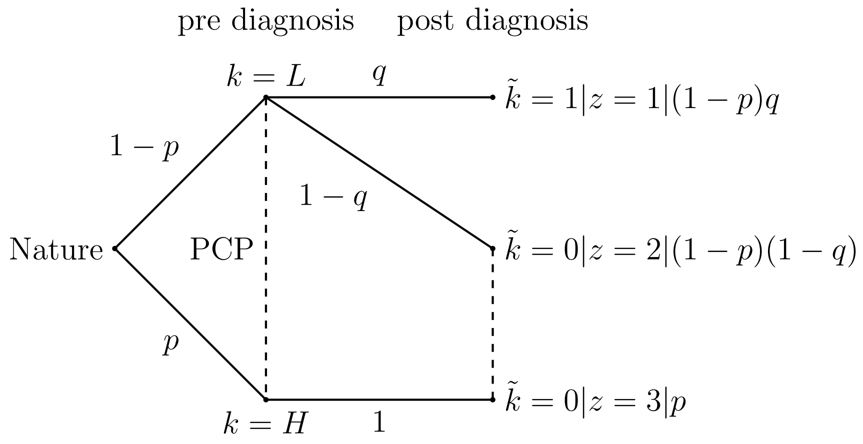


Figure 3.1: Cases z, dashed lines indicate that the PCP can not differentiate between these cases and therefore cannot identify a high-type patient with certainty

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 3.6.3, 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 $z \in \{1, 2, 3\}$:

$$z = \begin{cases} 1, & k = L,\ \tilde{k} = 1 \\ 2, & k = L,\ \tilde{k} = 0 \\ 3, & k = H,\ \tilde{k} = 0 \end{cases} \tag{3.1}$$

[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 3.1 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 3.6.2 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

$$\vec{T} = \begin{pmatrix} T_1 & T_2 & T_3 \end{pmatrix}^T \tag{3.2}$$

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:

$$U_j^k(T_z) = \beta_j b_k^{T_z} + \Pi_j(T_z), \; j \in \{P, S\}, \; k \in \{L, H\} \tag{3.3}$$

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 (3.3) 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 (3.3). However, if she can not identify the patient's type, she chooses treatment T^0 so as to maximize her conditional expected utility

$$\mathbf{E}U_P(\tilde{k}=0) = p_L^0 U_P^L(T^0) + p_H^0 U_P^H(T^0),\ j \in \{P, S\}, \tag{3.4}$$

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:

$$\mathbf{E}\Pi_j = q(1-p)\Pi_j(T_1^*) + (1-p)(1-q)\Pi_j(T_2^*) + p\Pi_j(T_3^*) \geq 0,\ j \in \{P, S\}, \tag{3.5}$$

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,

$$\mathbf{E}W = \mathbf{E}b - \mathbf{E}\Pi - \mathbf{E}c. \tag{3.6}$$

Figure 3.2 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

[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.

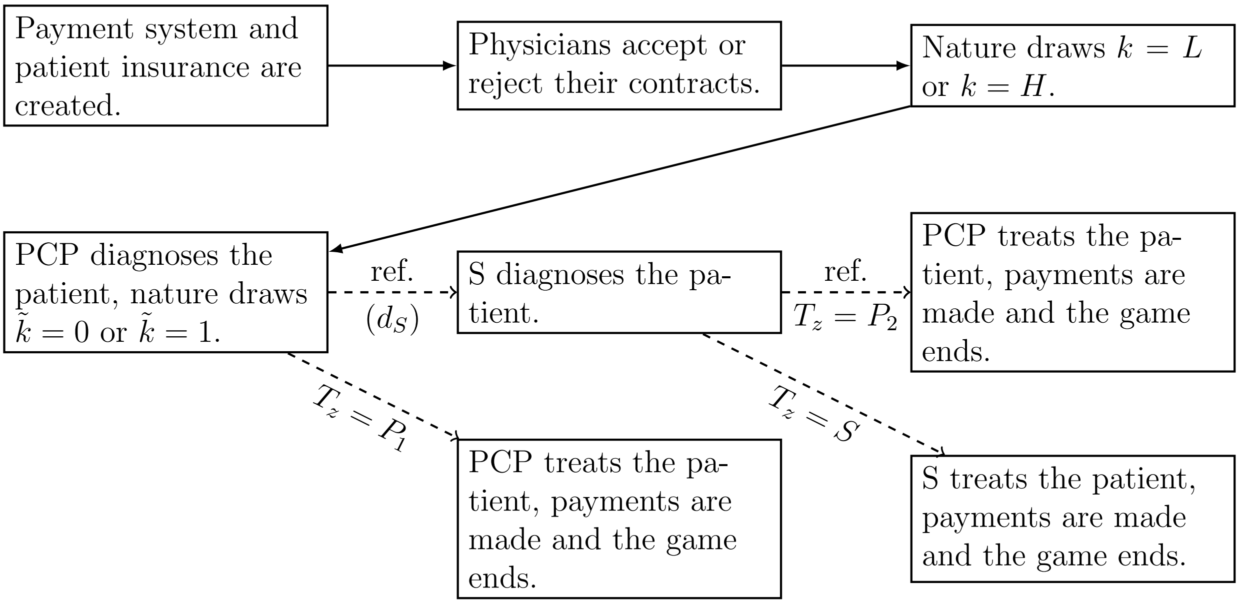


Figure 3.2: Sequence of Events

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.

3.4 First-Best Benchmark

In this section, we derive the first-best treatment paths $\vec{T}^{FB}$ which maximize patient welfare (3.6) 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 3.A.1, we prove Theorem 1 which derives the first-best treatment path vectors.

[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.

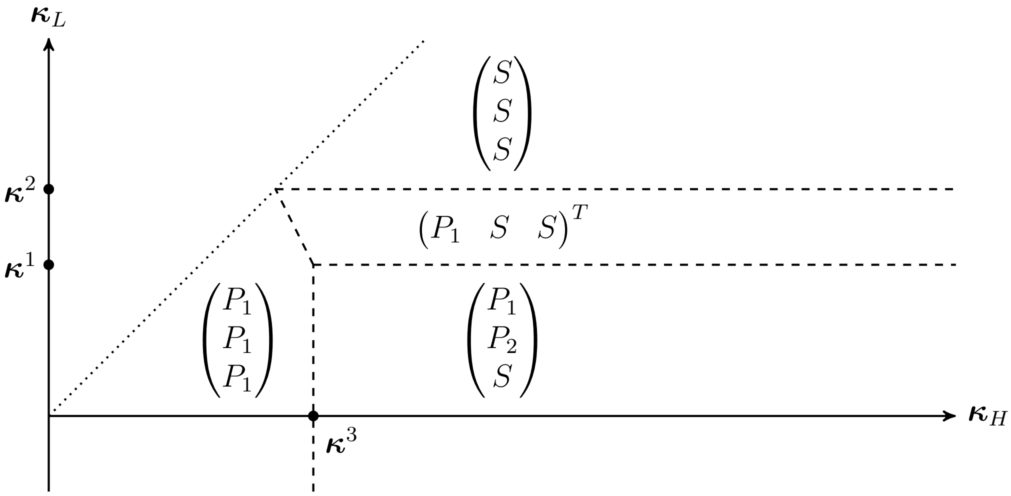


Figure 3.3: First-best (dashed lines) treatment path vector depending on the specialist surplus benefit. Dotted line indicates $\kappa_H = \kappa_L$. $\kappa^1 = c_S - c_P$, $\kappa^2 = d_S + c_S - c_P$, $\kappa^3 = c_S - c_P + d_S/p_H^0$.

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

$$\vec{T}^{FB} = \begin{cases} (P_1\ P_1\ P_1)^T \text{ if} & p_L^0\kappa_L + p_H^0\kappa_H \leq d_S + c_S - c_P, \\ & \kappa_H \leq c_S - c_P + d_S/p_H^0; \\ (P_1\ P_2\ S)^T \text{ if} & \kappa_H \geq c_S - c_P + d_S/p_H^0, \\ & \kappa_L \leq c_S - c_P; \\ (P_1\ S\ S)^T \text{ if} & c_S - c_P \leq \kappa_L \leq d_S + c_S - c_P, \\ & p_L^0\kappa_L + p_H^0\kappa_H \geq d_S + c_S - c_P; \\ (S\ S\ S)^T \text{ if} & \kappa_L \geq d_S + c_S - c_P. \end{cases} \tag{3.7}$$

Figure 3.3 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.
- 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.

3.5 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 (3.7) at minimal rents for the physicians. We impose the *non-negative* *payments* condition

$$\text{NNP: } \gamma_j^T \geq 0,\ j \in \{P, S\},\ T \in \{P_1, P_2, S\}. \tag{3.8}$$

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 (3.9).

$$\begin{aligned}
IC_P^1 &: \gamma_P^{P_1} \geq \gamma_P^{P_2} \\
IC_P^2 &: \gamma_P^{P_1} - c_P \leq p_H^0(\beta_P \kappa_H + \gamma_P^S) + p_L^0(\gamma_P^{P_2} - c_P) \\
PC_P &: (1-p)q(\gamma_P^{P_1} - c_P) + (1-p)(1-q)(\gamma_P^{P_2} - c_P) \\
&\quad + p\gamma_P^S \geq 0 \\
IC_S^1 &: \beta_S \kappa_H + \gamma_S^S - c_S \geq \gamma_S^{P_2} \\
IC_S^2 &: \beta_S \kappa_L + \gamma_S^S - c_S \leq \gamma_S^{P_2} \\
PC_S &: (1-p)(1-q)(\gamma_S^{P_2} - d_S) + p(\gamma_S^S - c_S - d_S) \geq 0 \\
NNP &: \gamma_j^T \geq 0
\end{aligned} \tag{3.9}$$

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.

[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 $(P_1\ S\ S)^T$.

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 3.4 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 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 3.4 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 3.4a. 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 3.4b). 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 3.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 3.3).

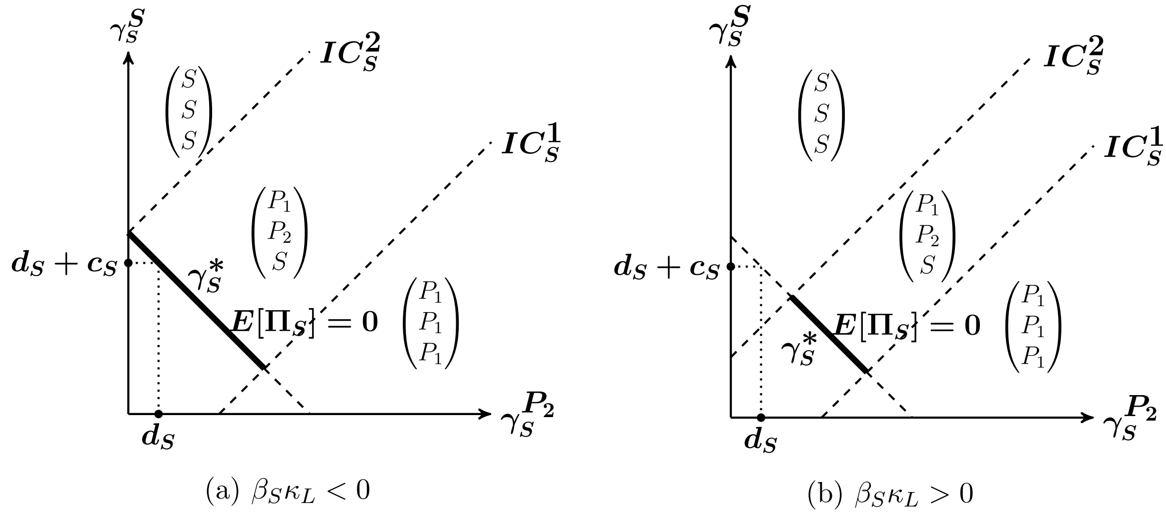


Figure 3.4: Second-best optimal contracts without rent payments to the specialist that implement $(P_1\ P_2\ S)^T$ given $(\gamma_P^{P_1}, \gamma_P^{P_2}, \gamma_P^S) = (c_P, c_P, 0)$. Dashed lines indicate incentive- and participation constraints, the solid and thick line indicates the set of optimal contracts.

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 PCP and assumed the specialist to behave independently of economic incentives (see for example Allard, Jelovac, 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.

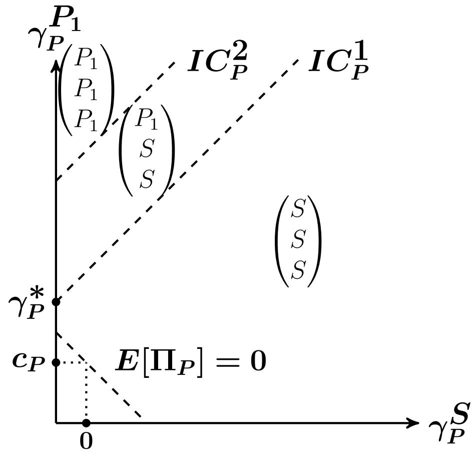


Figure 3.5: Second-best optimal contract for the PCP that implements $(P_1\ S\ S)^T$ given $(\gamma_S^{P_2}, \gamma_S^S) = (d_S, d_S + c_S)$ and $\beta_P \kappa_L > 0$. Dashed lines indicate incentive- and participation constraints.

Finally, we derive an optimal contract that implements $(P_1\ S\ S)^T$. This treatment path can be implemented by fulfilling System of Inequations (3.10).

$$\begin{aligned}
IC_P^1 &: \gamma_P^{P_1} - c_P \geq \beta_P \kappa_L + \gamma_P^S \\
IC_P^2 &: \gamma_P^{P_1} - c_P \leq \beta_P (p_H^0 \kappa_H + p_L^0 \kappa_L) + \gamma_P^S \\
PC_P &: (1-p)q(\gamma_P^{P_1} - c_P) + [(1-p)(1-q) + p]\gamma_P^S \geq 0 \\
IC_S^1 &: \beta_S \kappa_H + \gamma_S^S - c_S \geq \gamma_S^{P_2} \\
IC_S^2 &: \beta_S \kappa_L + \gamma_S^S - c_S \geq \gamma_S^{P_2} \\
PC_S &: \gamma_S^S - c_S - d_S \geq 0 \\
NNP &: \gamma_j^T \geq 0
\end{aligned} \tag{3.10}$$

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 3.3). 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 3.A.3).

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

$$\begin{aligned} \gamma_P^{P_1*} &= c_P + \beta_P \kappa_L \\ \gamma_P^{P_2*} &= 0 \\ \gamma_P^{S*} &= 0 \\ \gamma_S^{P_2*} &= d_S \\ \gamma_S^{S*} &= d_S + c_S \end{aligned} \tag{3.11}$$

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 3.5 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 T. G. 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 3.A.4).

Theorem 4. *If* $0 < \beta_P \leq \beta_P^* := \frac{(1-q)d_S}{c_S - c_P + q d_S}$*, 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.*

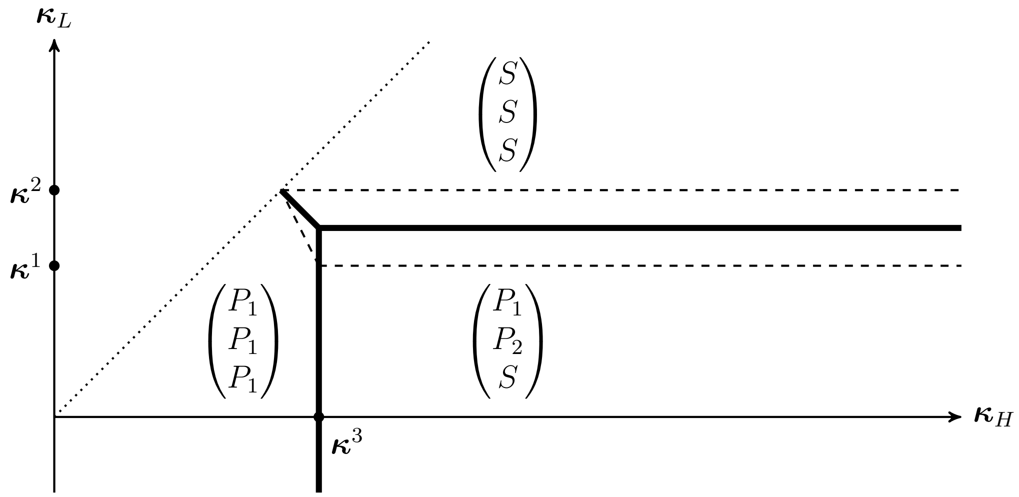


Figure 3.6: Second-best (solid and thick) vs. first-best (dashed) treatment path vector depending on the specialist surplus benefits, $\beta_P \geq \beta_P^*$.

Figure 3.6 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.

3.6 Extensions

In this section, we explore the impact of a change in the model's assumptions. In Subsection 3.6.1 we drop the assumption that the payer has perfect knowledge about the size of the surplus benefits κ_k. In Subsection 3.6.2, we study the effect of referral costs for the patient. In Subsection 3.6.3, we explore an alternative assumption on the information structure of the game.

3.6.1 Uncertain Surplus Benefits

The second-best optimal payment contracts derived in Section 3.5 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 3.5:

(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 3.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 3.7 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, $\mathbf{OT^j}$

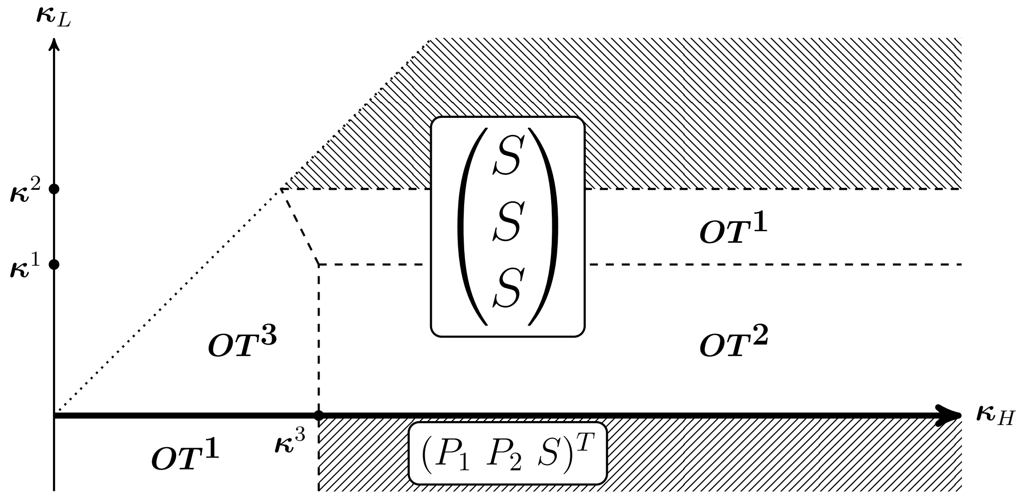


Figure 3.7: Contract (I) – Cost-based FFS (solid and thick) vs first-best (dashed) treatment path vectors.

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

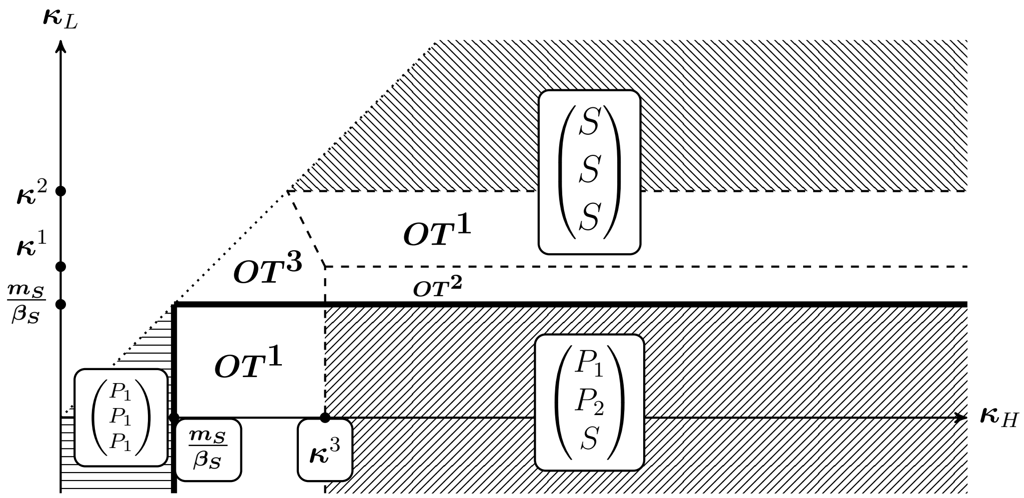


Figure 3.8: Contract (II) – Treatment path vectors resulting from a markup m_S for specialist referral (solid and thick) vs first-best (dashed) treatment path vectors. Shaded regions indicate congruence between first-best and treatment paths resulting from markup.

for specialist referrals can help to alleviate over-treatment. Figure 3.8 shows the impact of the introduction of a small markup for specialist back-referral. As explained in Section 3.5, 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 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 3.8 to Figure 3.7, 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 3.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).

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 3.9 ($\boldsymbol{UT}$ 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 3.7 to Figure 3.9 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

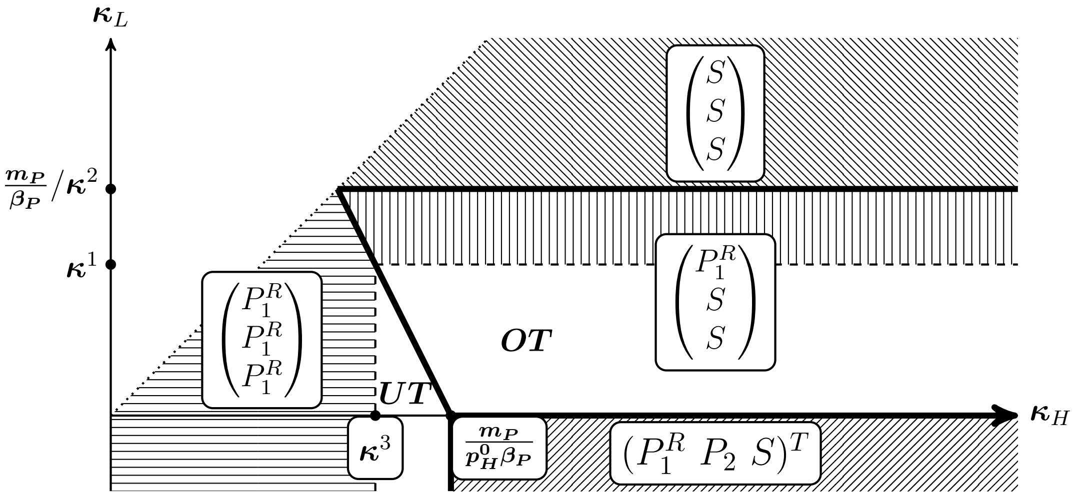


Figure 3.9: Contract (III) – Markup for PCP treatment without referral (solid and thick) vs first-best (dashed) treatment path vectors.

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 3.A.7).

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.

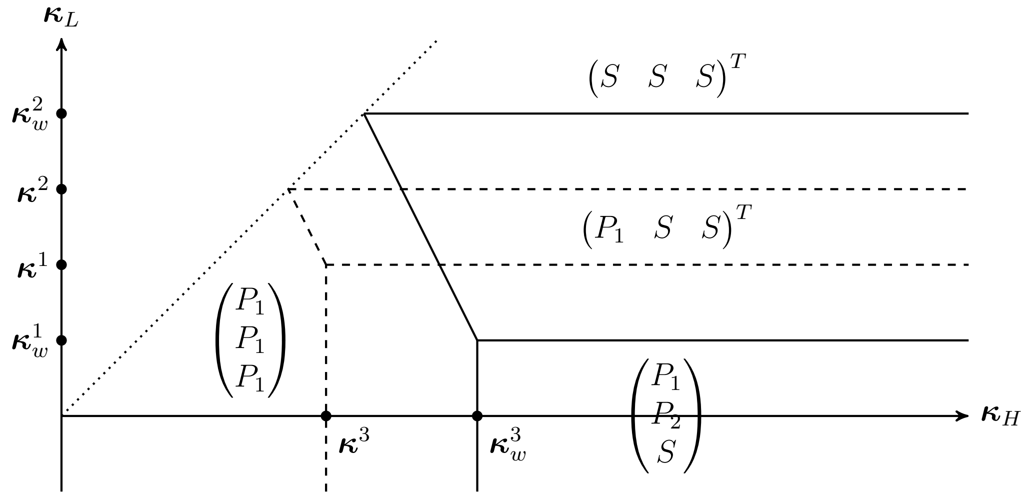


Figure 3.10: First-best treatment path vector with (regular lines) and without (dashed lines) waiting costs depending on the specialist surplus benefit. Dotted line indicates $\kappa_H = \kappa_L$. $\kappa_w^1 := c_S - c_P - w$, $\kappa_w^2 := d_S + c_S - c_P + w$, and $\kappa_w^3 := d_S/p_H^0 + (\frac{2p_L^0}{p_H^0} + 1)w + c_S - c_P$.

3.6.2 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 3.10. 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 $\begin{pmatrix} P_1 & P_1 & P_1 \end{pmatrix}^T$ 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 (3.9):

$$\begin{aligned}
IC_P^{1,w} &: \gamma_P^{P_1} \geq \gamma_P^{P_2} - 2\beta_P w \\
IC_P^{2,w} &: \gamma_P^{P_1} - c_P \leq p_L^0(\gamma_P^{P_2} - 2\beta_P w - c_P) + p_H^0[\gamma_P^S + \beta_P(\kappa_H - w)] \\
IC_S^{1,w} &: \gamma_S^S + \beta_S\kappa_H - c_S \geq \gamma_S^{P_2} - \beta_S w \\
IC_S^{2,w} &: \gamma_S^S + \beta_S\kappa_L - c_S \leq \gamma_S^{P_2} - \beta_S w
\end{aligned} \tag{3.12}$$

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 3.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:*

$$\begin{aligned}
\gamma_P^{P_1*} &= c_P \\
\gamma_P^{P_2*} &= c_P \\
\gamma_P^{S*} &= 0 \\
\gamma_S^{S*} &= d_S + c_S + p_L^0 \min(-\beta_S(\kappa_L + w), 0) \\
\gamma_S^{P_2*} &= d_S + p_H^0 \max(\beta_S(\kappa_L + w), 0)
\end{aligned} \tag{3.13}$$

To implement $\begin{pmatrix} P_1 & S & S \end{pmatrix}^T$ with waiting time costs, the following incentive constraints change compared to (3.10):

$$\begin{aligned} IC_P^{1,w} &: \gamma_P^{P_1} - c_P \geq \gamma_P^S + \beta_P(\kappa_L - w) \\ IC_P^{2,w} &: \gamma_P^{P_1} - c_P \leq \gamma_P^S + \beta_P[p_L^0(\kappa_L - w) + p_H^0(\kappa_H - w)] \\ IC_S^{1,w} &: \gamma_S^S + \beta_S\kappa_H - c_S \geq \gamma_S^{P_2} - \beta_S w \\ IC_S^{2,w} &: \gamma_S^S + \beta_S\kappa_L - c_S \geq \gamma_S^{P_2} - \beta_S w \end{aligned} \tag{3.14}$$

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 3.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

$$\begin{aligned} \gamma_P^{P_1*} &= c_P + \beta_P(\kappa_L - w) \\ \gamma_P^{P_2*} &= 0 \\ \gamma_P^{S*} &= 0 \\ \gamma_S^{S*} &= d_S \\ \gamma_S^{P_2*} &= d_S + c_S \end{aligned} \tag{3.15}$$

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.

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.

3.6.3 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 3.11 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 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 3.12 depicts the first-best under the new information structure.

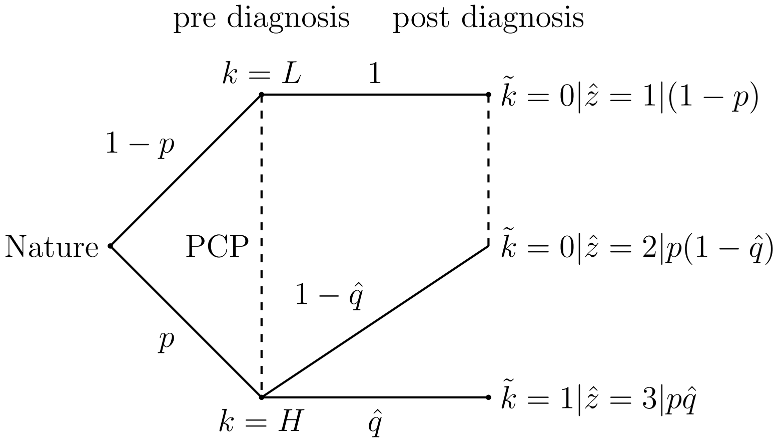


Figure 3.11: Cases z if the PCP can identify high-type patients, dashed lines indicate that the PCP can not differentiate between these cases.

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

$$\begin{aligned} \kappa^2 = c_S - c_P + d_S \leq \kappa_H \leq c_S - c_P + \frac{d_S}{\hat{p}_H^0} =: \hat{\kappa}^3, \\ \kappa^2 \geq \hat{p}_L^0 \kappa_L + \hat{p}_H^0 \kappa_H. \end{aligned} \tag{3.16}$$

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

$$\begin{aligned} \kappa_H &\geq \hat{\kappa}^3, \\ \kappa_L &\leq c_S - c_P = \kappa^1. \end{aligned} \tag{3.17}$$

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

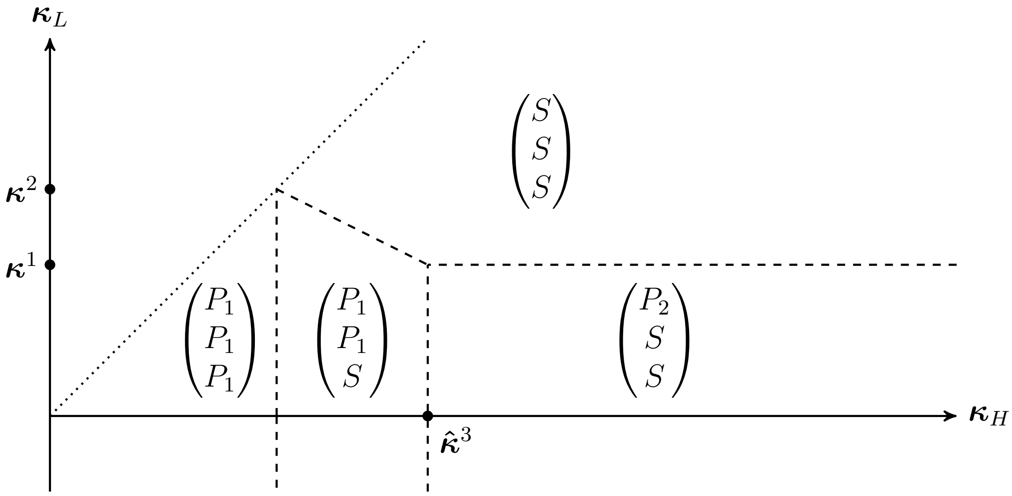


Figure 3.12: First-best (dashed lines) treatment path vector given the alternative information structure. Dotted line indicates $\kappa_H = \kappa_L$. $\hat{\kappa}^3 := c_S - c_P + \frac{d_S}{\hat{p}_H^0}$.

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

$$\begin{aligned}
IC_P^1 &: \hat{p}_L^0(\gamma_P^{P_2} - c_P) + \hat{p}_H^0(\gamma_P^S + \beta_P \kappa_H) \geq \gamma_P^{P_1} - c_P \\
IC_P^2 &: \gamma_P^S + \beta_P \kappa_H \geq \gamma_P^{P_1} - c_P \\
PC_P &: (1-p)(\gamma_P^{P_2} - c_P) + p\gamma_P^S \geq 0 \\
IC_S^1 &: \gamma_S^{P_2} \geq \beta_S \kappa_L + \gamma_S^S - c_S \\
IC_S^2 &: \gamma_S^{P_2} \leq \beta_S \kappa_H + \gamma_S^S - c_S \\
PC_S &: (1-p)(\gamma_S^{P_2} - d_S) + p(\gamma_S^S - c_S - d_S) \geq 0 \\
NNP &: \gamma_j^T \geq 0
\end{aligned} \tag{3.18}$$

Theorem 9 (proof in Appendix 3.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*

$$\begin{aligned}
{\gamma_P^{P_1}}^* &= c_P \\
{\gamma_P^{P_2}}^* &= c_P \\
{\gamma_P^S}^* &= 0 \\
{\gamma_S^{P_2}}^* &= d_S + p\beta_S \kappa_L \\
{\gamma_S^S}^* &= d_S + c_S - (1-p)\beta_S \kappa_L
\end{aligned} \tag{3.19}$$

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 3.6.1 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 13 in Appendix 3.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

$$\begin{aligned}
IC_P^1 &: \gamma_P^{P_1} - c_P \geq \hat{p}_L^0(\gamma_P^{P_2} - c_P) + \hat{p}_H^0(\gamma_P^S + \beta_P \kappa_H) \\
IC_P^2 &: \gamma_P^{P_1} - c_P \leq \gamma_P^S + \beta_P \kappa_H \\
PC_P &: (1 - p\hat{q})(\gamma_P^{P_1} - c_P) + p\hat{q}\gamma_P^S \geq 0 \\
IC_S^1 &: \gamma_S^{P_2} \geq \gamma_S^S + \beta_S \kappa_L - c_S \\
IC_S^2 &: \gamma_S^{P_2} \leq \gamma_S^S + \beta_S \kappa_H - c_S \\
PC_S &: \gamma_S^S - c_S - d_S \geq 0 \\
NNP &: \gamma_j^T \geq 0
\end{aligned} \tag{3.20}$$

and in the second case

$$\begin{aligned}
IC_P^1 &: \gamma_P^{P_1} - c_P \geq \hat{p}_L^0(\gamma_P^S + \beta_P\kappa_L) + \hat{p}_H^0(\gamma_P^S + \beta_P\kappa_H) \\
IC_P^2 &: \gamma_P^{P_1} - c_P \leq \gamma_P^S + \beta_P\kappa_H \\
PC_P &: (1 - p\hat{q})(\gamma_P^{P_1} - c_P) + p\hat{q}\gamma_P^S \geq 0 \\
IC_S^1 &: \gamma_S^{P_2} \leq \gamma_S^S + \beta_S\kappa_L - c_S \\
IC_S^2 &: \gamma_S^{P_2} \leq \gamma_S^S\beta_S\kappa_H - c_S \\
PC_S &: \gamma_S^S - c_S - d_S \geq 0 \\
NNP &: \gamma_j^T \geq 0.
\end{aligned} \tag{3.21}$$

It follows Theorem 10 (more details and proof in Appendix 3.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.

Concluding, 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.

3.7 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.

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 counterintuitive 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.

Appendices

3.A Proof of Theorems and Lemmas

3.A.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*

$$\vec{T}^1 \succ (\succeq) \vec{T}^2. \tag{3.22}$$

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

$$\left\{ \begin{pmatrix} P_1 \\ P_1 \\ P_1 \end{pmatrix}, \begin{pmatrix} P_1 \\ S \\ P_2 \end{pmatrix}, \begin{pmatrix} P_1 \\ P_2 \\ S \end{pmatrix}, \begin{pmatrix} P_1 \\ S \\ S \end{pmatrix}, \begin{pmatrix} S \\ P_1 \\ P_1 \end{pmatrix}, \begin{pmatrix} S \\ S \\ P_2 \end{pmatrix}, \begin{pmatrix} S \\ S \\ S \end{pmatrix} \right\}. \tag{3.23}$$

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

$$\begin{aligned}(T_1\ S\ P_2)^T \succeq (T_1\ P_1\ P_1)^T &\iff \kappa_L \geq c_S - c_P + d_S[1 + \frac{p}{(1-p)(1-q)}],\\ \text{and } \kappa_H > \kappa_L &\implies (T_1\ S\ S)^T \succ (T_1\ S\ P_2)^T.\end{aligned} \tag{3.24}$$

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

$$\kappa_H > \kappa_L \geq d_S + c_S - c_P. \tag{3.25}$$

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:

$$\vec{T}^{FB} = \begin{cases} (P_1\ P_1\ P_1)^T \text{ if } & (P_1\ P_1\ P_1)^T \succeq (P_1\ S\ S)^T,\\ & (P_1\ P_1\ P_1)^T \succeq (P_1\ P_2\ S)^T;\\ (P_1\ P_2\ S)^T \text{ if } & (P_1\ P_2\ S)^T \succeq (P_1\ P_1\ P_1)^T,\\ & (P_1\ P_2\ S)^T \succeq (P_1\ S\ S)^T;\\ (P_1\ S\ S)^T \text{ if } & (P_1\ S\ S)^T \succeq (P_1\ P_2\ S)^T\\ & (P_1\ S\ S)^T \succeq (S\ S\ S)^T,\\ & (P_1\ S\ S)^T \succeq (P_1\ P_1\ P_1)^T;\\ (S\ S\ S)^T \text{ if } & (S\ S\ S)^T \succeq (P_1\ S\ S)^T. \end{cases} \tag{3.26}$$

It is simple to verify that the missing inequations are already implied by (3.26):

$$\begin{aligned}(P_1\ P_1\ P_1)^T \succeq (P_1\ S\ S)^T &\implies \kappa_L < d_S + c_S - c_P\\ &\implies (P_1\ S\ S)^T \succ (S\ S\ S)^T,\\ (P_1\ P_2\ S)^T \succeq (P_1\ S\ S)^T &\iff \kappa_L < c_S - c_P\\ &\implies (P_1\ S\ S)^T \succ (S\ S\ S)^T,\\ (S\ S\ S)^T \succeq (P_1\ S\ S)^T &\iff \kappa_L \geq d_S + c_S - c_P\\ &\implies (S\ S\ S)^T \succeq (P_1\ P_2\ S)^T, (P_1\ P_1\ P_1)^T.\end{aligned} \tag{3.27}$$

□

3.A.2

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

Specialist:

The specialist's participation constraint is fulfilled:

$$\begin{aligned}\mathbf{E}\Pi_S &= p(\gamma_S^{S^*} - c_S - d_S) + (1-p)(1-q)(\gamma_S^{P_2{}^*} - d_S) \\ &= pp_L^0 \min(-\beta_S\kappa_L, 0) \\ &- (1-p)(1-q)p_H^0 \min(-\beta_S\kappa_L, 0) \\ &= 0.\end{aligned} \tag{3.28}$$

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 $\gamma_S^{S^*} > 0$.

□

3.A.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.

□

3.A.4

Theorem 11. *If* $\beta_P \leq \beta_P^* := \frac{(1-q)d_S}{c_S - c_P + q d_S}$, *the second-best vector of treatment paths* $\vec{T}^{SB}$ *is given by*

$$\vec{T}^{SB} = \begin{cases} (P_1\ P_1\ P_1)^T \text{ if } & (1-p)[(1-q) - q\beta_P]\kappa_L + p\kappa_H \leq \\ & [(1-p)(1-q)+p](d_S + c_S - c_P), \\ & \kappa_H \leq c_S - c_P + d_S/p_H^0, \\ & (1-p)\kappa_L + p\kappa_H \leq d_S + c_S - c_P; \\ (P_1\ P_2\ S)^T \text{ if } & \kappa_H \geq c_S - c_P + d_S/p_H^0, \\ & \kappa_L \leq \frac{(1-q)(c_S - c_P)/(1-p)}{1-q-\beta_P q}; \\ (P_1\ S\ S)^T \text{ if } & \kappa_L \geq \frac{(1-q)(c_S - c_P)/(1-p)}{1-q-\beta_P q}, \\ & \kappa_L \leq (d_S + c_S - c_P)/(1+\beta_P), \\ & (1-p)[(1-q) - q\beta_P]\kappa_L + p\kappa_H \geq \\ & [(1-p)(1-q)+p](d_S + c_S - c_P); \\ (S\ S\ S)^T \text{ if } & \kappa_L \geq (d_S + c_S - c_P)/(1+\beta_P), \\ & (1-p)\kappa_L + p\kappa_H \geq d_S + c_S - c_P, \end{cases} \tag{3.29}$$

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

$$\vec{T}^{SB} = \begin{cases} (P_1\ P_1\ P_1)^T \text{ if } & (1-p)\kappa_L + p\kappa_H \leq d_S + c_S - c_P, \\ & \kappa_H \leq c_S - c_P + d_S/p_H^0; \\ (P_1\ P_2\ S)^T \text{ if } & \kappa_H \geq c_S - c_P + d_S/p_H^0, \\ & \kappa_L \leq (q d_S + c_S - c_P); \\ (S\ S\ S)^T \text{ if } & \kappa_L \geq (q d_S + c_S - c_P), \\ & (1-p)\kappa_L + p\kappa_H \geq d_S + c_S - c_P. \end{cases} \tag{3.30}$$

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:

$$\begin{aligned}
&(P_1\ S\ S)^T \succeq (S\ S\ S)^T \iff \kappa_L \leq (d_S + c_S - c_P)/(1+\beta_P) \\
&(P_1\ S\ S)^T \succeq (P_1\ P_2\ S)^T \iff \\
&\qquad q(-\beta_P \kappa_L) + (1-q)[\kappa_L - c_S + c_P] \geq 0 \iff \\
&\qquad \kappa_L \geq \frac{(1-q)(c_S - c_P)/(1-p)}{1 - q - \beta_P q}
\end{aligned} \tag{3.31}$$

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

$$\begin{aligned}
&(d_S + c_S - c_P)/(1+\beta_P) \geq \frac{(1-q)(c_S - c_P)/(1-p)}{1 - q - \beta_P q} \iff \\
&(1+\beta_P)[(1-q)(c_S - c_P)/(1-p)] \leq \\
&\quad (d_S + c_S - c_P)[1 - q - \beta_P q] \iff \\
&\qquad \beta_P[c_S - c_P + q d_S] \leq (1-q) d_S \iff \\
&\qquad \beta_P \leq \frac{(1-q) d_S}{c_S - c_P + q d_S}
\end{aligned} \tag{3.32}$$

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

$$\begin{aligned}
&(P_1\ P_1\ P_1)^T \succeq (S\ S\ S)^T \iff (1-p)(d_S + c_S - c_P - \kappa_L) \\
&\qquad + p(d_S + c_S - c_P - \kappa_H) \geq 0 \iff \\
&\qquad (1-p)\kappa_L + p\kappa_H \leq d_S + c_S - c_P, \\
&(P_1\ P_1\ P_1)^T \succeq (P_1\ P_2\ S)^T \iff p(\kappa_H - d_S - c_S + c_P) \\
&\qquad + (1-p)(1-q)(-d_S + c_P) \leq 0 \iff \\
&\qquad \kappa_H \leq c_S - c_P + d_S/p_H^0, \\
&(P_1\ P_2\ S)^T \succeq (S\ S\ S)^T \iff q(-\kappa_L + c_S - c_P + d_S) \\
&\qquad + (1-q)[-\kappa_L + c_S - c_P] \leq 0 \iff \\
&\qquad \kappa_L \leq q d_S + c_S - c_P.
\end{aligned} \tag{3.33}$$

□

3.A.5

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

$$
\begin{aligned}
&(P_1\ P_1\ P_1)^T,\ \textit{if} \begin{cases} m_S \geq \beta_S \kappa_H & \textit{or} \\ \beta_S \kappa_H \geq m_S \geq \beta_S \kappa_L \textit{and } m_P \geq p_H^0 \beta_P \kappa_H & \textit{or} \\ \beta_S \kappa_L \geq m_S, \textit{and } m_P \geq p_H^0 \beta_P \kappa_H + p_L^0 \beta_P \kappa_L. \end{cases} \\
&(P_1\ P_2\ S)^T,\ \textit{if } \beta_S \kappa_H \geq m_S \geq \beta_S \kappa_L \textit{ and } m_P \leq p_H^0 \beta_P \kappa_H \\
&(P_1\ S\ S)^T,\ \textit{if } \beta_S \kappa_L \geq m_S \textit{ and } \beta_P \kappa_L \leq m_P \leq p_H^0 \beta_P \kappa_H + p_L^0 \beta_P \kappa_L \\
&(S\ S\ S)^T,\ \textit{if } \beta_S \kappa_L \geq m_S \textit{ and } m_P \leq \beta_P \kappa_L.
\end{aligned}
\tag{3.34}
$$

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.

□

3.A.6

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 3.8). 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.

$$EW^{FB}[(P_1\ P_2\ S)^T] \geq EW^{FB}[(S\ S\ S)^T] \iff \kappa_L \leq qd_S + c_S - c_P, \tag{3.35}$$

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

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

$$\begin{aligned} &EW^{FB}[(P_1\ P_1\ P_1)^T] \geq EW^{FB}[(S\ S\ S)^T] \iff (1-p)\kappa_L + p\kappa_H \leq \kappa^2 \text{ and} \\ &EW^{FB}[(P_1\ P_1\ P_1)^T] \geq EW^{FB}[(P_1\ P_2\ S)^T] \iff \kappa_L \leq \kappa^3. \end{aligned} \tag{3.36}$$

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.

□

3.A.7

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 3.9).

$$\begin{aligned}\frac{m_P}{\beta_P} &= p_H^0\kappa_H, \\ \frac{m_P}{\beta_P} &= p_H^0\kappa_H + p_L^0\kappa_L, \text{ and} \\ \frac{m_P}{\beta_P} &= \kappa_L\end{aligned} \tag{3.37}$$

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

$$\iff EW^{FB}[(P_1\ P_1\ P_1)^T] \geq EW^{FB}[(P_1\ P_2\ S)^T] \iff \kappa_H \leq \kappa^3. \tag{3.38}$$

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

$$EW^{FB}[(P_1\ P_1\ P_1)^T] = EW^{FB}[(S\ S\ S)^T] \iff \kappa_L = (\kappa^2 - p\kappa_H)/(1-p). \tag{3.39}$$

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 3.9), 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:

$$EW[(P_1\ P_1\ P_1)^T] \geq EW[(S\ S\ S)^T] \iff \kappa_L \leq (\kappa^2 - p\kappa_H - m_P)/(1-p). \tag{3.40}$$

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 $(S\ S\ S)^T$:

$$EW[(P_1\ S\ S)^T] \geq EW[(S\ S\ S)^T] \iff \kappa_L \leq \kappa^2 - m_P. \tag{3.41}$$

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

□

3.A.8

Proof. PCP:

$$\begin{aligned} \gamma_P^{P_1} - \gamma_P^{P_2} + 2\beta_P w \overset{!}{\geq} 0 \\ p_L^0(\gamma_P^{P_2} - \beta_P w) + p_H^0(\gamma_P^S + \beta_P \kappa_H + c_P) - \gamma_P^{P_1} - \beta_P w \overset{!}{\geq} 0 \end{aligned} \tag{3.42}$$

Inserting the cost-based FFS contract yields

$$\begin{aligned} \beta_P w \overset{!}{\geq} 0 \\ \beta_P[p_H^0 \kappa_H - (1 + p_l^0)w] \overset{!}{\geq} 0. \end{aligned} \tag{3.43}$$

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:

$$\begin{aligned} \beta_S(\kappa_H + w) - c_S + \gamma_S^S - \gamma_S^{P_2} \overset{!}{\geq} 0 \\ \beta_S(\kappa_L + w) - c_S + \gamma_S^S - \gamma_S^{P_2} \overset{!}{\leq} 0 \end{aligned} \tag{3.44}$$

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

$$\gamma_S^S - \gamma_S^{P2} = c_S - \beta_S(\kappa_L + w). \tag{3.45}$$

For $\kappa_L + w \leq 0$:

$$\gamma_S^S - \gamma_S^{P2} = c_S. \tag{3.46}$$

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 $\kappa_L \leq c_S - c_P - w$. □

3.A.9

Proof. PCP:

$$\begin{aligned}(I):&\gamma_P^{P_1}-\gamma_P^S-c_P-\beta_P(\kappa_L-w)\overset{!}{\geq}0\\(II):&\gamma_P^{P_1}-\gamma_P^S-c_P-\beta_P[p_L^0(\kappa_L-w)+p_H^0(\kappa_H-w)]\overset{!}{\leq}0\end{aligned} \tag{3.47}$$

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

$$d_S+c_S-c_P=p_L^0(\kappa_L-w)+p_H^0(\kappa_H-w). \tag{3.48}$$

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

Specialist:

$$\gamma_S^S-\gamma_S^{P_2}+\beta_S(\kappa_L+w)-c_S\overset{!}{\geq}0 \tag{3.49}$$

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. □

3.A.10

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

$$\begin{aligned}(I):&\gamma_S^{P_2}-\gamma_S^S\geq\beta_S\kappa_L-c_S\\(II):&\gamma_S^{P_2}-\gamma_S^S\leq\beta_S\kappa_H-c_S\end{aligned} \tag{3.50}$$

Fulfilling (I) exactly, implies (II). Setting the rent equal to 0 delivers the result for the specialist for $\kappa_L\geq0$. This contract fulfills the non-negativity constraints since $\kappa_L\leq\kappa^2=d_S+c_S-c_P$. For $\kappa_L\leq0$ 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. □

3.A.11

Theorem 12. *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*

$$\begin{aligned} \gamma_P^{P_1*} &= \hat{p}_H^0(\beta_P \kappa_H + \gamma_P^{S*} + c_P) \\ \gamma_P^{P_2*} &= 0 \\ \gamma_P^{S*} &= \frac{0-(1-p\hat{q})[\hat{p}_H^0 \beta_P \kappa_H - \hat{p}_L^0 c_P]}{p} \\ \gamma_S^{P_2*} &= d_S + p\beta_S \kappa_L \\ \gamma_S^{S*} &= d_S + c_S - (1-p)\beta_S \kappa_L \end{aligned} \tag{3.51}$$

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*

$$\begin{aligned} \gamma_P^{P_1*} &= \hat{p}_H^0(\beta_P \kappa_H + c_P) \\ \gamma_P^{P_2*} &= 0 \\ \gamma_P^{S*} &= 0 \\ \gamma_S^{P_2*} &= d_S + p\beta_S \kappa_L \\ \gamma_S^{S*} &= d_S + c_S - (1-p)\beta_S \kappa_L \end{aligned} \tag{3.52}$$

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

$$\begin{aligned} \max(-p\hat{q}(\beta_S \kappa_L + d_S), 0) + \max((1-p\hat{q})\beta_P[\hat{p}_L^0 \kappa_L + \hat{p}_H^0 \kappa_H], 0) \geq \\ \max((1-p\hat{q})[\beta_P \hat{p}_H^0 \kappa_H - \hat{p}_L^0 c_P], 0), \end{aligned} \tag{3.53}$$

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*

$$\begin{aligned} \gamma_P^{P_1*} &= \beta_P(\hat{p}_L^0\kappa_L + \hat{p}_H^0\kappa_H) + c_P \\ \gamma_P^{P_2*} &= 0 \\ \gamma_P^{S*} &= 0 \\ \gamma_S^{P_2*} &= 0 \\ \gamma_S^{S*} &= c_S - \beta_S\kappa_L. \end{aligned} \tag{3.54}$$

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 3.A.10). The ICs that need to be fulfilled for P are

$$\begin{aligned} (I): \hat{p}_L^0(\gamma_P^{P_1} - \gamma_P^{P_2}) + \hat{p}_H^0(\gamma_P^{P_1} - \gamma_P^S - \beta_P\kappa_H - c_P) &\geq 0 \\ (II): \gamma_P^S - \gamma_P^{P_1} &\geq -\beta_P\kappa_H - c_P \end{aligned} \tag{3.55}$$

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

$$(I^*): \gamma_P^{P_1} \geq \hat{p}_H^0(\beta_P\kappa_H + \gamma_P^S + c_P) \tag{3.56}$$

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

$$\mathbf{E}\Pi_P = (1-p\hat{q})(\hat{p}_H^0\beta_P\kappa_H - \hat{p}_L^0 c_P) \overset{!}{\geq} 0 \tag{3.57}$$

and the rent-minimizing contract is

$$\begin{aligned} \gamma_P^{P_1*} &= \hat{p}_H^0(\beta_P\kappa_H + c_P) \\ \gamma_P^{P_2*} &= 0 \\ \gamma_P^{S*} &= 0 \\ \gamma_S^{P_2*} &= d_S + p\beta_S\kappa_L \\ \gamma_S^{S*} &= d_S + c_S - (1-p)\beta_S\kappa_L \end{aligned} \tag{3.58}$$

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

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

$$\begin{aligned} \mathbf{E}\Pi_P = (1-p\hat{q})[\hat{p}_H^0(\beta_P\kappa_H + \gamma_P^S + c_P) - c_P] + p\hat{q}\gamma_P^S \stackrel{!}{=} 0 \iff \\ \gamma_P^S = \frac{-(1-p\hat{q})[\hat{p}_H^0\beta_P\kappa_H - \hat{p}_L^0 c_P]}{p}. \end{aligned} \tag{3.59}$$

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

$$\begin{aligned} (I) : \gamma_P^{P_1} - \gamma_P^S \geq \beta_P(\hat{p}_L^0\kappa_L + \hat{p}_H^0\kappa_H) + c_P \\ (II) : \gamma_P^S - \gamma_P^{P_1} \geq -\beta_P\kappa_H - c_P \end{aligned} \tag{3.60}$$

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

$$\begin{aligned} \iff \max(-p\hat{q}(\beta_S\kappa_L - d_S), 0) + \max((1-p\hat{q})\beta_P[\hat{p}_L^0\kappa_L + \hat{p}_H^0\kappa_H], 0) \leq \\ \max((1-p\hat{q})[\beta_P\hat{p}_H^0\kappa_H - \hat{p}_L^0 c_P], 0), \end{aligned} \tag{3.61}$$

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

3.A.12

Theorem 13. *Consider the alternative information structure. Consider the contract*

$$\begin{aligned} \gamma_P^{P_1*} &= c_P \\ \gamma_P^{P_2*} &= c_P \\ \gamma_P^{S*} &= 0 \\ \gamma_S^{P_2*} &= d_S + pm_S \\ \gamma_S^{S*} &= d_S + c_S - (1-p)m_S \end{aligned} \tag{3.62}$$

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

$$\begin{cases} \begin{pmatrix} S & S & S \end{pmatrix}^T & \text{for } \kappa_L > 0 \\ \begin{pmatrix} P_2 & S & S \end{pmatrix}^T & \text{for } \kappa_L \leq 0. \end{cases} \tag{3.63}$$

Under contract (3.62) it is

$$\begin{cases} \begin{pmatrix} S & S & S \end{pmatrix}^T & \text{for } \kappa_L > \frac{m_S}{\beta_S} \\ \begin{pmatrix} P_2 & S & S \end{pmatrix}^T & \text{for } \kappa_L \leq \frac{m_S}{\beta_S}. \end{cases} \tag{3.64}$$

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

$$\kappa_L \leq c_S - c_P. \tag{3.65}$$

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

4 Should Physicians Team Up to Treat Chronic Diseases?

Malte Griebenow

Abstract

This paper studies referral strategy and effort provision of a primary care physician (PCP) and a specialist who are responsible for the treatment of chronically ill patients. Two organizational settings are compared, a team in which physicians cooperate and solo practices in which they do not. Optimal treatment fees are derived for each setting. If the difference in expected treatment costs between disease severities is relatively larger for the PCP, an efficient flow of patients can be achieved in the physician team. In this case, effort is incentivized by a markup on PCP treatment and below-cost fees for specialist treatment. However, if this condition is not fulfilled, care may be delivered second-best efficiently in solo practices with a gatekeeping PCP.

4.1 Introduction

Chronic diseases are a costly burden on health care systems. In the US, for example, 60% of the population suffer from at least one chronic disease. Further, chronically ill patients account for 90% of health care expenditures (Buttorff, 2017). According to Bodenheimer, Chen, et al. (2009), one barrier to the efficient provision of chronic care is that health care systems are built around the treatment of acute problems rather than the long-term health of the patient. One of the measures they propose to improve chronic care is to provide it in teams rather than solo practices. In a meta analysis, Pascucci et al. (2020) find that inter-professional collaboration improves a number of health related outcomes for chronically ill patients indeed. Lemieux-Charles and W. L. McGuire (2006) analyze studies relating to the effectiveness of health care teams compared to usual care. They also find improvements in patient treatment for some interventions. However, they also find that team care may increase costs. This begs the question whether team care is more cost-efficient than care delivered in solo practices. This paper's aim is to provide theoretical guidance regarding this question.

In this paper, I consider two important aspects of chronic care. Firstly, chronically ill patients should receive treatment from an appropriate physician. Which physician should treat a patient crucially depends on the patient's disease severity. Whereas a primary care physician (PCP) is able to cost-efficiently treat a patient in mild condition, a specialist's services are required for a more severe case. This aspect is especially important for chronic diseases as the disease severity of chronically ill patients may change over time. Secondly, treatment efforts exerted today impact health outcomes and costs in the future. The preventive effort of a PCP can decrease the need for future treatment and, thus, decrease costs for the health care system (Dusheiko, Gravelle, Martin, et al., 2011; Bruin et al., 2001; R. Li et al., 2010). Similarly, high-quality specialist treatment can lead to the quicker recovery of the patient.

A key agency issue considered in this paper is physician self-interest. Even though physicians are expected to care about their patients' welfare (Ellis and T. G. McGuire, 1986; Arrow, 1963), empirical evidence indicates that financial incentives strongly affect physicians' provision behavior (Einav et al., 2018; Clemens and Gottlieb, 2014; Epstein and Johnson, 2012). According to some authors (Relman, 2007; Jones, 2002), medical professionals (Perry, 2009;

Whynes et al., 1999), and medical students (Civaner et al., 2016) the medical care market is increasingly commercialized and, as a result, medical professionalism and altruism are in decline. Indeed, J. Li et al. (2017) find, for a sample of US medical students, that the students are substantially less altruistic than the average American. Thus, in the main part of the paper I consider profit-oriented physicians who do not internalize patient benefit[1]. A profit-oriented physician may provide too little effort because she suffers the costs of effort provision but does not internalize the future health losses of patients. Furthermore, she ignores patient health losses when considering whether to refer a patient[2]. As an extension I consider partially altruistic physicians.

If physicians work in solo practices, they do not consider the other physician's profit in their treatment and effort decisions. This can be problematic if patients who could be treated less expensively by the other physician are not referred. If there is no asymmetric information between the physicians, this *coordination problem* can be solved by delivering chronic care in health care teams that are reimbursed by (risk-adjusted) per patient payments as Bodenheimer, Chen, et al. (2009) propose[3]. However, solving the coordination problem is not necessarily socially efficient because the least expensive treatment is not necessarily cost-efficient. Furthermore, organizing physicians in a team provides them an opportunity to collude in order to earn larger profits. For example, if specialists are paid larger treatment fees than PCPs, there is an incentive for the PCP to over-refer patients to the specialist. However, assuming that kickback payments between physicians are not allowed, the PCP faces no such incentive in the solo practices. The trade-off between coordination and collusion is the main focus of this paper[4].

[1]For simplicity, I refer to the physicians with female pronouns and the payer and patient with male pronouns.

[2]Empirical evidence suggests that financial incentives indeed influence physicians' patient selection (Sarma et al., 2018; Iversen and Lurås, 2000), which may result in both over- and under-referrals (Mehrotra et al., 2011).

[3]For example, US accountable care organizations (ACOs) use shared budgetary responsibility between physicians in order to make them financially responsible for their own referrals (Song et al., 2014).

[4]Interestingly, the way physicians are organized varies drastically from country to country. For example, in Germany 60% of ambulatory care physicians work in solo practices (Blümel and Busse, 2015). In the US only 18% of physicians work in solo practices (Kash and Tan, 2016).

The aim of this paper is to answer the following question. Under which conditions should a chronically ill patient receive care from a physician team (PCP and specialist) or from independent physicians who work in solo practices? In order to answer this question I derive optimal treatment fees for both physicians in each setting and compare the second-best optimal outcomes between both organizational forms. As an extension I consider the restriction that the team is paid by a flat treatment fee which does not differ between physicians. The main difference between organizational forms is that in a team, physicians coordinate their referral and effort decisions, whereas in solo practices they do not.

To answer the research question I develop a model with a PCP and a specialist who treat a fixed number of chronic patients for an indefinite time frame. Patients can either be in a mild condition, which is inexpensive to treat, or in a severe condition, which is costly to treat. The severe condition could, for example, correspond to a hyperglycemic diabetes patient or a patient with neuropathic or retinopathic complications. Physicians can exert tertiary preventive effort (time spent on patient, self-help support, appropriate medication, support personnel...) in order to lower the probability that a patient's condition deteriorates and they can exert curative effort in order to increase the probability that a patient's condition improves. Further, physicians refer patients between each other and can accept or reject each other's referrals.

The main innovation of the paper is to analyze physicians' agency problems relating to effort and referral efficiency in a model that captures the dynamic nature of the chronic care market. Patients' severity in each period is determined by a competitive Markov decision process (see Filar and Vrieze, 1996, for the theoretical background for this type of game). A patient's probability distribution over the disease severity in the next period is determined by the patient's current severity, the type of physician treating him, and the exerted effort of the treating physician. Consequently, treatment decisions made by physicians in one period affect the expected costs of care and expected patient health losses in all periods to follow.

Both the severity of the patient (hidden information) and the effort exerted by the physicians (hidden action) are unknown to the payer contracting with the physicians. In order to achieve efficient outcomes, both physicians should be incentivized to exert effort. Further, patients in the mild condition should receive care from the PCP and patients in the severe condition should receive specialist care. I derive conditions under which physicians in each organizational form exert more effort and/or more adequately refer patients.

There are several advantages of using a dynamic model rather than a static model. First, it provides an explanation for the provision of non-contractible effort without using altruism or pay-for-performance mechanisms that require the payer to have information on outcomes ex-post. Instead, physicians provide effort in order to reduce their own (or their team's) future costs of care. Second, a dynamic model allows for the study of the complete treatment path of the patient. This includes a back-referral to the PCP after successful specialist treatment.

I find that if profit maximizing physicians work in solo practices it is not possible for the payer to implement optimal referral patterns, though it may be possible in the team. If the first-best is implementable in the team, markups should be used for PCP treatment, whereas the specialist should be paid below-cost. Nevertheless, it can be optimal to organize the physicians in solo practices. The reason for this is that only in the solo practices a potentially second-best outcome can be implemented, in which the PCP acts as a gatekeeper for the specialist.

Hey and Patel (1983) are the first to develop a Markov model in order to analyze prevention and cure investments of an individual. Hennessy (2008) extends the analysis of Hey and Patel and finds that prevention efforts and cure efforts can be both complements and substitutes. In particular, a subsidy for curative effort may increase the prevalence of an adverse health state as individuals exert less prevention effort. A limitation of these papers is that they do not consider physician agency issues relating to the provision of chronic care.

This paper is also related to the literature on organizational design (Jelovac and Macho-Stadler, 2002; Grassi and Ma, 2016). Grassi and Ma (2016) study a referral market between experts who each provide cost-efficient treatment for one type of client. They find that forming an organization is beneficial for referral efficiency. However, this reduces incentives for cost-control. In their model cost is the only factor that determines which expert should optimally serve a client. By contrast, in my model the physicians not only differ in their costs but also in their ability to treat patient types. Consequently, cost-minimizing treatment may not be socially efficient. Furthermore, they focus on information asymmetry between physicians, which is not a factor in my model. Jelovac and Macho-Stadler (2002) finds that delegating a hospital to contract with its physician may be superior to contracting with both parties simultaneously. This is the case if the hospital's investment decisions matter sufficiently much for the quality of care. Instead of considering subcontracting between hospital and physician, I consider cooperation between two physicians.

Garcia-Mariñoso and Jelovac (2003), Shumsky and Pinker (2003), and Allard, Jelovac, et al. (2011) study referrals in health care markets. The focus of these papers is on setting up efficient payment mechanisms to incentivize appropriate referrals from a PCP to a specialist in a static context. Informational requirements of the optimal contracts can be large. For example, information on the ex-post benefit of a treatment needs to be available in some cases in Garcia-Mariñoso and Jelovac (2003) and Allard, Jelovac, et al. (2011). In contrast, I consider simple treatment fees that may not always be able to incentivize first-best solutions but are easy to implement by the payer. Furthermore, I consider a dynamic game with the possibility of back-referrals from specialists to PCPs.

Malcomson (2005) studies the optimal treatment choice of a physician that can provide one of two treatments to patients that differ in their disease severity and treatment costs. The payer is unable to verify the severity of the patients, though he is able to verify which treatment has been provided. It is only possible to implement the optimal treatment choice of the physician if the costs of the treatment that is appropriate for less severely ill patients rise more strongly in patient severity than the costs of the treatment that is appropriate for the more severely ill patients. The model setup in this paper is similar to Malcomson's model adapted to a dynamic framework with two physicians, in which each physician offers one of the treatment options.

The remainder of the paper is structured as follows. Section 5.2 describes the model used in this paper. Section 4.3 defines the first-best benchmark. In Section 4.4 treatment fees for both the team and the solo practice are derived to implement potentially second-best optimal outcomes under the assumption that the payer can not verify effort provision or the type of the patient. Subsequently, the second-best optimal outcomes for team and solo practice cases are compared. Conditions are derived under which either organizational form is superior. In Section 4.5 the case that teams are paid with flat fees and the case that physicians are partially altruistic are considered as extensions. Section 4.6 concludes.

4.2 Model

The payer contracts with two physicians, a PCP (P) and a specialist (S), to treat a fixed number of chronic patients for an indefinite number of periods. At the end of each period, there is a probability of $0 < \beta < 1$ that the game will continue. Furthermore, there exists an outside provider (in the following: hospital) (H) who treats all patients who are not treated by either physician. I assume that the disease is an ambulatory care sensitive condition. Because hospital treatment tends to be expensive, I assume that the payer always prefers ambulatory treatment to hospital treatment. To save on notation, I do not model the output of the hospital in detail.

In each period, a patient receives treatment from one physician only. There are two types of patients, patients in mild condition (l-types) and patients in severe condition (h-types). For every period that a patient remains in severe condition, he suffers a health loss $L \geq 0$. The payer can not observe the patients' types.

The health state of each patient changes probabilistically over time. The base transition probabilities between state i and j when receiving treatment from physician k are denoted by $p_k^{ij} \in (0, 1)$. Physicians $k \in \{P, S\}$ can exert effort $e_k^i \in \{0, 1\}, i \in \{l, h\}$ in order to reduce the probability that a patient's condition deteriorates from condition l to h (tertiary prevention effort) or increase the probability that a patient's condition improves from condition h to l by Δ_p (treatment effort) in a given period:

$$p_k^{ll}(e_k^l = 1) = p_k^{ll} + \Delta_p, \qquad p_k^{lh}(e_k^l = 1) = p_k^{lh} - \Delta_p$$
$$p_k^{hh}(e_k^h = 1) = p_k^{hh} - \Delta_p, \qquad p_k^{hl}(e_k^h = 1) = p_k^{hl} + \Delta_p$$

In order to ensure that all probabilities stay within zero and one, I assume $\Delta_p < p_k^{ij}$ $\forall i, j \in \{l, h\}, k \in \{P, S\}$. Physician treatment comes at costs $c_k^i + c_e$, where $c_k^i \geq 0$ are the treatment costs for physician k to treat a patient in condition i and $c_e \geq 0$ is the cost of exerting effort[5]. Thus, both the costs and benefits of providing effort are identical between the physicians. However, both the treatment costs and base transition probabilities differ between them.

I assume that for mildly ill patients PCP treatment is cheaper than specialist treatment ($c_P^l < c_S^l$). This is often the case due to the more intensive specialist treatment. For severely ill patients either provider can be cheaper. Furthermore, I assume that severely ill patients are more expensive to treat than mildly ill patients ($\Delta c_k := c_k^h - c_k^l > 0$) because patients in severe condition tend to require more intensive treatment and are at greater risk for complications[6]

Specialists are more effective at treating severely ill patients. This means that with equal physician effort a patient in severe condition will be more likely to improve his condition if he is treated by the specialist rather than the PCP ($p_S^{hh} < p_P^{hh}$). For mildly ill patients, either provider may be more effective. However, if the specialist has a treatment advantage for mildly ill patients, it is smaller than the treatment advantage for severely ill patients ($p_P^{hh} - p_S^{hh} > p_S^{ll} - p_P^{ll}$).

Both physicians can observe each patient's type at any time. At the beginning of a period the physician who treated a patient in the last period is responsible for treating the patient in this period. Physicians decide on whether to treat or refer patients who they are currently responsible for. If the physician decides to refer the patient, the referral only completes if the receiving physician accepts it. In this case, responsibility for the patient's treatment shifts towards the receiving physician. Otherwise, if the referral is rejected, the patient receives treatment from the hospital from then on forward. Figure 4.2.1 shows the possible states a patient can be in and the transition probabilities between the states. PCP behavior is at the top and specialist behavior is at the bottom.

Physicians receive a treatment fee $\gamma_k \geq 0$ in each period per patient whom they treat in

[5]A part of the physicians costs are time costs. The physician's valuation of his time may depend on his opportunity costs. For example, a physician who treats both privately and publicly insured patients may have higher opportunity costs than a physician who only treats publicly insured patients.

[6]This is a common assumption in health economic models (Grassi and Ma, 2016; Hafsteinsdottir and Siciliani, 2010; Malcomson, 2005; Eggleston, 2000; Ellis, 1998).

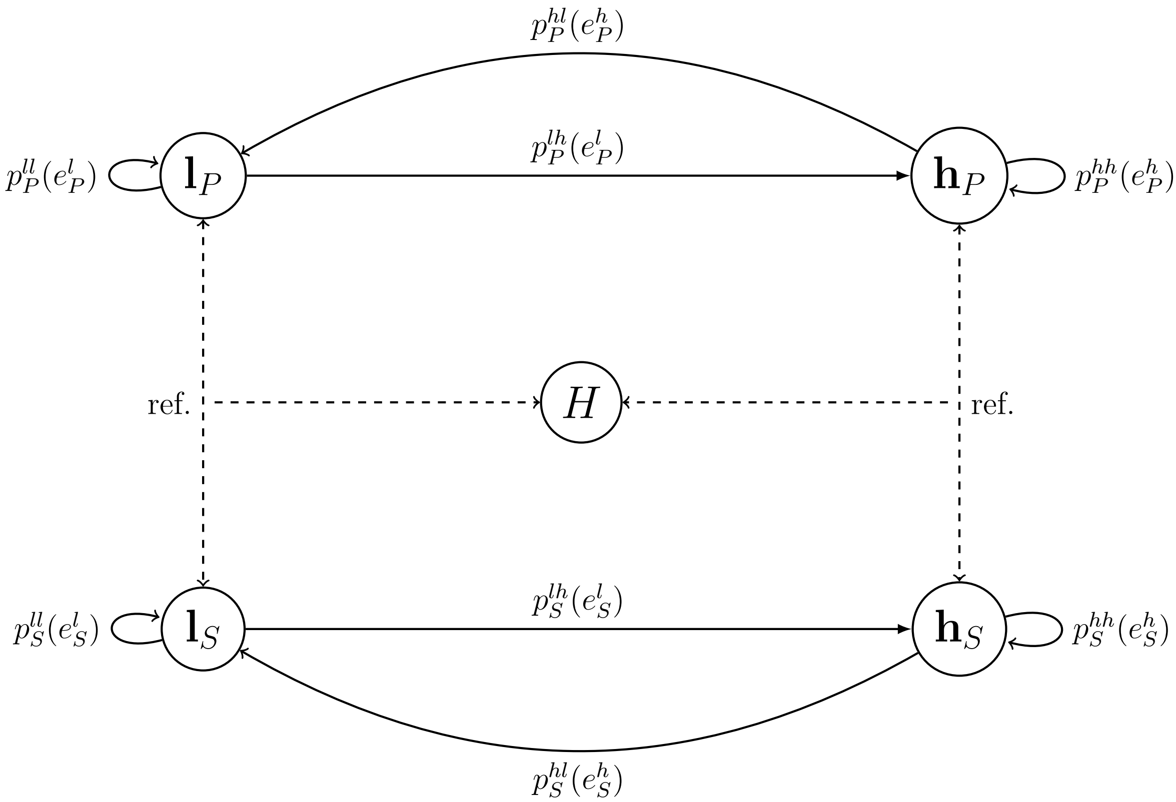


Figure 4.2.1: Patient states and transition probabilities

addition to a fixed payment of F_k which they receive after signing the contract. Periodical treatment fees are common components of real world payment systems[7]. The fixed payment represents subsidies that physicians may receive upon opening a practice.

Physicians are assumed to be profit maximizers and to not be capacity constrained. In order for them to accept their contracts, they need to earn at least a minimum expected discounted profit with the common discount factor β[8]. For simplicity, I assume that any rent can be extracted by lowering the fixed payments. Patients are fully insured and follow their physicians' recommendations. Because the costs, benefits, and treatment fees per patient are independent of the number of treated patients, I analyze a single representative patient. Figure 4.2.2

[7]Capitations and service fees can be budgeted for a certain time frame. For example, in the German social health insurance system, physicians receive strongly reduced fees once they exceed a certain amount of services in a quarter.

[8]Physicians discount the possibility that the game ends. For simplicity, I abstract from other reasons to discount the future.

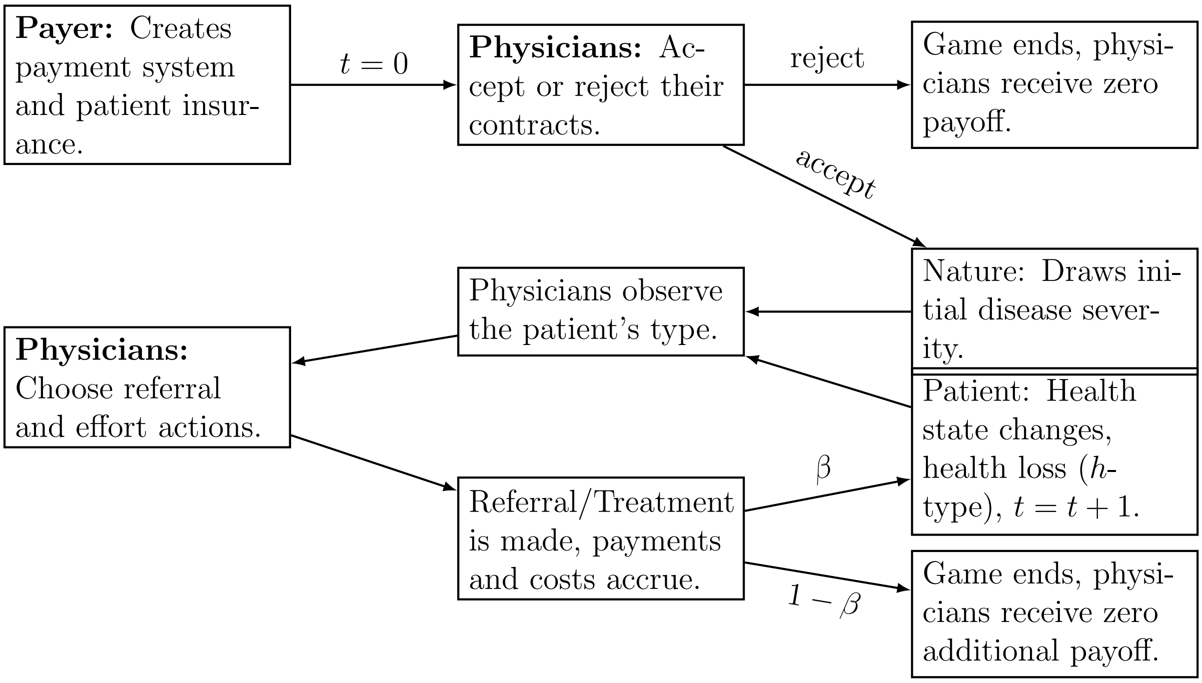


Figure 4.2.2: Sequence of Events

depicts the sequence of events. Before the treatments start, the payment system is created and physicians accept or reject their contracts and receive their fixed payments if they accepted the contract. If either physician rejects, the game ends, physicians receive no payoff, and patients are treated in the hospital. After accepting the contracts, the payer decides which provider is initially responsible for the treatment of the patient. Then the initial disease severity of the patient is drawn by nature.

The following describes the periodical patient treatment. First, the physicians observe the patient's type. The second step depends on the organizational form of the health care market.

Solo Practices: If the physicians work in solo practices, the physicians simultaneously decide non-cooperatively on patient type-dependent referral and effort actions in this period. If a physician chooses to refer the patient, the referral will only be completed if the other physician accepts it. If the receiving physician rejects the referral, the patient is treated in the hospital.

Team: If the physicians form a team, they commit to a cooperative strategy in the second step instead. I assume that this strategy maximizes the discounted expected joint profit of the team. Further, physicians can freely transfer money between each other, provided that both physicians' participation constraints are fulfilled.

After referrals have been made, the patient is treated, costs accrue, and payments are made. With probability β the game continues, with probability $1 - \beta$ it ends. If the game ends, the physicians receive no additional payoff. If the game continues, the state of the patient changes and the patient suffers a health loss if he is in severe condition.

4.3 First-Best Benchmark

In this section I describe the optimal outcome under the condition that the payer can implement any outcome without leaving profits to the physicians. Excluding the degenerate state H, there are four patient state variables $x \in \mathcal{X} := \{l_P, h_P, l_S, h_S\}$ in the game depicted in Figure 4.2.1, namely a patient in mild (l-types)/severe (h-types) condition who receives treatment in this period from the PCP/specialist respectively. Note that the patient *type* is the patient's disease severity, whereas the patient *state* also includes information about which physician is responsible for the treatment of the patient.

Both physicians know the state of the game whenever they take an action. I restrict the analysis of the game to Markov strategies, i.e. strategies that are conditioned only on the state of the game x and not on the history of the game in general[9]. Thus, each physician's strategy is fully described by their referral and effort decisions for each of the four states. The payer minimizes the discounted sum of expected health losses and physician payments (including fixed payments). Rents are optimally set to equal zero, thus the payments (γ, F) just cover the physicians' costs.

The game proposed in Section 5.2 fulfills the definition of a discounted stochastic game[10]. Let us consider first the *continuation welfare* $W(s, x)$. It describes the expected welfare generated

[9]Considering non-Markovian equilibrium strategies would give rise to Folk Theorem type strategies. This undermines the ability of the model to produce unique predictions.

[10]I follow the solution method for this type of game in Filar and Vrieze (1996).

by the physicians' strategies $s = (s_P, s_S)$ for a patient in a given state x:

$$W(s,x) = \underbrace{-\mathcal{L}(x) - \mathcal{C}(s,x)}_{\text{per period health/treatment/effort costs}} + \underbrace{\beta \sum_{\tilde{x}\in\mathcal{X}} p(\tilde{x}|s,x)W(s,\tilde{x})}_{\text{discounted expected future welfare}} \quad ,x \in \mathcal{X}. \tag{4.1}$$

Here, a physician strategy $s_k, k \in \{P, S\}$ assigns an action from the space

$$\{\text{treat, refer}\} \times \{\text{effort, no effort}\}$$

to every state $x \in \mathcal{X}$. Note that continuation welfare is independent of the period t. Thus, from the four equations corresponding to the four states, ex-ante expected welfare can be calculated as

$$\mathbf{E}W(s) = l_P^0 W(s,l_P) + l_S^0 W(s,l_S) + h_P^0 W(s,h_P) + h_S^0 W(s,h_S), \tag{4.2}$$

where $l_P^0, l_S^0, h_P^0, h_S^0 \in [0,1)$ are the initial probabilities for the patient to be in a given state. The probability for the patient to be in severe condition is given by $p^0 \in (0,1)$. The payer decides which physician is responsible for the patient initially. Thus, $h_P^0 + h_S^0 = p^0$, $l_P^0 + l_S^0 = 1 - p^0$. I assume that the patient does not have prior information on his type, thus the payer can not improve by letting the patient choose his own initial physician[11].

Let the patient have a fixed disease severity in some period. The type of physician who treated the patient in the last period is irrelevant to the question of which physician should treat the patient in this period and which effort level he should receive. Therefore, the patient should receive the same treatment and effort for a given type. Thus, any potentially optimal physician strategy can be denoted by

$$s_k = (D_k^l, D_k^h)^{e_k^l e_k^h} \text{ with } D_k^l, D_k^h \in \{T, R\}, \tag{4.3}$$

where D_k^i is the decision of physician k to treat or refer the patient of type i and where e_k^i is the effort decision of the physician. The joint treatment and referral decisions of the physicians define the *treatment paths* for all patients. Table 4.3.1 provides an overview over all possible treatment paths. A treatment path is defined by the physician(s) who treat(s) a patient of type l and h.

[11] In all cases discussed in this paper the payer (weakly) prefers the PCP to receive patients initially. All main results are robust as long as some patients visit the PCP initially.

If the PCP chooses to treat a patient type i and the specialist chooses to refer ($D_P^i = T$, $D_S^i = R$), that patient type will be treated by the PCP. This is indicated by the letter P. Conversely, if the specialist chooses to treat a patient type and the PCP chooses to refer ($D_S^i = T$, $D_P^i = R$), the specialist treats this type. This is indicated by the letter S. If both physicians choose to treat a patient type ($D_P^i = D_S^i = T$), whichever physician is responsible for treating the patient continues to treat him. This is indicated by the letter M. If both physicians choose to refer a patient type ($D_P^i = D_S^i = R$), the patient receives hospital treatment. This is indicated by the letter H. Once a patient receives treatment from the hospital, he does not return to the physicians. I assume that hospital costs are large enough such that implementing a treatment path including H is never optimal from the payer's perspective. To summarize,

P: Treatment by the PCP.

S: Treatment by the specialist.

M: Treatment by the physician who treated the patient in the last period.

H: Treatment by the hospital.

Excluding the paths in which the patient receives hospital treatment, nine different treatment paths emerge.

PS: l-types get referred to the PCP, h-types get referred to the specialist,.

PP: All patients get treated by the PCP.

SS: All patients get treated by the specialist,.

PM: l-types get referred to the PCP, h-types stay at their resp. provider.

MS: h-types get referred to the specialist, l-types stay at their resp. provider.

MM: All patients stay at their resp. provider.

MP: h-types get referred to the PCP, l-types stay at their resp. provider.

SM: l-types get referred to the specialist, h-types stay at their resp. provider.

Table 4.3.1: All possible treatment paths.

		Specialist			
		(T,T)	(T,R)	(R,T)	(R,R)
PCP	(T,T)	MM	MP	PM	PP
	(T,R)	MS	MH	PS	PH
	(R,T)	SM	SP	HM	HP
	(R,R)	SS	SH	HS	HH

SP: h-types get referred to the PCP, l-types get referred to the specialist.

The strategies of the physicians determine the *outcome* of the game. Outcomes are defined by the treatment path that the patient takes and the physicians' efforts. They will be denoted by $\{\text{treatment path}\}^{e_P e_S}$ with $e_k = e_k^l = e_k^h$. In principle physicians could provide effort for one patient type but not the other. However, as will be shown in Subsection 4.4.2, this is never a second-best equilibrium outcome. For proofs I use the following extended notation for outcomes whenever necessary: $\{\text{treatment path}\}^{e_P^l e_P^h e_S^l e_S^h}$. If a physician k does not treat a patient type i, her chosen effort level for this patient type does not impact expected welfare. In this case effort is denoted by e_k^i.

Contracting with the physicians is always in the payer's interest. I assume, consistently with chronic care recommendations (Gask, 2005), that specialist treatment is only socially efficient for patients in severe condition. Furthermore, the cost of effort provision is low enough, such that effort provision for both physicians is always first-best optimal. Therefore, high-effort PCP treatment for low types and high-effort specialist treatment for high types is first-best in all periods. This outcome is denoted by PS^{11}. For this outcome to emerge, the PCP needs to play the strategy $s_P = (T,R)^{1e_P^h}$ and the specialist needs to play the strategy $s_S = (R,T)^{e_S^l 1}$.

Formally, the conditions for PS^{11} to be optimal are (for any initial probabilities x^0):

$$\boldsymbol{E}W(PS^{11}) \geq \boldsymbol{E}W(PS^{00}) \iff c_e \leq \frac{\Delta_p \beta(L + c_S^h - c_P^l)}{1 + \beta(1 - p_P^{ll} - p_S^{hh})} \tag{4.4}$$

$$\boldsymbol{E}W(PS^{11}) \geq \boldsymbol{E}W(PP^{1e_S}) \iff c_S^h - c_P^h \leq \frac{(p_P^{hh} - p_S^{hh})\beta(L + c_P^h - c_P^l)}{1 + \beta(1 - p_P^{ll} - p_P^{hh})} \tag{4.5}$$

$$\boldsymbol{E}W(PS^{11}) \geq \boldsymbol{E}W(SS^{e_P 1}) \iff c_S^l - c_P^l \geq \frac{(p_S^{ll} - p_P^{ll})\beta(L + c_S^h - c_P^l)}{1 + \beta(1 - p_P^{ll} - p_S^{hh})} \tag{4.6}$$

Condition (4.4) implies that effort provision is optimal given that the PCP treats l-types and the specialist treats h-types. Intuitively, the costs of effort may not exceed its discounted benefits.

Condition (4.5) implies that it is first-best optimal to let the specialist treat h-types given that the PCP treats l-types. Here, the cost difference between the specialist and PCP treatment of h-types may not be larger than the additional benefits of specialist treatment.

Condition (4.6) implies that it is first-best optimal to let the PCP treat l-types given that the specialist treats h-types. Note that if $p_P^{ll} \geq p_S^{ll}$, Condition (4.6) is always fulfilled. Otherwise, Condition (4.6) delivers a lower bound on L, whereas Condition (4.5) delivers an upper bound on L. In this case Condition (4.7) needs to hold in order for feasible parameters to exist.

$$\frac{c_S^h - c_P^h}{p_P^{hh} - p_S^{hh}} \leq \frac{c_S^l - c_P^l}{p_S^{ll} - p_P^{ll}} \tag{4.7}$$

Intuitively, Condition (4.7) states that potential cost savings from PCP care relative to specialist care may not be too large for h-types relative to l-types.

An outcome is *superior* (denoted by $\geq$) to another outcome in terms of welfare if and only if it delivers at least the same continuation welfare for any state. It is *strictly superior* if it is superior and delivers strictly greater welfare for some state. For the second-best welfare comparison between organizational forms, I assume the strict superiority of PS^{11} over all other strategies. It follows Lemma 2 (proof in Appendix 4.A.1).

Lemma 2. *If Conditions (4.4) to (4.6) hold (strictly), PS^{11} is (strictly) superior to all other outcomes.*

4.4 Second-Best Analysis

4.4.1 Physician Behaviour

From now on I assume that the payer is not able to observe the type of a patient; however, he is still able to observe which physician provides treatment. The *continuation profit* of physician k for state x, given both physicians strategies s, is

$$u_k^x(s) = \underbrace{\Gamma_k(s,x) - \mathcal{C}_k(s,x)}_{\text{per period profit}} + \underbrace{\beta \sum_{\tilde{x}\in\mathcal{X}} p(\tilde{x}|s,x) u_k^{\tilde{x}}(s)}_{\text{discounted expected future profit}} \quad , x \in \mathcal{X} \tag{4.8}$$

with $\Gamma_k(s,x)$ being the physicians' per period treatment fees. The discounted expected profit in period $t = 0$ amounts to

$$U_k(s) = F_k + l_P^0 u_k^{l_P}(s) + l_S^0 u_k^{l_S}(s) + h_P^0 u_k^{h_P}(s) + h_S^0 u_k^{h_S}(s). \tag{4.9}$$

In order to participate, a physician/the physician team needs to earn minimum profits. For simplicity they are normalized to zero. Thus, an equilibrium strategy s^* must satisfy the *ex-ante participation constraints*:

$$U_k(s^*) \geq 0, k \in \{P, S\} \text{ (PC)} \tag{4.10}$$

In the team, participation depends on the sum of the physicians' profits because the team can always ensure that both physicians earn non-negative profits with the help of internal transfers.

$$U_P(s^*) + U_S(s^*) \geq 0 \text{ (PC - T)} \tag{4.11}$$

Condition (4.8) defines a system of linear equations with four equations that can be used to determine the four unknown continuation profits. They are made up of two parts; the per-period profit of the physician and her discounted expected profit over the remaining periods. In the solo practices, physicians non-cooperatively aim to maximize their discounted expected profit (4.9), whereas the physician team aims to maximize the sum over both physicians' profits.

In order to ensure that the hospital does not receive patients, solo practice physicians need to receive non-negative continuation profit for each patient type, whereas in the team it is sufficient that the team's continuation profit is non-negative.

$$u_k^x(s^*) \geq 0, x \in \mathcal{X}, k \in \{P, S\} \tag{4.12}$$

$$u_P^x(s^*) + u_S^x(s^*) \geq 0, x \in \mathcal{X} \tag{4.13}$$

The payer's goal is to determine the treatment fees and a gatekeeping rule (i.e. a rule determining which physician is initially responsible for treating the patient) that maximize expected welfare (4.2) subject to the constraint that physicians maximize their expected profits (4.9), given the participation constraints (4.10) respectively (4.11) and interim participation constraints (4.12) respectively (4.13). I derive the contract that fulfills the above condition with minimal treatment fees. The fixed payments F_k do not impact incentives. Therefore, F_k is set to exactly fulfill the ex-ante participation constraints:

$$F_k = -[l_P^0 u_k^{l_P}(s) + l_S^0 u_k^{l_S}(s) + h_P^0 u_k^{h_P}(s) + h_S^0 u_k^{h_S}(s)]$$

For the solo practices I derive Markov-Perfect-Equilibria (MPE) in pure strategies, i.e. both physician's strategies need to be best responses to each other in each state:

$$u_P^x(s_P^*|s_S^*) \geq u_P^x(s_P|s_S^*) \forall s_P, \forall x \in \mathcal{X},$$

$$u_S^x(s_S^*|s_P^*) \geq u_S^x(s_S|s_P^*) \forall s_S, \forall x \in \mathcal{X}$$

A strategy is *superior* (denoted by $\geq$) to another strategy in terms of utility if and only if it delivers at least the same continuation utility for each state. It is *strictly superior* (denoted by $>$) if it is superior and it delivers strictly greater continuation utility for some state. For the team I derive outcomes that maximize the teams joint profit. For the comparison between organizational forms, I only consider strategies that can be implemented as a unique equilibrium, in order to ensure that they can be reliably implemented by the payer.

As assumed, a patient of a given type has the same costs and transition probabilities if he receives the same treatment, regardless of which physician treated him last period. Thus, a physician can never improve by treating the patient differently based on the previously treating physician. I therefore only consider strategies that assign the same treatment and effort decision to the same patient type (as in Condition (4.3)).

If possible, the payer aims to implement both the first-best treatment path PS as well as effort provision for both physicians $e_P^l = e_S^h = 1$. However, both implementing effort provision and implementing the desired treatment path PS are not always possible. Thus, a second-best outcome needs to be implemented in this case. Lemma 3 is useful in the second-best analysis. It follows from Conditions (4.5) and (4.6), see Appendix 4.A.2.

Lemma 3. *It holds that*

$$\begin{aligned} &\boldsymbol{E}W(PS^{e_Pe_S}) \geq \boldsymbol{E}W(PM^{e_Pe_S}) \geq \boldsymbol{E}W(PP^{e_Pe_S}) \text{ and} \\ &\boldsymbol{E}W(PS^{e_Pe_S}) \geq \boldsymbol{E}W(MS^{e_Pe_S}) \geq \boldsymbol{E}W(SS^{e_Pe_S}) \end{aligned} \tag{4.14}$$

for $e_P = e_S$. If Conditions (4.4) to (4.6) hold strictly, so does Condition (4.14).

Furthermore, paths MM, SP, MP and SM are weakly dominated by PP and SS for fixed e_P, e_S.

Intunitively, Lemma 3 states that given equal levels of effort, an outcome is superior to another if its treatment path deviates less from the first-best treatment path PS.

In Subsection 4.4.2 and 4.4.3 I derive the set of potentially second-best outcomes which can be implemented by the payer for the solo practices and the team respectively. Furthermore, I derive conditions on the treatment fees that allow the payer to implement his desired outcome. In Subsection 4.4.4 I compare the second-best optimal outcomes from both organizational forms and derive conditions under which either one is optimal.

4.4.2 Solo Practices

In this subsection I derive the set of potentially second-best outcomes, i.e. the set of outcomes that are not dominated in terms of welfare and which can be implemented by the payer. Let us consider first which treatment paths can not be implemented in the solo practices. This is shown in and Proposition 1.

Proposition 1. *Strategies $s_k = (R, T), k \in \{P, S\}$ are never a part of an MPE in the solo-practices.*

Table 4.4.1: All possible treatment paths (solo practices), bold: first-best, single crossed-out: not second-best optimal, double crossed-out: impossible to implement as a unique MPE.

		Specialist			
		(T,T)	(T,R)	~~(R,T)~~	(R,R)
PCP	(T,T)	~~MM~~	~~MP~~	~~PM~~	PP
	(T,R)	MS	~~MH~~	~~**PS**~~	~~PH~~
	~~(R,T)~~	~~SM~~	~~SP~~	~~HM~~	~~MP~~
	(R,R)	SS	~~SH~~	~~HS~~	~~HH~~

Proposition 1 is true because l-types are cheaper to treat than h-types. Therefore, treating an l-type now is always more profitable than treating an h-type now or later. Hence, physicians never refer l-types when they are willing to treat h-types. The affected treatment paths have been double crossed out in Table 4.4.1. Not being able to incentivize $s_S = (R, T)$ is problematic because the first-best calls for the specialist to refer back l-types and to treat h-types.

In contrast to the first-best treatment path, the "blind" treatment paths in which only one physician treats all patients (PP and SS), can trivially be implemented by the payer by only paying one physician sufficient treatment fees for treating both patient types. If a physician treats all patients, she is willing to exert effort for both patient types if the costs of effort are lower than the discounted future cost savings from keeping the patient in the mild condition. For the PCP and specialist the conditions are respectively,

$$c_e \leq \tilde{c_e}^P := \frac{\Delta_p \beta \Delta_{c_P}}{1 + \beta(1 - p_P^{ll} - p_P^{hh})},$$
$$c_e \leq \tilde{c_e}^S := \frac{\Delta_p \beta \Delta_{c_S}}{1 + \beta(1 - p_S^{ll} - p_S^{hh})}.$$

Thus, if there is a large difference in the treatment costs between l- and h-types (i.e large Δ_{c_P} or Δ_{c_S}) physicians are more willing to exert effort. For both treatment paths, the exerted effort is independent from the treatment fees. It follows Proposition 2 (proof in Appendix 4.A.3).

Proposition 2. *Treatment paths PP and SS can always be implemented by the payer. Effort* *for both patient types is exerted by the PCP (specialist) if and only if* $c_e \leq \tilde{c_e}^P (\tilde{c_e}^S)$. *These two treatment paths together weakly dominate paths* MM *and* MP *in terms of social welfare.*

As PP and SS can always be implemented, any treatment path that is dominated by them in terms of social welfare, is never second-best optimal. Lemma 3 has already shown that MM and MP are weakly dominated for fixed levels of effort. Whenever effort can be incentivized for MM and MP, it can also be incentivized for PP and SS. Thus, the latter paths weakly dominate the former. MM and MP have been crossed out once in Table 4.4.1. Furthermore, as assumed, hospital treatment is never efficient from the payer's perspective. All remaining paths containing hospital treatment have also been crossed out once.

Proposition 2 shows that despite the fact that physicians are profit oriented in this model, they may still exert non-contractible effort in order to reduce their future treatment costs. Thus, the dynamic model of physician treatment offers an alternative explanation to altruism for effort in a credence goods market. Note that in the blind treatment paths physicians fully internalize the cost savings from high effort provision, though they do not internalize the patient's health losses. Therefore, only under-provision of effort is a potential issue but not over-provision.

It may seem surprising that the threshold for providing effort is independent of patient severity. As Hennessy (2008) point out, prevention effort and curative effort can be both complements and substitutes. If the probability that a patient remains healthy increases, so does the future gain from increasing curative effort. The reason for this is that patients stay healthy for longer once they are cured. If the probability that a patient improves his condition increases, the future gain from increasing prevention effort decreases. This is so because patients return to the healthy state more quickly. In this model, both effects just cancel each other out because both effort costs and the effect of effort coincide between disease severities.

There remains only a single treatment path from Table 4.4.1 which has not been analyzed yet, namely MS. In this path the PCP treats l-type patients until their condition deteriorates to h and then refers them to the specialist (as in the first-best). The specialist then continuously treats the patient (as in the blind specialist treatment). This treatment path differs from the first-best path PS insofar that patients who have been cured by the specialist are not referred back to the PCP. In order to implement this path, the PCP plays strategy $(T, R)^{e^l_P e^h_P}$ and the specialist plays strategy $(T,T)^{e^l_S e^h_S}$. Even though the specialist would prefer to treat l-types, the PCP does not refer them. Thus, the PCP acts as a gatekeeper for the patient. In order to maximally utilize this gatekeeping effect, all patients should initially visit the PCP.

The specialist can be easily incentivized to treat all referred patients by setting her treatment fee higher or equal to her expected costs. Her effort provision is independent of the payments. Effort will be provided for both patient types if and only if $c_e \leq \tilde{c_e}^S$ (see Proposition 2).

Let the specialist play $(T,T)^{e^l_S e^h_S}$. Incentivizing the PCP to play $(T,R)^{e^l_P e^h_P}$ requires that the treatment fee is larger than her costs for l-types but not too large so that her continuation profits for h-types are not positive.

With PCP effort:

$$U_P[(T,R)^{1e^h_P}] \geq U_P[(R,R)^{e^l_P e^h_P}] \iff \gamma_P \geq c^l_P + c_e \tag{4.15}$$

$$U_P[(T,R)^{1e^h_P}] \geq U_P[(T,T)^{11}] \iff \gamma_P \leq c^h_P + c_e - \frac{\beta\Delta_{c_P}(1 - p^{hh}_P + \Delta_p)}{1 + \beta(1 - p^{ll}_P - p^{hh}_P)} \tag{4.16}$$

Without PCP effort:

$$U_P[(T,R)^{0e^h_P}] \geq U_P[(R,R)^{e^l_P e^h_P}] \iff \gamma_P \geq c^l_P \tag{4.17}$$

$$U_P[(T,R)^{0e^h_P}] \geq U_P[(T,T)^{00}] \iff \gamma_P \leq c^h_P - \frac{\beta\Delta_{c_P}(1 - p^{hh}_P)}{1 + \beta(1 - p^{ll}_P - p^{hh}_P)} \tag{4.18}$$

Both pairs of conditions can be fulfilled at the same time respectively by a contract that just covers the PCP's treatment (and in the first case effort) costs for the treatment of l-types.

In order to incentivize effort, the PCP needs to make profits on l-types. This ensures that she has an incentive to keep patients in the mild state.

$$U_P[(T,R)^{1e^h_P}] \geq U_P[(T,R)^{0e^h_P}] \iff \gamma_P \geq c^l_P + \frac{c_e(1 - \beta p^{ll}_P)}{\beta\Delta_p} \tag{4.19}$$

Condition (4.19) shows that in order to incentivize PCP effort, it is not sufficient to only pay her treatment and effort costs. Instead, a markup on these costs needs to be paid. Continuation profits for treating l-types with effort rise more strongly in the treatment fee γ_P than the continuation profits for treating l-types without effort. The reason for this is that effort increases the expected amount of periods that a patient will be treated by the PCP. However, it is not possible to increase γ_P indefinitely because this would incentivize the PCP to not refer h-types (see Condition (4.16)). The larger the cost differences between types are for the PCP (Δ_{c_P}), the more room exists for increasing γ_P and, thus, for incentivizing effort. This effect is only present due to the multi-period nature of the model. In a one-shot interaction, the PCP would not have to consider the impact of her effort decision on future costs, and thus, paying a markup on her costs would be ineffective for incentivizing effort. Proposition 3 (proof in Appendix 4.A.4) describes when and how effort incentivization for MS is possible.

Proposition 3. *In the solo practices, the treatment path in which both physicians treat l-types and the specialist treats h-types (MS) can always be implemented as a (unique) MPE. If and only if $c_e(<) \leq \tilde{c}_e{}^P$, the PCP can be incentivized to provide effort and if and only if $c_e(<) \leq \tilde{c}_e{}^S$, the specialist provides effort.*

In order for the PCP to provide effort, she needs to be paid a markup on both her treatment and effort costs of l-types and the specialist can to be paid her expected treatment and effort costs for the treatment of h-types:

$$\gamma_P^* = c_P^l + \frac{c_e(1-\beta p_P^{ll})}{\beta\Delta_p},$$
$$\gamma_S^* \geq c_S^h + c_e - \frac{\beta\Delta_{c_S}(1-p_S^{hh}+\Delta_p)}{1+\beta(1-p_S^{ll}-p_S^{hh})}.$$

If effort can not be incentivized ($c_e > \tilde{c}_e{}^S, \tilde{c}_e{}^P$):

$$\gamma_P^* = c_P^l,$$
$$\gamma_S^* \geq c_S^h - \frac{\beta\Delta_{c_S}(1-p_S^{hh})}{1+\beta(1-p_S^{ll}-p_S^{hh})}.$$

Let $\epsilon > 0$ be sufficiently small. The equilibria are unique for

$$\gamma_P^* = c_P^l + \frac{c_e(1-\beta p_P^{ll})}{\beta\Delta_p} + \epsilon,$$
$$\gamma_S^* > c_S^h + c_e - \frac{\beta\Delta_{c_S}(1-p_S^{hh}+\Delta_p)}{1+\beta(1-p_S^{ll}-p_S^{hh})},$$

respectively

$$\gamma_P^* = c_P^l + \epsilon,$$
$$\gamma_S^* > c_S^h - \frac{\beta\Delta_{c_S}(1-p_S^{hh})}{1+\beta(1-p_S^{ll}-p_S^{hh})}.$$

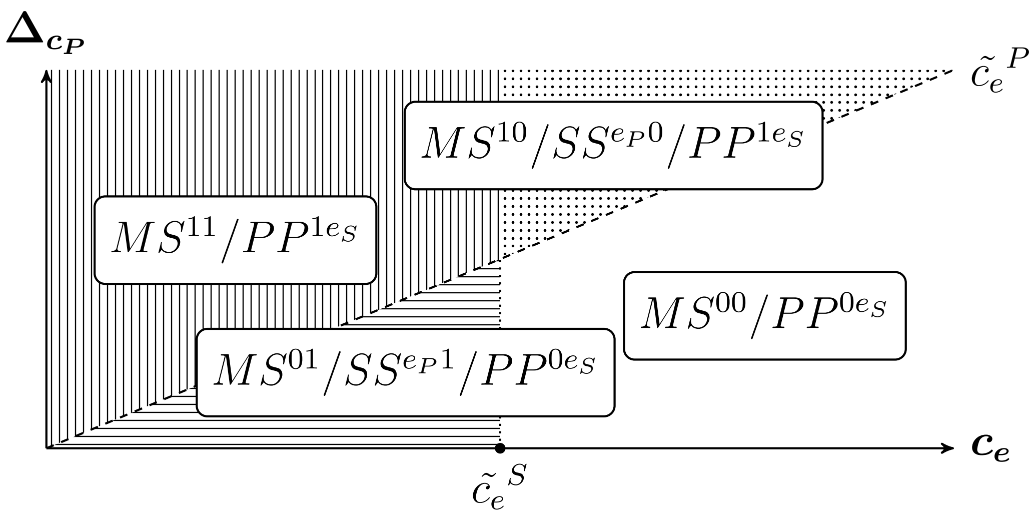


Figure 4.4.1: Potentially second-best outcomes in the solo practices.

Figure 4.4.1 summarizes the results from Propositions 2 and 3. It shows, depending on Δ_{c_k} and c_e, the set of potentially second-best outcomes in the four differently shaded areas. These sets consist of PP, SS, and MS with PCP effort provision for $c_e \leq \tilde{c_e}^P$ and specialist effort provision for $c_e \leq \tilde{c_e}^S$. Note that given $c_e \leq \min(\tilde{c_e}^S, \tilde{c_e}^P)$ or $c_e \geq \max(\tilde{c_e}^S, \tilde{c_e}^P)$, blind specialist treatment is not second-best optimal as MS dominates it (see Lemma 3). In the other two areas either blind treatment or MS can be second-best optimal. Corollary 1 summarizes the results for the solo practices.

Corollary 1. *The first-best treatment path can not be implemented in the solo practices.*

In addition to blind PCP or specialist treatment PP and SS, there is another potentially second-best treatment path MS that can always be implemented. In this path the PCP acts as a gatekeeper for patients in mild condition and only refers severely ill patients to the specialist who continues treatment until the game ends.

4.4.3 Team Care

Physicians who work in a team coordinate their treatment and effort decisions via internal profit-sharing rules. I assume that there is no asymmetric information between the physicians

Table 4.4.2: All possible treatment paths (team), bold: first-best, single crossed-out: not second-best optimal, double crossed-out: impossible to implement as a unique equilibrium.

		Specialist			
		(T,T)	(T,R)	(R,T)	(R,R)
PCP	(T,T)	$\xcancel{MM}$	$\xcancel{MP}$	$\xcancel{PM}$	PP
	(T,R)	$\xcancel{MS}$	$\xcancel{MH}$	$\boldsymbol{PS}$	~~PH~~
	(R,T)	$\xcancel{SM}$	SP	$\xcancel{HM}$	$\xcancel{MP}$
	(R,R)	SS	~~SH~~	~~HS~~	~~HH~~

and that they can freely transfer profits between them. Thus, they maximize their joint profit $U_T = U_P + U_S$ and split it in some manner[12]. For each patient of the same type, the continuation profit for the team (i.e. the sum of the continuation profits of both physicians) is identical. Therefore, the physician team never strictly prefers a mixed treatment path (i.e. a treatment path for which at least one patient type receives treatment from both physicians) to a treatment path in which only one physician treats a certain patient type. Furthermore, according to Condition (4.13), as long as the team makes a non-negative continuation profit from each type, the team will never refer a patient to the hospital. This can be ensured by setting treatment fees sufficiently high. Consequently, the possible treatment paths from Table 4.3.1 can be reduced to those in Table 4.4.2. These paths have in common that for each patient type there is exactly one physician who provides treatment.

This subsection proceeds as follows. First, I derive under which conditions PS^{11} can be implemented. Second, I derive second-best candidates given that PS^{11} can not be implemented. Finally, I derive upper boundaries on L such that PS^{11} is first-best optimal and can be implemented in the team.

Let us consider first the first-best outcome PS^{11}. In order to implement it, after eliminating

[12]For example, they may use Nash Bargaining.

all inactive incentive constraints (see Appendix 4.A.5), three constraints need to be fulfilled

$$PS^{11} \geq SS^{e_P 1} : (\gamma_S - c_S^h) - (\gamma_P - c_P^l) \leq -\frac{[1+\beta(1-p_P^{ll}-p_S^{hh})]\Delta_{c_S}}{1+\beta(1-p_S^{ll}-p_S^{hh})} \tag{4.20}$$

$$PS^{11} \geq PS^{00} : (\gamma_S - c_S^h) - (\gamma_P - c_P^l) \leq -\frac{c_e[1+\beta(1-p_P^{ll}-p_S^{hh})]}{\beta\Delta_p} \tag{4.21}$$

$$PS^{11} \geq PP^{1e_S} : (\gamma_S - c_S^h) - (\gamma_P - c_P^l) \geq -\frac{[1+\beta(1-p_P^{ll}-p_S^{hh})]\Delta_{c_P}}{1+\beta(1-p_P^{ll}-p_P^{hh})} \tag{4.22}$$

Condition (4.20) ensures that l-types are treated by the PCP. This is a lower bound on the team's profits when treating l-types. Condition (4.21) demonstrates that in order for the team to provide effort, it is necessary that the team makes greater profits when the PCP treats l-types rather than when the specialist treats h-types. Earning a markup for PCP treatment relative to specialist treatment increases the continuation profits of the team when treating l-types. Thus, they are incentivized to exert effort in order to increase the expected number of l-types in future periods. Which of these two condition is more strict depends on the effort costs c_e and the difference in treatment costs between types for the specialist (Δ_{c_s}). If $c_e \leq \tilde{c_e}^S$, Condition (4.20) is stricter than (4.21). Otherwise, the opposite is true.

However, markups for PCP treatment may not be too large relative to the specialist's markups. Otherwise, the PCP treats all patients. Condition (4.22) ensures that this does not happen. All three constraints can not always be fulfilled together. Proposition 4 (proof in Appendix 4.A.5) describes how, and under which conditions, the first-best outcome PS^{11} can be implemented in the team. Because Condition (4.22) is the only lower boundary on γ_S, the payer can simply fulfill it exactly in order to implement the outcome whenever this is possible.

Proposition 4. *The first-best treatment path (PS) can be implemented as a (unique) equilibrium in the team if and only if the difference in the expected treatment costs between mild and severe cases is (strictly) larger for the PCP than the specialists:*

$$\Delta_{c_P}(>) \geq \tilde{\Delta}_{c_P} := \frac{\Delta_{c_S}[1+\beta(1-p_P^{ll}-p_P^{hh})]}{1+\beta(1-p_S^{ll}-p_S^{hh})}. \tag{4.23}$$

Then, if and only if effort costs are low enough, i.e.

$$c_e(<) \leq \tilde{c_e}^P, \tag{4.24}$$

the following contract implements PS^{11}:

$$\gamma_P^* = c_P^l + c_e + \frac{\beta \Delta_{c_P}(1 - p_P^{ll} - \Delta_p)}{1 + \beta(1 - p_P^{hh} - p_P^{ll})}$$
$$\gamma_S^* = c_S^h + c_e - \frac{\beta \Delta_{c_P}(1 - p_S^{hh} + \Delta_p)}{1 + \beta(1 - p_P^{hh} - p_P^{ll})}$$

Let $\epsilon > 0$ be sufficiently small. Then, the following contract implements PS^{11} as a unique equilibrium:

$$\gamma_P^* = c_P^l + c_e + \frac{\beta \Delta_{c_P}(1 - p_P^{ll} - \Delta_p)}{1 + \beta(1 - p_P^{hh} - p_P^{ll})} - \epsilon$$
$$\gamma_S^* = c_S^h + c_e - \frac{\beta \Delta_{c_P}(1 - p_S^{hh} + \Delta_p)}{1 + \beta(1 - p_P^{hh} - p_P^{ll})}$$

According to Proposition 4, given that Conditions (4.23) and (4.24) hold, the first-best outcome PS^{11} can be implemented by paying the PCP a markup on her costs and paying below-cost fees to the specialist. This ensures that the team does not shift care towards the more expensive specialist and that the team exerts effort because they benefit from a patient in mild condition. Furthermore, the team's continuation profits for h-types are set to zero.

Condition (4.23) requires that the difference in expected costs when continuously treating an l-type patient rather than an h-type patient is larger for the PCP than the specialist. In this case the PCP has the *(expected) relative cost-advantage* when treating l-type patients, and the specialist has the relative cost-advantage when treating h-type patients. If, conversely, the specialist has the relative cost advantage for treating l-types, the team is only willing to provide specialist care to h-types if the payment for specialist care is large. In this case, however, l-type patients are also treated by the specialist and, thus, the first-best outcome is not implemented. Condition (4.24) requires that effort costs are lower than the expected cost differences between the patient types for the PCP. This is necessary because effort is incentivized by a markup on PCP care. The larger the effort costs are, the larger is the markup required to incentivize effort. However, if the expected cost differences between types are small for the PCP, the team shifts care of h-types towards the PCP rather than exerting effort.

The second-best outcomes in the team are depicted in Figure 4.4.2. If the PCP has the relative cost-advantage for l-types ($\Delta_{c_P} > \tilde{\Delta}_{c_P}$), treatment path PS can always be implemented.

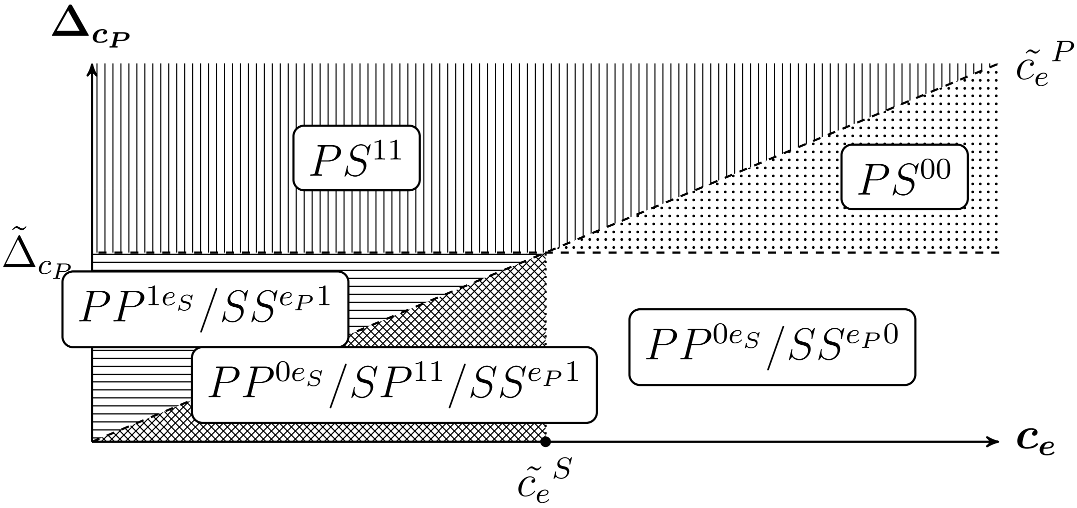


Figure 4.4.2: Second-best implementation in the team if $p_P^{ll} \geq p_S^{ll}$.

For low enough effort costs, the first-best outcome PS^{11} is implemented. If PS^{11} can not be implemented because effort costs are too large, PS^{00} can be implemented instead (see Appendix 4.A.5). The alternative to implementing PS^{00} is to implement one of the blind outcomes without effort. Thus, according to Lemma 3, implementing PS^{00} is always second-best optimal. If the specialist has the relative cost-advantage for l-types ($\Delta_{c_P} < \tilde{\Delta}_{c_P}$), either blind treatment or SP^{11} can be second-best. Details are discussed in Appendix 4.A.6.

So far I have considered whether PS^{11} can be implemented by the payer. However, the first-best can not be implemented whenever this is desirable, i.e. whenever Conditions (4.4) to (4.6) are fulfilled. Here, a large health loss L is problematic. If $L = 0$, the cost-minimizing outcome is first-best optimal. A simple flat fee $\gamma_P = \gamma_S$ implements it in the team. However, with growing L the first-best benchmark is influenced more by concern for the patient's health and less by cost considerations. Thus, the strong focus on cost reduction in the team turns into a disadvantage. Differing markups for PCP and specialist can alleviate but not necessarily solve this problem by partly internalizing patient health losses. In Subsection 4.5.1 I consider the additional problems that emerge when fees have to be flat. Proposition 5 (proof in Appendix 4.A.7) shows a sufficient condition under which PS^{11} can be implemented in the team for all feasible parameters.

Proposition 5. *Let λ^* be the set of parameter combinations satisfying Conditions (4.4) to*

(4.6). The first-best outcome PS^{11} can be implemented in the team for all $\lambda \in \lambda^*$ that satisfy

$$L \leq \min(L^1, L^2) \text{ with}$$
$$L^1 := [1 + \beta(1 - p_P^{ll} - p_P^{hh})] \left(\frac{c_S^h - c_P^h}{\beta(p_P^{hh} - p_S^{hh})} - \frac{\Delta_{c_S}}{1 + \beta(1 - p_S^{ll} - p_S^{hh})} \right),$$
$$L^2 := c_P^l - c_S^h + \frac{\Delta_{c_P}(1 + \beta(1 - p_P^{ll} - p_S^{hh}))}{1 + \beta(1 - p_P^{ll} - p_P^{hh}))}.$$

Conversely, if $L > \min(L^1, L^2)$, there exist $\lambda \in \lambda^*$ such that PS^{11} can not be implemented in the team.

Corollary 2 summarizes the results for the team.

Corollary 2. *Providing chronic care in a team can incentivize appropriate referrals between physicians if the PCP has the relative cost-advantage for treating patients in mild condition.*

If effort costs are low compared to the difference in the PCP's expected cost of care between severely ill and mildly ill patients, effort can be incentivized as well. In this case, markups for the PCP's treatment of mildly ill patients and below-cost fees for the specialist incentivizes the team to play the first-best strategies.

If and only if the health losses from the severe conditions are sufficiently small, the first-best can always be implemented in the team.

4.4.4 Optimal Organization of Care

Comparing the second-best outcomes from solo practice care and team care yields the results depicted in Figure 4.4.3. If the PCP has the strict relative cost-advantage when treating l-types ($\Delta_{c_P} > \tilde{\Delta}_{c_P}$) the first-best treatment path can always be implemented in the team but not in the solo practices. Furthermore, the physicians always provide effort if they provide effort under solo practice care. Thus, team care strictly dominates solo practice care.

Conversely, if the specialist has the strict relative cost-advantage when treating l-types ($\Delta_{c_P} < \tilde{\Delta}_{c_P}$), there is an adverse incentive for the specialist to treat the mildly ill patients instead of,

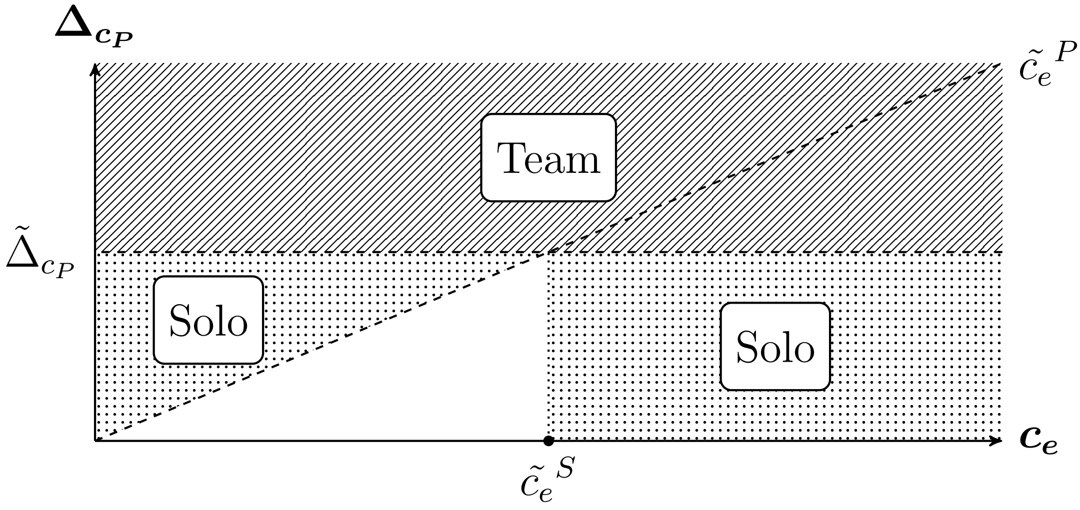


Figure 4.4.3: Second-best optimal organization of care (dots: weak dominance for solo practices, lines: strict dominance for the team, no pattern: no general dominance relation)

or in addition to, the patients in severe condition. In the solo practices this incentive is isolated on the specialist, whereas in the team this incentive also affects the PCP. Thus, only the blind treatment paths PP and SS or, if $\tilde{c_e}^P < c_e < \tilde{c_e}^S$, the reversed first-best path SP will be implemented in the team. Not being able to implement PS^{11} is problematic if this outcome is first-best optimal given $\Delta_{c_P} < \tilde{\Delta}_{c_P}$, which is the case only if the health loss in the severe condition L is sufficiently large ($L > L^1$, see Appendix 4.A.7).

In the solo practices the payer can potentially improve on the team's outcomes. In addition to implementing the blind outcomes it is possible to incentivize the PCP to treat only patients in mild condition and to incentivize the specialist to treat all referred patients (treatment path MS). In the team it is not possible to implement this path. Therefore, if $c_e < \tilde{c_e}^P$ or $c_e > \tilde{c_e}^S$, the set of implementable outcomes is larger in the solo practices than in the team. Thus, solo practice care weakly dominates team care in this case.

Strict superiority of solo practice care can also be demonstrated for some parameters. Let $\Delta_{c_P} < \tilde{\Delta}_{c_P}$. If $c_e < \tilde{c_e}^P$, either MS^{11} or PP^{1e_S} are second-best optimal. If $c_e > \tilde{c_e}^S$, either MS^{00} or PP^{0e_S} are second-best optimal. Whenever MS is strictly optimal, the solo practices are strictly superior to the team. As shown in Appendix 4.A.8, MS is strictly superior to

PP whenever the PCP's treatment costs for treating h-types are sufficiently large. Thus, cooperation between physicians is not desirable for the payer in this case. If $\tilde{c_e}^P < c_e < \tilde{c_e}^S$, no dominance relationship between organizational forms can be established for all parameters because outcome SP^{11} can not be implemented in the solo practices. However, as shown via simulation in Appendix 4.A.8, there also exist parameters in this case such that solo practice care is strictly superior to team care.

To illustrate the case in which solo practice care strictly dominates team care, consider the following example. Let specialist care of high-severity patients be both necessary (L very large, $p_S^{hh} << p_P^{hh}$) and expensive (c_S^h large). Note that in this case, it is likely that the specialist has the relative cost-advantage for l-types. If so, and if additionally the PCP only has a slight advantage when treating l-types ($c_P^l \approx c_S^l, p_S^{ll} \approx p_P^{ll}$), MS will be strictly superior to PP for levels of c_P^h that are sufficiently large but not large enough to shift the relative cost-advantage for l-types to the PCP. Figures 4.A.1 and 4.A.2 in Appendix 4.A.8 showcase this example by means of simulation. Proposition 6 summarizes the main results of this paper.

Proposition 6. *Team care is (strictly) superior to solo practice care (given that Conditions (4.4) to (4.6) hold strictly) if the PCP has the relative cost advantage for treating patients in mild condition, i.e.* $\Delta_{c_P} > \tilde{\Delta}_{c_P}$*. Otherwise, if* $\Delta_{c_P} < \tilde{\Delta}_{c_P}$ *and* $c_e < \tilde{c_e}^P$ *or* $c_e > \tilde{c_e}^S$*, solo practice care is weakly superior to team care. In this case, if the PCP's costs for treating h-types is sufficiently large, solo practice care is strictly superior to team care. A necessary condition for solo practice superiority is that health losses in severe condition are sufficiently large,* $L > L^1$.

4.5 Extensions

4.5.1 Flat Fee in the Team

In this Subsection I consider the effects of paying a flat periodical treatment fee to the physician team, i.e. a fee that is not differentiated by the type of physician who provided treatment during the period ($\gamma_P = \gamma_S$). This is relevant if practices are paid in bundled, periodical payments for the provision of both primary and specialty care. Naturally, we can expect to implement the first-best outcome in a smaller parameter region than before because the optimization problem includes an additional condition. In fact, the size of the flat fee does not impact incentives in the team because the team's behavior is influenced only by the difference in treatment fees between physicians (see Conditions (4.20) to (4.22)). The solo practices are still paid by treatment fees that may differ. Consequently, solo practices perform relatively better against the team than before.

Hospital treatment can be prevented by paying sufficiently large flat fees. After eliminating all non-binding incentive constraints, the following constraints remain in order for PS^{11} to be played:

$$U_T(PS^{11}) \geq U_T(PS^{00}) \iff c_e \leq \frac{\Delta_p \beta(c_S^h - c_P^l)}{1 + \beta(1 - p_P^{ll} - p_S^{hh})} \tag{4.25}$$

$$U_T(PS^{11}) \geq U_T(PP^{1e_S}) \iff c_S^h - c_P^h \leq \frac{(p_P^{hh} - p_S^{hh})\beta(c_P^h - c_P^l)}{1 + \beta(1 - p_P^{ll} - p_P^{hh})} \tag{4.26}$$

$$U_T(PS^{11}) \geq U_T(SS^{e_P 1}) \iff c_S^l - c_P^l \geq \frac{(p_S^{ll} - p_P^{ll})\beta(c_S^h - c_P^l)}{1 + \beta(1 - p_P^{ll} - p_S^{hh})} \tag{4.27}$$

Note that these conditions are identical to the first-best Conditions (4.4) to (4.6) given that $L = 0$. Thus, the team minimizes expected treatment costs without internalizing the patient's health loss. Inefficient treatment is thus provided if and only if providing PS^{11} is not cost-minimizing but still socially efficient due to patient health losses. PS^{00} is played if Condition (4.25) holds with flipped inequality. Providing effort to one type but not the other is always weakly dominated by providing effort for both types or not providing effort for either type.

Condition (4.27) states that the additional costs of specialist treatment are greater than any cost savings from specialist treatment that accrue in the future. This is already implied by first-best Condition (4.6). Because fees are flat, the team would only use expensive specialists treatment for l-types, if specialist treatment is much superior to PCP treatment. In this case, however, specialist treatment for l-types would be preferred by the payer because the payer also considers the health benefits of the patient. As shown in Appendix 4.A.9, SP is also never played in the team with flat fees. Thus, specialist over-treatment is not an issue.

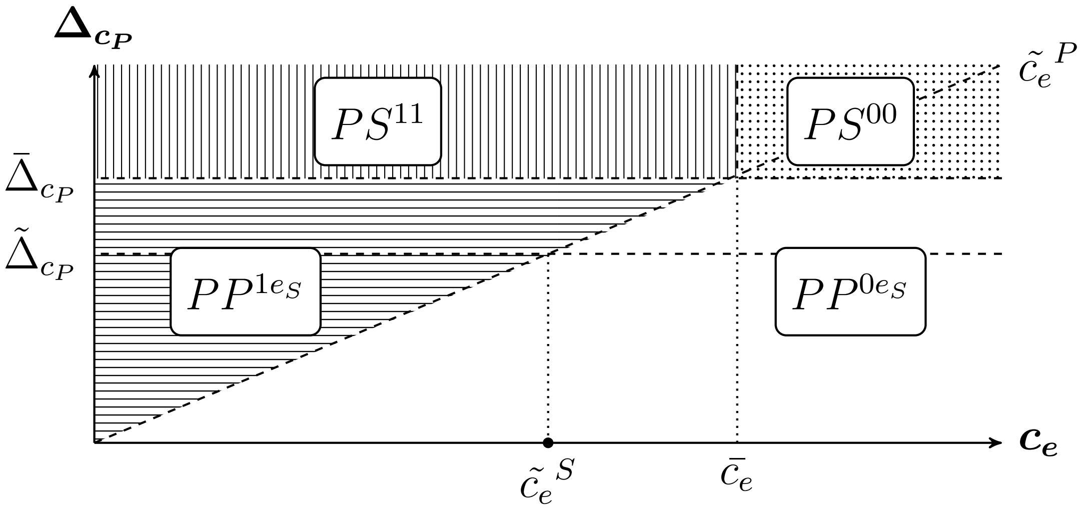


Figure 4.5.1: Outcomes in the team with flat fees ($\gamma_P = \gamma_S$).

Let the minimum c_e implied by Condition (4.25) be defined as $\bar{c_e}$ and the maximum Δ_{c_P} implied by Condition (4.26) as $\bar{\Delta}_{c_P}$, i.e.

$$\bar{c_e} := \frac{\Delta_p \beta (c_S^h - c_P^l)}{1 + \beta(1 - p_P^{ll} - p_S^{hh})}$$
$$\bar{\Delta}_{c_P} := \frac{(c_S^h - c_P^h)[1 + \beta(1 - p_P^{ll} - p_P^{hh})]}{\beta(p_P^{hh} - p_S^{hh})}.$$

Figure 4.5.1 depicts the outcomes in the team given flat fees. We can make the following observations in Lemma 4 (follows directly from Conditions (4.6) & (4.5)).

Lemma 4. *Compared to the optimal treatment fees, under flat fees, a larger difference in expected treatment costs for the PCP is necessary in order to implement the first-best outcome* PS^{11}*:*

$$\bar{\Delta}_{c_P} \geq \tilde{\Delta}_{c_P}$$

Given that implementing treatment path PS is cost-efficient ($\Delta_{c_P} \geq \bar{\Delta}_{c_P}$*), effort is implemented only for lower effort costs:*

$$\bar{c_e} \leq \tilde{c_e}^P$$

with

$$\bar{c_e} = \tilde{c_e}^P \textit{ if } \Delta_{c_P} = \bar{\Delta}_{c_P}.$$

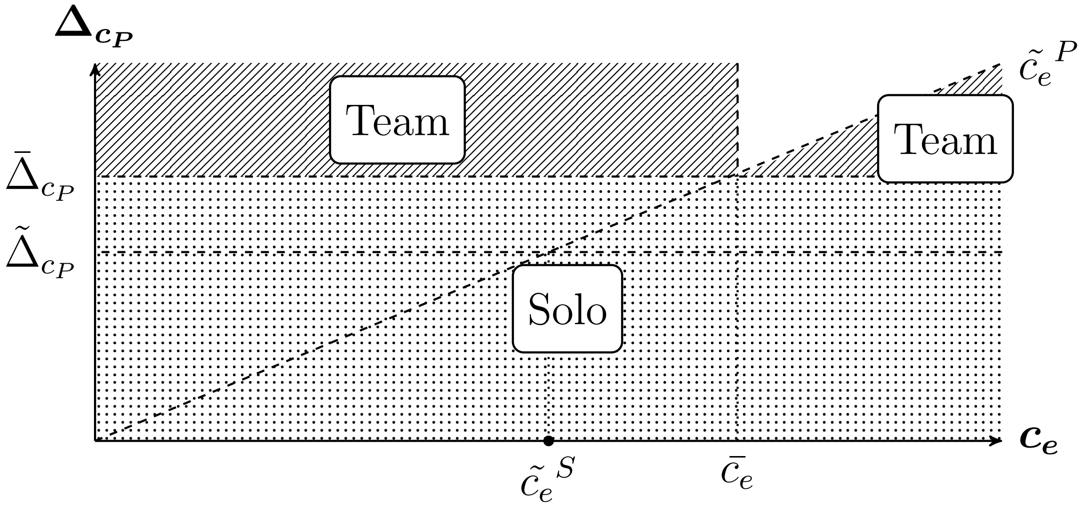


Figure 4.5.2: Second-best optimal organization of care given $\gamma_P = \gamma_S$: weak dominance for solo practices (dots), strict dominance for the team (lines), no general dominance relation (no pattern)

Furthermore, if $\bar{\Delta}_{c_P} < \tilde{\Delta}_{c_P}$, only blind PCP treatment is provided.

Lemma 4 and Figure 4.5.1 highlight the three problems that exist when fees are flat. Firstly, for $\Delta_{c_P} \in [\tilde{\Delta}_{c_P}, \bar{\Delta}_{c_P}]$, the treatment fee for the PCP is too large compared to the specialist's fee. The result is too little specialist treatment. This could have been prevented by a relative increase in the specialist's fee. Secondly, SP^{11} can not be implemented anymore. Consequently, as Figure 4.5.2 shows, the area for which solo practices weakly dominate team care is increased. Thirdly, for $\Delta_{c_P} \geq \bar{\Delta}_{c_P}$ and $c_e \in [\bar{c_e}, \tilde{c_e}^P]$, the treatment fee for the PCP is too small compared to the specialist's fee. The result of this is an undersupply of effort, which could have been prevented by a relative increase in the PCP's fee.

As Figure 4.5.2 shows, team care does not strictly dominate solo practices for all parameters given $\Delta_{c_P} > \bar{\Delta}_{c_P}$. Instead, there exists a parameter region $c_e \in [\bar{c_e}, \tilde{c_e}^P]$, in which the team plays PS^{00} and the solo practices play either PP^{1es} or MS^{10}. Depending on the exact parameters, either organizational form may perform better. Here, the team has the advantage for the referral efficiency, while the solo practices can provide larger PCP effort. Proposition 7 summarizes the results.

Proposition 7. *Let teams be paid by flat fees and solo practices by differing fees. Then the following observations can be made.*

1. *Solo practices weakly dominate team care whenever treatment path PS is not cost-efficient* ($\Delta_{c_P} < \bar{\Delta}_{c_P}$).
2. *Given that PS is cost-efficient, team care (strictly) dominates solo practice care (given that Conditions (4.4) to (4.6) hold strictly), except when effort costs are moderate* ($c_e \in [\bar{c_e}, \tilde{c_e}^P]$)*. In this case, neither organizational form is dominant for all parameters.*

4.5.2 Physician Altruism

Let us now consider how the main result of Section 4.4 changes if physicians are partially altruistic. As in the seminal model of Ellis and T. G. McGuire (1986) physicians are assumed to consider the patient's benefit in addition to their own profits. In order to ensure a fair comparison between the organizational settings, I assume that both physicians and the physician team internalize patient benefit to the same degree α which is commonly known, i.e. the solo practice physicians' continuation utility is

$$u_k^{x\alpha}(s) = -\alpha\mathcal{L}(s,x) + \Gamma_k(s,x) - \mathcal{C}_k(s,x) + \beta \sum_{\tilde{x}\in\mathcal{X}} p(\tilde{x}|s,x)u_k^{\tilde{x}\alpha}(s), x \in \mathcal{X} \tag{4.28}$$

and the discounted expected utility in period $t = 0$ is

$$U_k^{\alpha}(s) = F_k + l_P^0 u_k^{l_P\alpha}(s) + l_S^0 u_k^{l_S\alpha}(s) + h_P^0 u_k^{h_P\alpha}(s) + h_S^0 u_k^{h_S\alpha}(s). \tag{4.29}$$

For the team, continuation utility is

$$\begin{aligned} u_T^{x\alpha}(s) = &-\alpha\mathcal{L}(s,x) + \Gamma_P(s,x) + \Gamma_S(s,x) - \mathcal{C}(s,x) + \\ &\beta \sum_{\tilde{x}\in\mathcal{X}} p(\tilde{x}|s,x)u_T^{\tilde{x}\alpha}(s), x \in \mathcal{X} \end{aligned} \tag{4.30}$$

and the discounted expected utility in period $t = 0$ is

$$U_T^{\alpha}(s) = F_T + l_P^0 u_T^{l_P\alpha}(s) + l_S^0 u_T^{l_S\alpha}(s) + h_P^0 u_T^{h_P\alpha}(s) + h_S^0 u_T^{h_S\alpha}(s). \tag{4.31}$$

I assume, for simplicity, that solo practice physicians are indifferent between the other physician or the hospital treating a patient. Using the following two-step argument I will show that physician altruism increases the set of parameters for which the team weakly dominates the solo practices in terms of social welfare.

1. If the team acts in an altruistic manner, the parameter region in which PS can be implemented grows.

2. Given that PS can be implemented in the team, effort is always as least as large in the team as it is in the solo practices.

Consider step 1 first. Including altruism changes Conditions (4.20) to (4.22) for the implementation of PS^{11} to:

$$PS^{11} \geq SS^{e_P1} : (\gamma_S - c_S^l) - (\gamma_P - c_P^l) \leq \frac{(p_P^{ll} - p_S^{ll})\beta(\alpha L + \Delta_{c_S})}{1 + \beta(1 - p_S^{ll} - p_S^{hh})} \tag{4.32}$$

$$PS^{11} \geq PS^{00} : (\gamma_S - c_S^h) - (\gamma_P - c_P^l) \leq \alpha L - \frac{c_e[1 + \beta(1 - p_P^{ll} - p_S^{hh})]}{\beta \Delta_p} \tag{4.33}$$

$$PS^{11} \geq PP^{1e_S} : (\gamma_S - c_S^h) - (\gamma_P - c_P^h) \geq -\frac{(p_P^{hh} - p_S^{hh})\beta(\alpha L + \Delta_{c_P})}{1 + \beta(1 - p_P^{ll} - p_P^{hh})} \tag{4.34}$$

If $c_e \leq \tilde{c_e}^{S\alpha}$ with

$$\tilde{c_e}^{k\alpha} := \frac{\Delta_p \beta(\alpha L + \Delta_{c_k})}{1 + \beta(1 - p_k^{ll} - p_k^{hh})}, k \in \{P, S\},$$

Condition (4.32) is stricter than (4.33). Otherwise, the opposite is true. PS can be implemented if and only if Conditions (4.32) and (4.34) can be fulfilled simultaneously:

$$\tilde{c_e}^{P\alpha} \geq \tilde{c_e}^{S\alpha} \iff$$
$$\frac{\Delta_{c_P} + \alpha L}{1 + \beta(1 - p_P^{ll} - p_P^{hh})} \geq \frac{\Delta_{c_S} + \alpha L}{1 + \beta(1 - p_S^{ll} - p_S^{hh})}.$$

As assumed, the specialist has a stronger treatment advantage for h-types: $p_P^{hh} - p_S^{hh} > p_S^{ll} - p_P^{ll}$. Thus follows step 1.:

$$\tilde{\Delta}_{c_P} > \tilde{\Delta}_{c_P}^{\alpha} := \frac{(\Delta_{c_S} + \alpha L)[1 + \beta(1 - p_P^{ll} - p_P^{hh})]}{1 + \beta(1 - p_S^{ll} - p_S^{hh})} - \alpha L$$

Then, if effort costs are low enough, i.e. $c_e \leq \tilde{c_e}^{P\alpha}$, PS^{11} can be implemented. Otherwise, PS^{00} can be implemented.

Figure 4.5.3 illustrates the result. Compared to the case without altruism, the area in which PS (especially PS^{11}) can be implemented grows. For $\Delta_{c_P} < \tilde{\Delta}_{c_P}^{\alpha}$, only blind treatment paths

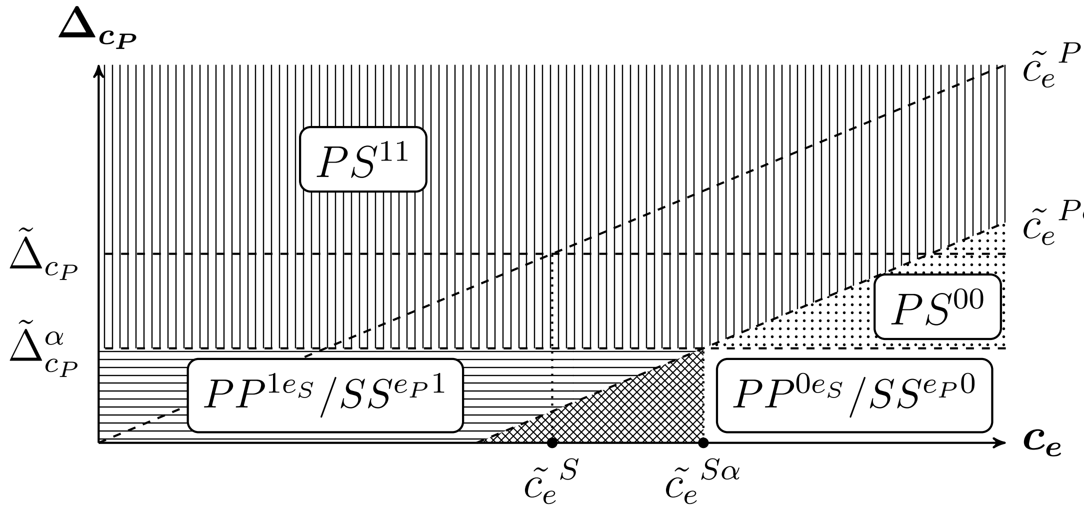


Figure 4.5.3: Second-best implementation in the altruistic team: crosshatch pattern = $PP^{0e_S}/SP^{11}/SS^{e_P1}$

or SP^{11} can be implemented second-best optimally. In the special case $\alpha = 1$, the first-best can be always be implemented by a flat fee $\gamma_P = \gamma_S$.

It remains to be proven that in the solo practices effort is provided only if $c_e \leq \tilde{c}_e^{\,P\alpha}$ for any treatment path. This is done in Appendix 4.A.10. In particular, PS^{11} can not be implemented for

$$c_e > \Delta\tilde{c}_e^{\,P} := \frac{\Delta_p \alpha L \beta}{1 + \beta(1 - p_P^{ll} - p_P^{hh})} = \tilde{c}_e^{\,P\alpha} - \tilde{c}_e^{\,P}.$$

It follows Proposition 8. Figure 4.5.4 visualizes the proposition.

Proposition 8. *If physicians are altruistic, the set of parameters for which the team weakly dominates the solo practices in terms of social welfare is increased. Furthermore, if*

$$c_e \in (\Delta\tilde{c}_e^{\,P}, \tilde{c}_e^{\,P\alpha}),$$

the team strictly dominates the solo practices given that Conditions (4.4) to (4.6) hold strictly.

This result can be interpreted in the following way. With increasing altruism, the team internalizes the patient's health loss. As there is no coordination problem between the physicians, the outcome approaches the first-best. In contrast, in the solo practices the coordination problem remains.

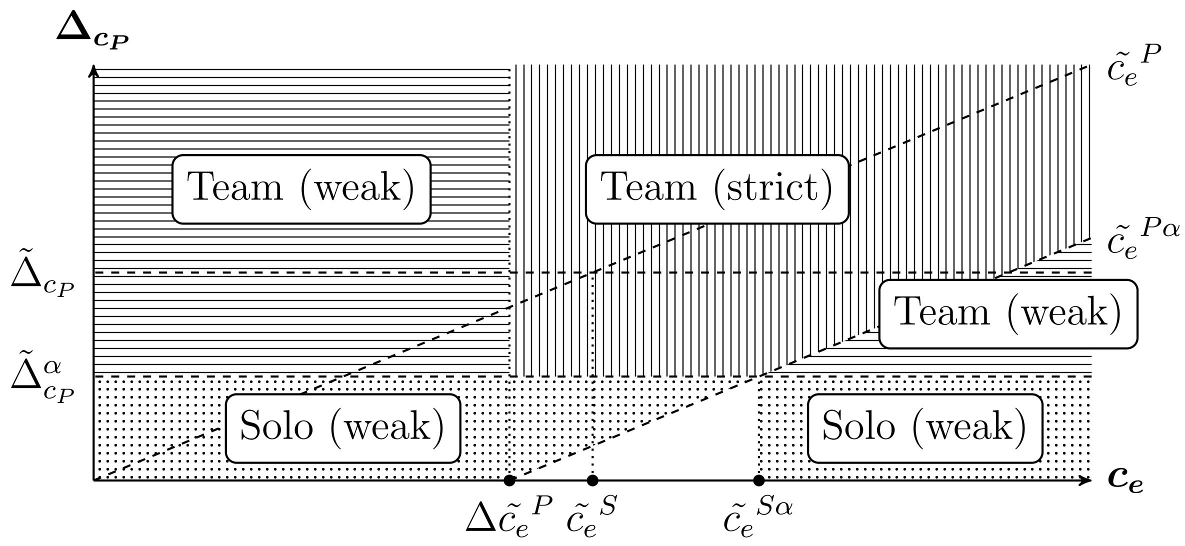


Figure 4.5.4: Second-best optimal organization of care for altruistic physicians (vertical lines: strict team dominance, horizontal lines: weak team dominance, dots: weak solo practice dominance, no pattern: no general dominance relation)

4.6 Conclusion

This paper studies the referral and effort provision of two physicians, a PCP and a specialist, who are responsible for treatment of chronically ill patients. I compare two organizational forms, solo practices and a physician team. In both organizational forms non-contractible effort can be incentivized for profit-maximizing physicians if effort costs are low enough. Effort can be incentivized because in this paper's dynamic setting effort provision today lowers physicians' future costs. If physicians are profit-maximizers, an optimal flow of patients between physicians (patients in mild conditions receive treatment from the PCP and patients in severe condition receive treatment from the specialist) can only be achieved in the team. However, it is necessary that the expected treatment cost differences between the patient types are relatively large for the PCP and small for the specialist, i.e. the PCP must have the relative cost-advantage when treating patients in mild condition. In this case, the first-best outcome is implemented by above-cost treatment fees for PCP treatment and below-cost fees for specialist treatment.

Consider the converse case in which the specialist has the relative cost advantage. If effort costs are either sufficiently small or sufficiently large, social welfare in the solo practice care is weakly superior to team care. For medium levels of effort there is a trade-off between better referral efficiency in the solo practices and higher effort in the team. Organizing physicians in solo practices allows the PCP to act as a gatekeeper, who treats all patients in mild condition until they are severely ill. Once they are severely ill, the PCP refers the patients to the specialist, who continues treatment. If the PCP's treatment costs for severely ill patients are sufficiently large, this gatekeeping outcome is strictly superior to all outcomes in the team. Thus, solving the coordination problem by organizing the physicians in a team does not necessarily enhance social welfare. A necessary condition for the superiority of solo practices is that health losses in severe condition are sufficiently large. Otherwise, the first-best can always be implemented in the team. For small health losses, profit-orientation of physicians is less of an issue, whereas the coordination problem remains, which is solved in the team but not the solo practices.

Let us turn to the extensions of the model. If treatment fees in the team are flat, solo practice care is weakly superior to team care if the first-best treatment path is not cost-minimizing. Furthermore, if the first-best treatment path is cost-minimizing, the team does not dominate the solo practices for medium levels of effort because effort incentives are stronger in the solo practices. Altruistic physicians were considered as a second extension. If physicians are partially altruistic, team care weakly dominates solo practice care for a greater parameter region. If they are perfectly altruistic, the first-best can always be implemented in the team.

Several policy implications follow from the analysis presented in this paper. Firstly, organizing physicians in separate solo practices may be superior to organizing them in a team. This is likely to be the case if both the costs and benefits of specialist treatment for severely ill patients are large. Secondly, regardless of organizational form, markups on PCP treatment can enhance effort provision. If, however, a team specialist generates most of the team's profits, this can have an adverse effect on effort provision because the team profits more from sick rather than healthy patients. In some physician teams and hospitals, PCPs have taken on the role of a loss-leader. Their main contribution to the team's profits is to refer patients to specialists for more costly procedures. According to model presented in this paper, this type of arrangement may lead to inferior care because effort incentives are muted and over-referral of patients in incentivized.

The main innovation of this paper is that it considers agency problems regarding referrals and effort provision in a dynamic framework. Because this paper focuses on the efficiency of different organizational forms of care, a simplified payment system is used. Future work could allow for more complexity. For example, the length of stay of the patient at a physician provides information on the state of the patient. Thus, the payer could use it as a basis for the size of treatment fees. Alternatively, the payer could consider a budgeted system. Furthermore, capacity constraints of physicians, diagnostic uncertainty, and informational asymmetries between the physicians regarding the type of the patient could be studied.

Appendices

4.A Proof of Propositions and Lemmas

A general note: When calculating expected welfare or profit between different strategies s, it is sufficient to only consider states whose continuation welfare/profit is different between strategies. For the sake of brevity I may omit states for which there is no such difference. The symbolic math toolbox of MATLAB (2016) was used in the calculation of the proofs.

4.A.1

Proof for Lemma 2 (First-best conditions):

Proof. I consider all possible outcomes of the game and show that, given Conditions (4.4) to (4.6), PS^{11} yields greater or equal expected welfare for any initial distribution of patients.

Optimal treatment effort:

First, I consider the condition under which providing effort is optimal given that PCPs treat l-types and specialists treat h-types. The system of equations for the continuation welfare of

$PS^{e_P e_S}$ in all states $x \in \mathcal{X}$ are given by:

$$\begin{aligned}
W^{l_P}_{PS^{e_P e_S}} &= -c^l_P - e_P c_e + \beta[(p^{ll}_P + e_P\Delta_p)W^{l_P}_{PS^{e_P e_S}} + (1 - p^{ll}_P - e_P\Delta_p)W^{h_S}_{PS^{e_P e_S}}] \\
W^{l_S}_{PS^{e_P e_S}} &= W^{l_P}_{PS^{e_P e_S}} \\
W^{h_P}_{PS^{e_P e_S}} &= W^{h_S}_{PS^{e_P e_S}} \\
W^{h_S}_{PS^{e_P e_S}} &= -L - c^h_S - e_S c_e + \beta[(p^{hh}_S - e_S\Delta_p)W^{h_S}_{PS^{e_P e_S}} + (1 - p^{hh}_S + e_S\Delta_p)W^{l_P}_{PS^{e_P e_S}}]
\end{aligned}$$

Solving this system leads to the following continuation welfare:

$$\begin{aligned}
W^{l_P}_{PS^{11}} = W^{l_S}_{PS^{11}} &= \frac{\frac{-\beta(1-p^{ll}_P-\Delta_p)(L+c^h_S-c^l_P)}{1+\beta(1-p^{ll}_P-p^{hh}_S)} - c^l_P - c_e}{1-\beta} \\
W^{h_P}_{PS^{11}} = W^{h_S}_{PS^{11}} &= \frac{\frac{-(1-\beta p^{ll}_P-\beta\Delta_p)(L+c^h_S-c^l_P)}{1+\beta(1-p^{ll}_P-p^{hh}_S)} - c^l_P - c_e}{1-\beta}
\end{aligned}$$

Analogously, solving the systems for the other combinations of effort levels leads to continuation welfare:

$$\begin{aligned}
W^{l_P}_{PS^{10}} = W^{l_S}_{PS^{10}} &= \frac{\frac{(1-\beta p^{hh}_S)(L+c^h_S-c^l_P-c_e)}{1+\beta(1-p^{ll}_P-p^{hh}_S-\Delta_p)} - L - c^h_S}{1-\beta} \\
W^{h_P}_{PS^{10}} = W^{h_S}_{PS^{10}} &= \frac{\frac{\beta(1-p^{hh}_S)(L+c^h_S-c^l_P-c_e)}{1+\beta(1-p^{ll}_P-p^{hh}_S-\Delta_p)} - L - c^h_S}{1-\beta} \\
W^{l_P}_{PS^{01}} = W^{l_S}_{PS^{01}} &= \frac{\frac{-\beta(1-p^{ll}_P)(L+c^h_S-c^l_P+c_e)}{1+\beta(1-p^{ll}_P-p^{hh}_S+\Delta_p)} - c^l_P}{1-\beta} \\
W^{h_P}_{PS^{01}} = W^{h_S}_{PS^{01}} &= \frac{\frac{-(1-\beta p^{ll}_P)(L+c^h_S-c^l_P+c_e)}{1+\beta(1-p^{ll}_P-p^{hh}_S+\Delta_p)} - c^l_P}{1-\beta} \\
W^{l_P}_{PS^{00}} = W^{l_S}_{PS^{00}} &= \frac{\frac{-\beta(1-p^{ll}_P)(L+c^h_S-c^l_P)}{1+\beta(1-p^{ll}_P-p^{hh}_S)} - c^l_P}{1-\beta} \\
W^{h_P}_{PS^{00}} = W^{h_S}_{PS^{00}} &= \frac{\frac{-(1-\beta p^{ll}_P)(L+c^h_S-c^l_P)}{1+\beta(1-p^{ll}_P-p^{hh}_S)} - c^l_P}{1-\beta}
\end{aligned}$$

It follows

$$\begin{aligned}
&\boldsymbol{E}W(PS^{11}) \geq \boldsymbol{E}W(PS^{10}) \iff W^{l_P}_{PS^{11}} \geq W^{l_P}_{PS^{10}} \iff W^{h_S}_{PS^{11}} \geq W^{h_S}_{PS^{10}} \iff \\
&\boldsymbol{E}W(PS^{11}) \geq \boldsymbol{E}W(PS^{01}) \iff W^{l_P}_{PS^{11}} \geq W^{l_P}_{PS^{01}} \iff W^{h_S}_{PS^{11}} \geq W^{h_S}_{PS^{01}} \iff \\
&\boldsymbol{E}W(PS^{11}) \geq \boldsymbol{E}W(PS^{00}) \iff W^{l_P}_{PS^{11}} \geq W^{l_P}_{PS^{00}} \iff W^{h_S}_{PS^{11}} \geq W^{h_S}_{PS^{00}} \iff \\
&\qquad c_e \leq \frac{\Delta_p\beta(L + c^h_S - c^l_P)}{1 + \beta(1 - p^{ll}_P - p^{hh}_S)}.
\end{aligned}$$

Clearly, varying a physician's effort level for a patient type that she does not treat, has no impact on expected welfare. These cases are omitted here.

For the remaining outcomes the procedure is analogous. I will present them in an abridged manner.

PS vs. PP:

If effort provision is efficient for one patient type, it is efficient for the other type as well:

$$\begin{aligned}
&\boldsymbol{E}W(PP^{11e_S^l e_S^h}) \geq \boldsymbol{E}W(PP^{10e_S^l e_S^h}) \iff \boldsymbol{E}W(PP^{11e_S^l e_S^h}) \geq \boldsymbol{E}W(PP^{01e_S^l e_S^h}) \iff \\
&\boldsymbol{E}W(PP^{11e_S^l e_S^h}) \geq \boldsymbol{E}W(PP^{00e_S^l e_S^h}) \iff c_e \leq \frac{\Delta_p \beta(L + c_P^h - c_P^l)}{1 + \beta(1 - p_P^{ll} - p_P^{hh})} \\
&\boldsymbol{E}W(PP^{00e_S^l e_S^h}) \geq \boldsymbol{E}W(PP^{10e_S^l e_S^h}) \iff \boldsymbol{E}W(PP^{00e_L^l e_S^h}) \geq \boldsymbol{E}W(PP^{01e_S^l e_S^h}) \iff \\
&\boldsymbol{E}W(PP^{00e_S^l e_S^h}) \geq \boldsymbol{E}W(PP^{11e_S^l e_S^h}) \iff c_e \geq \frac{\Delta_p \beta(L + c_P^h - c_P^l)}{1 + \beta(1 - p_P^{ll} - p_P^{hh})}
\end{aligned}$$

Therefore, only PP^{1e_S} and PP^{0e_S} need consideration.

$$\begin{aligned}
\boldsymbol{E}W(PS^{11}) \geq \boldsymbol{E}W(PP^{1e_S}) &\iff \boldsymbol{E}W(PS^{00}) \geq \boldsymbol{E}W(PP^{0e_S}) \\
&\iff c_S^h - c_P^h \leq \frac{(p_P^{hh} - p_S^{hh})\beta(L + c_P^h - c_P^l)}{1 + \beta(1 - p_P^{ll} - p_P^{hh})}
\end{aligned}$$

PS vs. SS:

If effort provision is efficient for one patient type, it is efficient for the other type as well:

$$\begin{aligned}
&\boldsymbol{E}W(SS^{e_P^l e_P^h 11}) \geq \boldsymbol{E}W(SS^{e_P^l e_P^h 10}) \iff \boldsymbol{E}W(SS^{e_P^l e_P^h 11}) \geq \boldsymbol{E}W(SS^{e_P^l e_P^h 01}) \iff \\
&\boldsymbol{E}W(SS^{e_P^l e_P^h 11}) \geq \boldsymbol{E}W(SS^{e_P^l e_P^h 00}) \iff c_e \leq \frac{\Delta_p \beta(L + c_S^h - c_S^l)}{1 + \beta(1 - p_S^{ll} - p_S^{hh})} \\
&\boldsymbol{E}W(SS^{e_P^l e_P^h 00}) \geq \boldsymbol{E}W(SS^{e_P^l e_P^h 10}) \iff \boldsymbol{E}W(SS^{e_P^l e_P^h 00}) \geq \boldsymbol{E}W(SS^{e_P^l e_P^h 01}) \iff \\
&\boldsymbol{E}W(SS^{e_P^l e_P^h 00}) \geq \boldsymbol{E}W(SS^{e_P^l e_P^h 11}) \iff c_e \geq \frac{\Delta_p \beta(L + c_S^h - c_S^l)}{1 + \beta(1 - p_S^{ll} - p_S^{hh})}
\end{aligned}$$

Therefore, only $SS^{e_P 1}$ and $SS^{e_P 0}$ need consideration.

$$\begin{aligned}
\boldsymbol{E}W(PS^{11}) \geq \boldsymbol{E}W(SS^{e_P 1}) &\iff \boldsymbol{E}W(PS^{00}) \geq \boldsymbol{E}W(SS^{e_P 0}) \\
&\iff c_S^l - c_P^l \geq \frac{(p_S^{ll} - p_P^{ll})\beta(L + c_S^h - c_P^l)}{1 + \beta(1 - p_P^{ll} - p_S^{hh})}
\end{aligned}$$

PS vs. PM: First note that if PCP effort provision in PM is efficient for one of the types, it must be efficient for the other as well. This is so because the continuation welfare of PM is identical to that of PP for any patient who is treated by the PCP. Therefore, it is sufficient to consider only cases in which physicians either provide effort to all patient types they receive or none.

$$\boldsymbol{E}W(PS^{11}) \geq \boldsymbol{E}W(PM^{11}) \iff \boldsymbol{E}W(PS^{00}) \geq \boldsymbol{E}W(PM^{00})$$
$$\iff c_S^h - c_P^h \leq \frac{(p_P^{hh} - p_S^{hh})\beta(L + c_P^h - c_P^l)}{1 + \beta(1 - p_P^{ll} - p_P^{hh})}$$

If $\boldsymbol{E}W(PM^{11}) \geq \boldsymbol{E}W(PM^{10})$, $\boldsymbol{E}W(PS^{11}) \geq \boldsymbol{E}W(PM^{10})$ follows immediately. Let instead $\boldsymbol{E}W(PM^{10}) \geq \boldsymbol{E}W(PM^{11})$. This implies a lower boundary $c_e \geq \hat{c_e}$ with

$$\hat{c_e} := \frac{\Delta_p\beta\{L - c_P^l + c_S^h - \beta(1 - p_P^{ll} - \Delta_p)(L + c_P^h - c_P^l)(1 + \beta[1 - p_p^{ll} - p_P^{hh}])\}}{1 - \beta(p_S^{hh} - \Delta_p)}.$$

Further,

$$\begin{aligned}
&\boldsymbol{E}W(PS^{10}) - \boldsymbol{E}W(PM^{10}) = \\
&\frac{\beta^2(1 - p_S^{hh})(1 - p_P^{ll} - \Delta_p)(L + c_P^h - c_P^l)(1 + \beta[1 - p_P^{ll} - p_P^{hh}])}{(1-\beta)(1 - \beta p_S^{hh})} - \\
&\frac{\beta^2(1 - p_S^{hh})(1 - p_P^{ll} - \Delta_p)(L - c_P^l + c_S^h - c_e)(1 + \beta[1 - p_P^{ll} - p_S^{hh} - \Delta_p])}{(1-\beta)(1 - \beta p_S^{hh})} \\
&\implies \frac{\partial(\boldsymbol{E}W(PS^{10}) - \boldsymbol{E}W(PM^{10}))}{\partial c_e} > 0.
\end{aligned}$$

Thus, if PS^{10} is superior to PM^{10} for some level of c_e, it is also superior for any larger level. Inserting $\hat{c_e}$ delivers

$$\boldsymbol{E}W(PS^{11}) \geq \boldsymbol{E}W(PM^{11}) \implies \boldsymbol{E}W(PS^{10}) \geq \boldsymbol{E}W(PM^{10}).$$

An analogue argument shows that

$$\boldsymbol{E}W(PS^{11}) \geq \boldsymbol{E}W(PM^{11}) \implies \boldsymbol{E}W(PS^{01}) \geq \boldsymbol{E}W(PM^{01}).$$

PS vs. MS: An analogue argument applies as in the previous case.

PS vs. MM: MM is weakly dominated by the set of all other outcomes due to the linearity of costs and benefits.

PS vs. SP:

First note that SP^{11} and SP^{00} are superior to SP^{10} and SP^{01}:

$$\begin{aligned}
\boldsymbol{E}W(SP^{11}) \geq \boldsymbol{E}W(SP^{10}) &\iff \boldsymbol{E}W(SP^{11}) \geq \boldsymbol{E}W(SP^{01}) \\
\iff \boldsymbol{E}W(SP^{11}) \geq \boldsymbol{E}W(SP^{00}) &\iff c_e \leq \frac{\Delta_p \beta(L + c_P^h - c_S^l)}{1 + \beta(1 - p_S^{ll} - p_P^{hh})} \\
\boldsymbol{E}W(SP^{00}) \geq \boldsymbol{E}W(SP^{10}) &\iff \boldsymbol{E}W(SP^{00}) \geq \boldsymbol{E}W(SP^{01}) \\
\iff \boldsymbol{E}W(SP^{00}) \geq \boldsymbol{E}W(SP^{11}) &\iff c_e \geq \frac{\Delta_p \beta(L + c_P^h - c_S^l)}{1 + \beta(1 - p_S^{ll} - p_P^{hh})}
\end{aligned}$$

Thus, it is sufficient to show the superiority of PS^{11} over SP^{11} and SP^{00}. $SP^{e_P e_S}$ is dominated by the blind outcomes:

$$\boldsymbol{E}W(PP^{1e_S}) \geq \boldsymbol{E}W(SP^{11}) \iff \boldsymbol{E}W(PP^{0e_S}) \geq \boldsymbol{E}W(SP^{00}) \iff$$
$$\frac{L + c_P^h - c_P^l}{1 + \beta(1 - p_P^{ll} - p_P^{hh})} \geq \frac{L + c_P^h - c_S^l}{1 + \beta(1 - p_S^{ll} - p_P^{hh})}$$

If $p_P^{ll} \geq p_S^{ll}$, this condition is always fulfilled. Otherwise, it bounds L from above. Further,

$$\boldsymbol{E}W(SS^{e_P 1}) \geq \boldsymbol{E}W(SP^{11}) \iff \boldsymbol{E}W(SS^{e_P 0}) \geq \boldsymbol{E}W(SP^{00}) \iff$$
$$L \geq c_S^l - c_P^h - \frac{(c_P^h - c_S^h)[1 + \beta(1 - p_S^{ll} - p_P^{hh})]}{\beta(p_P^{hh} - p_S^{hh})}.$$

One of the conditions is fulfilled if

$$\frac{c_S^h - c_P^h}{p_P^{hh} - p_S^{hh}} \leq \frac{c_S^l - c_P^l}{p_S^{ll} - p_P^{ll}}.$$

This is just Condition (4.7). As PS^{11} is superior to both $PP^{e_P e_S}$ and $SS^{e_P e_S}$, it follows $PS^{11} \geq SP^{e_P e_S}$.

PS vs. SM:

First note that if specialist effort provision in SM is efficient for one of the types, it must be efficient for the other as well. This is so because the continuation welfare of SM is identical to that of SS for any patient who is treated by the specialist. Therefore, it is sufficient to consider only cases in which physicians either provide effort to all patient types they receive or none.

$SM^{e_P e_S}$ is dominated:

$$\boldsymbol{E}W(SS^{e_P 1}) \geq \boldsymbol{E}W(SM^{11}) \iff c_S^h - c_P^h \leq \frac{(p_P^{hh} - p_S^{hh})\beta(L + c_P^h - c_S^l)}{1 + \beta(1 - p_S^{ll} - p_P^{hh})}$$

$$\boldsymbol{E}W(SP^{11}) \geq \boldsymbol{E}W(SM^{11}) \iff c_S^h - c_P^h \geq \frac{(p_P^{hh} - p_S^{hh})\beta(L + c_P^h - c_S^l)}{1 + \beta(1 - p_S^{ll} - p_P^{hh})}$$

$$\boldsymbol{E}W(SS^{e_P 0}) \geq \boldsymbol{E}W(SM^{10}) \iff c_S^h - c_P^h - c_e \leq \frac{(p_P^{hh} - p_S^{hh} - \Delta_p)\beta(L + c_S^h - c_S^l)}{1 + \beta(1 - p_S^{ll} - p_S^{hh})}$$

$$\boldsymbol{E}W(SP^{10}) \geq \boldsymbol{E}W(SM^{10}) \iff c_S^h - c_P^h - c_e \geq \frac{(p_P^{hh} - p_S^{hh} - \Delta_p)\beta(L + c_S^h - c_S^l)}{1 + \beta(1 - p_S^{ll} - p_S^{hh})}$$

$$\boldsymbol{E}W(SS^{e_P 1}) \geq \boldsymbol{E}W(SM^{01}) \iff c_S^h - c_P^h + c_e \leq \frac{(p_P^{hh} - p_S^{hh} + \Delta_p)\beta(L + c_S^h - c_S^l)}{1 + \beta(1 - p_S^{ll} - p_S^{hh})}$$

$$\boldsymbol{E}W(SP^{01}) \geq \boldsymbol{E}W(SM^{01}) \iff c_S^h - c_P^h + c_e \geq \frac{(p_P^{hh} - p_S^{hh} + \Delta_p)\beta(L + c_S^h - c_S^l)}{1 + \beta(1 - p_S^{ll} - p_S^{hh})}$$

$$\boldsymbol{E}W(SS^{e_P 0}) \geq \boldsymbol{E}W(SM^{00}) \iff c_S^h - c_P^h \leq \frac{(p_P^{hh} - p_S^{hh})\beta(L + c_S^h - c_S^l)}{1 + \beta(1 - p_S^{ll} - p_S^{hh})}$$

$$\boldsymbol{E}W(SP^{00}) \geq \boldsymbol{E}W(SM^{00}) \iff c_S^h - c_P^h \geq \frac{(p_P^{hh} - p_S^{hh})\beta(L + c_S^h - c_S^l)}{1 + \beta(1 - p_S^{ll} - p_S^{hh})}$$

As PS^{11} is superior to $PP^{e_P e_S}$ and $SS^{e_P e_S}$, and $SP^{e_P e_S}$, it follows $PS^{11} \geq SM^{e_P e_S}$.

PS vs. MP: An analogue argument applies as in the previous case.

The proof for the strict superiority of PS^{11} follows by replacing inequalities with strict inequalities. □

4.A.2

Proof for Lemma 3 (welfare ordering):

Proof. Part 1:

$$\begin{aligned}
\boldsymbol{E}W(PS^{11}) \geq \boldsymbol{E}W(PM^{11}) &\iff \boldsymbol{E}W(PS^{00}) \geq \boldsymbol{E}W(PM^{00}) \\
\boldsymbol{E}W(PM^{11}) \geq \boldsymbol{E}W(PP^{1e_S}) &\iff \boldsymbol{E}W(PM^{00}) \geq \boldsymbol{E}W(PP^{0e_S}) \\
&\iff c_S^h - c_P^h \leq \frac{(p_P^{hh} - p_S^{hh})\beta(L + c_P^h - c_P^l)}{1 + \beta(1 - p_P^{ll} - p_P^{hh})} \\
\boldsymbol{E}W(PS^{11}) \geq \boldsymbol{E}W(MS^{11}) &\iff \boldsymbol{E}W(PS^{00}) \geq \boldsymbol{E}W(MS^{00}) \\
\boldsymbol{E}W(MS^{11}) \geq \boldsymbol{E}W(SS^{e_P1}) &\iff \boldsymbol{E}W(MS^{00}) \geq \boldsymbol{E}W(SS^{e_P0}) \\
&\iff c_S^l - c_P^l \geq \frac{(p_S^{ll} - p_P^{ll})\beta(L + c_S^h - c_P^l)}{1 + \beta(1 - p_P^{ll} - p_S^{hh})}
\end{aligned}$$

The part of the Proposition concerning strict inequality follows analogously.

Part 2: The dominance of PP and SS over MP and SM has been demonstrated in Appendix 4.A.1 already. In treatment path MM one part of the patients receives continuous PCP treatment and the other part receives continuous specialist treatment. Clearly, either PP or SS is weakly preferred by the payer. □

4.A.3

Proof for Proposition 2 (implementation of blind treatment paths):

Proof. The system of equations for the continuation profit for the PCP of $PP^{e_P^l e_P^h e_S^l e_S^h}$ in all states $x \in \mathcal{X}$ are given by:

$$\begin{aligned}
u_P^{l_P} &= -c_P^l - e_P^l c_e + \beta[(p_P^{ll} + e_P^l\Delta_p)u_P^{l_P} + (1 - p_P^{ll} - e_P^l\Delta_p)u_P^{h_P}] \\
u_P^{l_S} &= u_P^{l_P} \\
u_P^{h_P} &= -c_P^h - e_P^h c_e + \beta[(p_P^{hh} - e_P^h\Delta_p)u_P^{h_P} + (1 - p_P^{hh} + e_P^h\Delta_p)u_P^{l_P}] \\
u_P^{h_S} &= u_P^{h_P}
\end{aligned}$$

Solving this system leads to continuation profits:

$$\begin{aligned}
u_P^{l_P}(PP^{11e_S^l e_S^h}) = u_P^{l_S} &= \frac{\gamma_P - c_e - c_P^l}{1-\beta} - \frac{\beta(c_P^h - c_P^l)(1-p_P^{ll}-\Delta_p)}{(1-\beta)[1+\beta(1-p_P^{ll}-p_P^{hh})]} \\
u_P^{h_P}(PP^{11e_S^l e_S^h}) = u_P^{h_S} &= \frac{\gamma_P - c_e - c_P^l}{1-\beta} - \frac{(c_P^h - c_P^l)(1-\beta p_P^{ll}-\beta\Delta_p)}{(1-\beta)[1+\beta(1-p_P^{ll}-p_P^{hh})]} \\
u_P^{l_P}(PP^{10e_S^l e_S^h}) = u_P^{l_S} &= \frac{\gamma_P - c_P^h}{1-\beta} - \frac{(c_e - c_P^h + c_P^l)(1-\beta p_P^{hh})}{(1-\beta)[1+\beta(1-p_P^{ll}-p_P^{hh}-\Delta_p)]} \\
u_P^{h_P}(PP^{10e_S^l e_S^h}) = u_P^{h_S} &= \frac{\gamma_P - c_P^h}{1-\beta} - \frac{\beta(c_e - c_P^h + c_P^l)(1-p_P^{hh})}{(1-\beta)[1+\beta(1-p_P^{ll}-p_P^{hh}-\Delta_p)]} \\
u_P^{l_P}(PP^{01e_S^l e_S^h}) = u_P^{l_S} &= \frac{\gamma_P - c_P^l}{1-\beta} - \frac{\beta(c_e + c_P^h - c_P^l)(1-p_P^{hh})}{(1-\beta)[1+\beta(1-p_P^{ll}-p_P^{hh}-\Delta_p)]} \\
u_P^{h_P}(PP^{01e_S^l e_S^h}) = u_P^{h_S} &= \frac{\gamma_P - c_P^l}{1-\beta} - \frac{(c_e + c_P^h - c_P^l)(1-\beta p_P^{ll})}{(1-\beta)[1+\beta(1-p_P^{ll}-p_P^{hh}-\Delta_p)]} \\
u_P^{l_P}(PP^{00e_S^l e_S^h}) = u_P^{l_S} &= \frac{\gamma_P - c_P^l}{1-\beta} - \frac{\beta(c_P^h - c_P^l)(1-p_P^{ll})}{(1-\beta)[1+\beta(1-p_P^{ll}-p_P^{hh})]} \\
u_P^{h_P}(PP^{00e_S^l e_S^h}) = u_P^{h_S} &= \frac{\gamma_P - c_P^l}{1-\beta} - \frac{(c_P^h - c_P^l)(1-\beta p_P^{ll})}{(1-\beta)[1+\beta(1-p_P^{ll}-p_P^{hh})]}
\end{aligned}$$

The specialist always earns zero profit in this outcome. In treatment path SS there exists an analogue set of continuation profits for the specialist.

If providing effort is preferred for one patient type, it is also preferred for the other type:

$$\begin{aligned}
&U_P(PP^{1e_S}) \geq U_P(PP^{0e_S}), U_P(PP^{10e_S^l e_S^h}), U_P(PP^{01e_S^l e_S^h}) \iff \\
&u_P^x(PP^{1e_S}) \geq u_P^x(PP^{0e_S}), u_P^x(PP^{10e_S^l e_S^h}), u_P^x(PP^{01e_S^l e_S^h}) \forall x \in \mathcal{X} \iff \\
&\qquad c_e \leq \tilde{c_e}^P \\
&U_P(PP^{0e_S}) \geq U_P(PP^{1e_S}), U_P(PP^{10e_S^l e_S^h}), U_P(PP^{01e_S^l e_S^h}) \iff \\
&u_P^x(PP^{0e_S}) \geq u_P^x(PP^{1e_S}), u_P^x(PP^{10e_S^l e_S^h}), u_P^x(PP^{01e_S^l e_S^h}) \forall x \in \mathcal{X} \iff \\
&\qquad c_e \geq \tilde{c_e}^P.
\end{aligned}$$

An analogous result holds for the specialist in treatment path SS:

$$\begin{aligned}
&U_S(SS^{e_P 1}) \geq U_S(SS^{e_P 0}), U_S(SS^{e_P^l e_P^h 01}), U_S(SS^{e_P^l e_P^h 10}) \iff c_e \leq \tilde{c_e}^S \\
&U_S(SS^{e_P 0}) \geq U_S(SS^{e_P 1}), U_S(SS^{e_P^l e_P^h 01}), U_S(SS^{e_P^l e_P^h 10}) \iff c_e \geq \tilde{c_e}^S.
\end{aligned}$$

This completes the proof of the first part of Proposition 2.

The weak dominance of PP and SS over MM has already been argued in Appendix 4.A.2. Effort provision functions the same in path MM as it does in PP for the PCP and in SS for the specialist.

The following two step argument can be made to prove that the blind treatment paths weakly dominate treatment path MP.

Step 1: PP and SS weakly dominate MP for the same effort levels (see Lemma 3).

Step 2: Whenever effort can be incentivized for MP, it can be incentivized in the blind treatment paths.

Step 2 is trivial for the PCP as her incentives are identical to the blind treatment paths. Let the PCP play (T,T). For the specialist, two conditions need to hold in order to implement MP^{11}:

$$\begin{aligned}
U_S[(T,R)^{1e_S^h}] \geq U_S[(T,R)^{0e_S^h}] &\iff u_S^{l_S}[(T,R)^{1e_S^h}] \geq u_S^{l_S}[(T,R)^{0e_S^h}] \\
&\iff \gamma_S \geq c_S^l + \frac{c_e(1-\beta p_S^{ll})}{\beta \Delta_p} \\
U_S[(T,R)^{1e_S^h}] \geq U_S[(T,T)^{11}] &\iff u_S^{l_S}[(T,R)^{1e_S^h}] \geq u_S^{l_S}[(T,T)^{11}] \\
&\iff u_S^{h_S}[(T,R)^{1e_S^h}] \geq u_S^{h_S}[(T,T)^{11}] \\
&\iff \gamma_S \leq c_S^l + c_e + \frac{(c_S^h - c_S^l)(1-\beta(p_S^{ll}-\Delta_p))}{1+\beta(1-p_S^{ll}-p_S^{hh})}
\end{aligned}$$

These conditions can be fulfilled together if and only if $c_e \leq \tilde{c_e}^S$. Thus, the specialist is always willing to provide effort in SS if she is providing effort in MP. □

4.A.4

Proof for Proposition 3 (gatekeeping treatment path MS):

Proof. Let the specialist play $s_S = (T,T)^{11}$. The following describes the conditions for $s_P = (T,R)^{1e_P^h}$ being a best response for the PCP for all states.

Optimal effort choice:

Continuation profits for $s_P = (T,R)^{e^l_P e^h_P}$ are given by:

$$u_P^{l_P} = \frac{\gamma_P - c^l_P - e^l_P c_e}{1-\beta(p^{ll}_P + e^l_P \Delta_p)}$$
$$u_P^{h_P} = u_P^{l_S} = u_P^{h_S} = 0$$

It follows

$$\begin{aligned} U_P[(T,R)^{1e^h_P}] \geq U_P[(T,R)^{0e^h_P}] &\iff u_P^{l_P}[(T,R)^{1e^h_P}] \geq u_P^{l_P}[(T,R)^{0e^h_P}] \\ &\iff \gamma_P \geq c^l_P + \frac{c_e(1-\beta p^{ll}_P)}{\beta \Delta_p}. \end{aligned} \tag{4.35}$$

$s_P = (T,R)^{e^l_P e^h_P}$ vs. $s_P = (T,T)^{e^l_P e^h_P}$: For $s_P = (T,T)^{e^l_P e^h_P}$, continuation profits, and thus incentives for effort provision, are identical to outcome $PP^{e^l_P e^h_P}11$ (see Appendix 4.A.3).

$$\begin{aligned} U_P[(T,R)^{1e^h_P}] \geq U_P[(T,T)^{11}] &\iff u_P^{l_P}[(T,R)^{1e^h_P}] \geq u_P^{l_P}[(T,T)^{11}] \\ &\iff u_P^{h_P}[(T,R)^{1e^h_P}] \geq u_P^{h_P}[(T,T)^{11}] \\ &\iff \gamma_P \leq c^l_P + c_e + \frac{\Delta_{c_P}(1-\beta(p^{ll}_P + \Delta_p))}{1+\beta(1-p^{ll}_P - p^{hh}_P)} \end{aligned} \tag{4.36}$$

$$U_P[(T,T)^{11}] \geq U_P[(T,T)^{00}], U_P[(T,T)^{10}], U_P[(T,T)^{01}] \iff c_e \leq \tilde{c_e}^P$$

Conditions (4.35) and (4.36) can only be simultaneously fulfilled if and only if $c_e \leq \tilde{c_e}^P$.

$s_P = (T,R)^{e^l_P e^h_P}$ vs. $s_P = (R,R)^{e^l_P e^h_P}$:

For $s_P = (R,R)^{e^l_P e^h_P}$ continuation profits for the PCP are 0.

$$\begin{aligned} U_P[(T,R)^{1e^h_P}] \geq U_P[(R,R)^{e^l_P e^h_P}] &\iff u_P^{l_P}[(T,R)^{1e^h_P}] \geq u_P^{l_P}[(R,R)^{e^l_P e^h_P}] \\ &\iff \gamma_P \geq c^l_P + c_e \end{aligned}$$

This is implied by Condition (4.35).

Now, let the PCP play $s_P = (T,R)^{1e^h_P}$. The following describes the conditions for $s_S = (T,T)^{11}$ being a best response for the specialist for all states. If the specialist is willing to treat h-types, then she is willing to treat l-types. Thus, it is sufficient to show that the specialist is willing to accept h-types and that she is willing to provide effort.

Optimal effort choice:

For any patient who is treated by the specialist, continuation profits are identical so outcome $SS^{e_P^l e_P^h e_S^l e_S^h}$. Continuation profits for h_P are identical to those for h_S. Continuation profits for l_P only include an additional discount factor.

$$U_S[(T,T)^{11}] \geq U_S[(T,T)^{00}], U_S[(T,T)^{10}], U_S[(T,T)^{01}] \iff$$
$$\iff c_e \leq \tilde{c_e}^S$$
$$U_S[(T,T)^{00}] \geq U_S[(T,T)^{11}], U_S[(T,T)^{10}], U_S[(T,T)^{01}] \iff$$
$$\iff c_e \geq \tilde{c_e}^S$$

$s_S = (T,T)^{11}$ vs. $s_S = (T,R)^{e_S^l e_S^h}$:

The only difference between $s_S = (T,T)^{e_S^l e_S^h}$ and $s_S = (T,R)^{e_S^l e_S^h}$ is whether the specialist accepts the treatment of h-types referred by the PCP. Thus, $s_S = (T,T)^{e_S^l e_S^h}$ is preferred whenever the continuous treatment of initial h-types is profitable for the specialist.

$$U_S[(T,T)^{11}] \geq U_S[(T,R)^{11}] \iff u_S^{h_P}[(T,T)^{11}] \geq u_S^{h_P}[(T,R)^{11}] \iff$$
$$u_S^{l_P}[(T,T)^{11}] \geq u_S^{l_P}[(T,R)^{11}] \iff$$
$$\gamma_S \geq c_S^h + c_e - \frac{\beta \Delta_{c_S}(1 - p_S^{hh} + \Delta_p)}{1 + \beta(1 - p_S^{ll} - p_S^{hh})} =: \gamma_S^{min}$$
$$U_S[(T,R)^{11}] \geq U_S[(T,R)^{00}], U_S[(T,R)^{10}], U_S[(T,R)^{01}] \iff c_e \leq \tilde{c_e}^S$$

Let us now turn to the proof for the uniqueness of the equilibrium. Consider the case with effort provision for both physicians. Conditions (4.35) is fulfilled strictly for $\epsilon > 0$. Furthermore, Condition (4.36) is fulfilled strictly as long as ϵ is small enough because $c_e < \tilde{c_e}^S$. Thus, $s_P = (T,R)^{1e_P^h}$ is the strict best response to $s_S = (T,T)^{11}$.

I will show that $s_S = (T,T)^{11}$ is the strict best response to $s_P = (T,T)^{e_l e_h}$ and $s_P = (R,R)^{e_l e_h}$ given $c_e < \tilde{c_e}^S$. Because $s_P = (T,R)^{1e_P^h}$ is the strict best response to $s_S = (T,T)^{11}$, the uniqueness result follows.

Let $s_P = (T,T)^{e_l e_h}$, then $s_S = (T,T)^{11}$ is the strict best response:

$$\begin{aligned}
&U_S[(T,T)^{11}] > U_S[(T,R)^{1e_S^h}] \iff \gamma_S > \gamma_S^{min} \underset{c_e<\tilde{c_e}^S}{\Longrightarrow} \\
&\qquad \gamma_S > c_S^h + c_e + \frac{c_e(1-\beta)}{\beta(1-p_S^{ll})} - \frac{\Delta_p(c_S^h - c_S^l)(1-\beta)}{(1-p_S^{ll})(1-\beta p_S^{hh})} \\
&\qquad\quad - \frac{\beta(c_S^h - c_S^l)(1-p_S^{hh})[1+\beta(\Delta_p - p_S^{hh})]}{(1-\beta p_S^{hh})[1+\beta(1-p_S^{ll}-p_S^{hh})]} \iff \\
&u_S^{l_S}[(T,T)^{11}] > u_S^{l_S}[(T,R)^{0e_S^h}] \iff \\
&u_S^{h_S}[(T,T)^{11}] > u_S^{h_S}[(T,R)^{0e_S^h}] \iff \\
&U_S[(T,T)^{11}] > U_S[(T,R)^{0e_S^h}]
\end{aligned}$$

Furthermore,

$$\gamma_S > \gamma_S^{min} \implies U_S[(T,T)^{11}] > U_S[(R,R)^{e_S^l e_S^h}]$$

since

$$u_S^{l_S}[(T,T)^{11}] > u_S^{h_S}[(T,T)^{11}] > 0.$$

$s_S = (R,T)^{e_l e_h}$ is never a best response as it is dominated by $s_S = (T,T)^{e_l e_h}$ and $s_S = (R,R)^{e_l e_h}$. I have already proved that $s_P = (T,R)^{1e_P^h}$ is a best response to $s_S = (T,T)^{11}$. If the conditions hold strictly, it is the strict best response. The same holds true in the case without effort. Thus, there exists no equilibrium with $s_P = (T,T)^{e_l e_h}$.

Let $s_P = (R,R)^{e_l e_h}$, then $s_S = (T,T)^{e_l e_h}$ is the only response that does not lead to hospital treatment. Otherwise, both physicians receive a payoff of zero. As the payoff of $s_S = (T,T)^{11}$ is positive for $c_e < \tilde{c_e}^S$, it is the strict best response. Again, $s_P = (T,R)^{1e_P^h}$ is the strict best response to $s_S = (T,T)^{11}$. Thus, there exists no equilibrium with $s_P = (R,R)^{e_l e_h}$.

$s_P = (R,T)^{e_l e_h}$ is never an equilibrium as it is dominated by $s_P = (T,T)^{e_l e_h}$ and $s_P = (R,R)^{e_l e_h}$.

The proof for the outcome without effort follows along the same lines. □

4.A.5

Proof for Proposition 4 (first-best implementation in the team):

Proof. To show: $U_T(PS^{11}) := U_P[PS^{11}] + U_S[PS^{11}] \geq U_T(s)$
$\forall s \in \{PS^{e_P e_S}, PP^{e_P e_S}, SS^{e_P e_S}, SP^{e_P e_S}\}$

Optimal effort choice:

First, I consider the condition under which effort is provided in the team given that PCPs treat l-types and specialists treat h-types. The system of equations for the continuation profit of $PS^{e_P e_S}$ in all states $x \in \mathcal{X}$ are given by:

$$\begin{aligned}
u_T^{l_P} &= \gamma_P - c_P^l - e_P c_e + \beta[(p_P^{ll} + e_P \Delta_p) u_T^{l_P} + (1 - p_P^{ll} - e_P \Delta_p) u_T^{h_S}] \\
u_T^{l_S} &= u_T^{l_P} \\
u_T^{h_P} &= u_T^{h_S} \\
u_T^{h_S} &= \gamma_S - c_S^h - e_S c_e + \beta[(p_S^{hh} - e_S \Delta_p) u_T^{h_S} + (1 - p_S^{hh} + e_S \Delta_p) u_T^{l_P}]
\end{aligned}$$

Continuation profits are given by:

$$\begin{aligned}
u_T^l[PS^{11}] &= \frac{\gamma_P - c_P^l - c_e}{1-\beta} + \frac{\beta(1 - p_P^{ll} - \Delta_p)(\gamma_S - \gamma_P - c_S^h + c_P^l)}{(1-\beta)(1 + \beta[1 - p_P^{ll} - p_S^{hh}])} \\
u_T^h[PS^{11}] &= \frac{\gamma_P - c_P^l - c_e}{1-\beta} + \frac{(1 - \beta[p_P^{ll} + \Delta_p])(\gamma_S - \gamma_P - c_S^h + c_P^l)}{(1-\beta)(1 + \beta[1 - p_P^{ll} - p_S^{hh}])} \\
u_T^l[PS^{10}] &= \frac{\gamma_S - c_S^h}{1-\beta} - \frac{(1 - \beta p_S^{hh})(\gamma_S - \gamma_P - c_S^h + c_P^l + c_e)}{(1-\beta)(1 + \beta[1 - p_P^{ll} - p_S^{hh} + \Delta_p])} \\
u_T^h[PS^{10}] &= \frac{\gamma_S - c_S^h}{1-\beta} - \frac{\beta(1 - p_S^{hh})(\gamma_S - \gamma_P - c_S^h + c_P^l + c_e)}{(1-\beta)(1 + \beta[1 - p_P^{ll} - p_S^{hh} + \Delta_p])} \\
u_T^l[PS^{01}] &= \frac{\gamma_P - c_P^l}{1-\beta} + \frac{\beta(1 - p_P^{ll})(\gamma_S - \gamma_P - c_S^h + c_P^l - c_e)}{(1-\beta)(1 + \beta[1 - p_P^{ll} - p_S^{hh} + \Delta_p])} \\
u_T^h[PS^{01}] &= \frac{\gamma_P - c_P^l}{1-\beta} + \frac{(1 - \beta p_P^{ll})(\gamma_S - \gamma_P - c_S^h + c_P^l - c_e)}{(1-\beta)(1 + \beta[1 - p_P^{ll} - p_S^{hh} + \Delta_p])} \\
u_T^l[PS^{00}] &= \frac{\gamma_P - c_P^l}{1-\beta} + \frac{\beta(1 - p_P^{ll})(\gamma_S - \gamma_P - c_S^h + c_P^l)}{(1-\beta)(1 + \beta[1 - p_P^{ll} - p_S^{hh}])} \\
u_T^h[PS^{00}] &= \frac{\gamma_P - c_P^l}{1-\beta} + \frac{(1 - \beta p_P^{ll})(\gamma_S - \gamma_P - c_S^h + c_P^l)}{(1-\beta)(1 + \beta[1 - p_P^{ll} - p_S^{hh}])}
\end{aligned}$$

In the team, one physician always treats all patients of the same type. Thus, continuation profits for l^P and l^S as well as h^P and h^S are identical.

$$U_T(PS^{11}) \geq U_T(PS^{10}), U_T(PS^{01}), U_T(PS^{00}) \iff$$
$$u^l_T[PS^{11}] \geq u^l_T[PS^{10}], u^l_T[PS^{01}], u^l_T[PS^{00}] \iff$$
$$u^h_T[PS^{11}] \geq u^h_T[PS^{10}], u^h_T[PS^{01}], u^h_T[PS^{00}] \iff$$
$$(\gamma_S - c^h_S) - (\gamma_P - c^l_P) \leq -\frac{c_e[1+\beta(1-p^{ll}_P - p^{hh}_S)]}{\beta\Delta_p}$$

PS vs PP:

For PP, continuation profits are given by:

$$u^l_T[PP^{11e^l_S e^h_S}] = \frac{\gamma_P - c^l_P - c_e}{1-\beta} - \frac{\beta(1-p^{ll}_P - \Delta_p)(c^h_P - c^l_P)}{(1-\beta)(1+\beta[1-p^{ll}_P - p^{hh}_P])}$$
$$u^h_T[PP^{11e^l_S e^h_S}] = \frac{\gamma_P - c^l_P - c_e}{1-\beta} - \frac{(1-\beta[p^{ll}_P + \Delta_p])(c^h_P - c^l_P)}{(1-\beta)(1+\beta[1-p^{ll}_P - p^{hh}_P])}$$
$$u^l_T[PP^{10e^l_S e^h_S}] = \frac{\gamma_P - c^h_P}{1-\beta} + \frac{(1-\beta p^{hh}_P)(c^h_P - c^l_P - c_e)}{(1-\beta)(1+\beta[1-p^{ll}_P - p^{hh}_P - \Delta_p])}$$
$$u^h_T[PP^{10e^l_S e^h_S}] = \frac{\gamma_P - c^h_P}{1-\beta} + \frac{\beta(1-p^{hh}_P)(c^h_P - c^l_P - c_e)}{(1-\beta)(1+\beta[1-p^{ll}_P - p^{hh}_P - \Delta_p])}$$
$$u^l_T[PP^{01e^l_S e^h_S}] = \frac{\gamma_P - c^l_P}{1-\beta} - \frac{\beta(1-p^{ll}_P)(c^h_P - c^l_P + c_e)}{(1-\beta)(1+\beta[1-p^{ll}_P - p^{hh}_P + \Delta_p])}$$
$$u^h_T[PP^{01e^l_S e^h_S}] = \frac{\gamma_P - c^l_P}{1-\beta} - \frac{(1-\beta p^{ll}_P)(c^h_P - c^l_P + c_e)}{(1-\beta)(1+\beta[1-p^{ll}_P - p^{hh}_P + \Delta_p])}$$
$$u^l_T[PP^{00e^l_S e^h_S}] = \frac{\gamma_P - c^l_P}{1-\beta} - \frac{\beta(1-p^{ll}_P)(c^h_P - c^l_P)}{(1-\beta)(1+\beta[1-p^{ll}_P - p^{hh}_P])}$$
$$u^h_T[PP^{00e^l_S e^h_S}] = \frac{\gamma_P - c^l_P}{1-\beta} - \frac{(1-\beta p^{ll}_P)(c^h_P - c^l_P)}{(1-\beta)(1+\beta[1-p^{ll}_P - p^{hh}_P])}$$

It follows:

$$
\begin{aligned}
U_T(PP^{11e_S^le_S^h}) \geq U_T(PP^{10e_S^le_S^h}), U_T(PP^{01e_S^le_S^h}), U_T(PP^{00e_S^le_S^h}) &\iff \\
u_T^l[PP^{11e_S^le_S^h}] \geq u_T^l[PP^{10e_S^le_S^h}], u_T^l[PP^{01e_S^le_S^h}], u_T^l[PP^{00e_S^le_S^h}] &\iff \\
u_T^h[PP^{11e_S^le_S^h}] \geq u_T^h[PP^{10e_S^le_S^h}], u_T^h[PP^{01e_S^le_S^h}], u_T^h[PP^{00e_S^le_S^h}] &\iff \\
c_e \leq \tilde{c_e}^P & \\
U_T(PP^{00e_S^le_S^h}) \geq U_T(PP^{10e_S^le_S^h}), U_T(PP^{01e_S^le_S^h}), U_T(PP^{11e_S^le_S^h}) &\iff \\
u_T^l[PP^{00e_S^le_S^h}] \geq u_T^l[PP^{10e_S^le_S^h}], u_T^l[PP^{01e_S^le_S^h}], u_T^l[PP^{11e_S^le_S^h}] &\iff \\
u_T^h[PP^{00e_S^le_S^h}] \geq u_T^h[PP^{10e_S^le_S^h}], u_T^h[PP^{01e_S^le_S^h}], u_T^h[PP^{11e_S^le_S^h}] &\iff \\
c_e \geq \tilde{c_e}^P & \\
U_T(PS^{11e_S^le_S^h}) \geq U_T(PP^{11e_S^le_S^h}) &\iff \\
U_T(PS^{00e_S^le_S^h}) \geq U_T(PP^{00e_S^le_S^h}) &\iff \\
u_T^l[PS^{11e_S^le_S^h}] \geq u_T^l[PP^{11e_S^le_S^h}] &\iff \\
u_T^l[PS^{00e_S^le_S^h}] \geq u_T^l[PP^{00e_S^le_S^h}] &\iff \\
u_T^h[PS^{11e_S^le_S^h}] \geq u_T^h[PP^{11e_S^le_S^h}] &\iff \\
u_T^h[PS^{00e_S^le_S^h}] \geq u_T^h[PP^{00e_S^le_S^h}] &\iff \\
(\gamma_S - c_S^h) - (\gamma_P - c_P^l) \geq \frac{-[1+\beta(1-p_P^{ll}-p_S^{hh})]\Delta_{c_P}}{1+\beta(1-p_P^{ll}-p_P^{hh})} &
\end{aligned}
$$

For all compared strategies, differences in continuation profits are identical for a fixed state. Therefore, I omit continuation profits in the following.

PS vs SS:

For SS the continuation profits of the team are identical to continuation profits for PP with all indices P switched so S.

$$U_T(SS^{e_P^l e_P^h 11}) \geq U_T(SS^{e_P^l e_P^h 10}), U_T(SS^{e_P^l e_P^h 01}), U_T(SS^{e_P^l e_P^h 00}) \iff$$
$$c_e \leq \tilde{c_e}^S$$
$$U_T(SS^{e_P^l e_P^h 00}) \geq U_T(SS^{e_P^l e_P^h 10}), U_T(SS^{e_P^l e_P^h 01}), U_T(SS^{e_P^l e_P^h 11}) \iff$$
$$c_e \geq \tilde{c_e}^S$$
$$U_T(PS^{11}) \geq U_T(SS^{e_P^l e_P^h 11}) \iff$$
$$U_T(PS^{00}) \geq U_T(SS^{e_P^l e_P^h 00}) \iff$$
$$(\gamma_S - c_S^h) - (\gamma_P - c_P^l) \leq -\frac{[1 + \beta(1 - p_P^{ll} - p_S^{hh})]\Delta_{c_S}}{1 + \beta(1 - p_S^{ll} - p_S^{hh})}$$

PS vs SP:

For SP the continuation profits of the team are identical to continuation profits for PS with all indices P switched so S and indices S switched so P.

$$U_T(SP^{11}) \geq U_T(SP^{10}), U_T(SP^{01}), U_T(SP^{00}) \iff$$
$$c_e \leq \frac{\Delta_p \beta(c_P^h - c_S^l - \gamma_P + \gamma_S)}{1 + \beta(1 - p_S^{ll} - p_P^{hh})}$$
$$U_T(SP^{00}) \geq U_T(SP^{10}), U_T(SP^{01}), U_T(SP^{11}) \iff$$
$$c_e \geq \frac{\Delta_p \beta(c_P^h - c_S^l - \gamma_P + \gamma_S)}{1 + \beta(1 - p_S^{ll} - p_P^{hh})}$$
$$U_T(SS^{e_P^l e_P^h 11}) \geq U_T(SP^{11}) \iff$$
$$U_T(SS^{e_P^l e_P^h 00}) \geq U_T(SP^{00}) \iff$$
$$(\gamma_S - c_S^h) - (\gamma_P - c_P^h) \geq \frac{\beta \Delta_{c_S}(p_S^{hh} - p_P^{hh})}{1 + \beta(1 - p_S^{ll} - p_S^{hh})}$$
$$U_T(PP^{11 e_S^l e_S^h}) \geq U_T(SP^{11}) \iff$$
$$U_T(PP^{00 e_S^l e_S^h}) \geq U_T(SP^{00}) \iff$$
$$(\gamma_S - c_S^l) - (\gamma_P - c_P^l) \leq \frac{\beta \Delta_{c_P}(p_P^{ll} - p_S^{ll})}{1 + \beta(1 - p_P^{ll} - p_P^{hh})}$$

Thus, SP is dominated if

$$\Delta_{c_P} \geq \frac{\Delta_{c_S}[1 + \beta(1 - p_P^{ll} - p_P^{hh})]}{1 + \beta(1 - p_S^{ll} - p_S^{hh})}$$

For outcome PS^{11} to be implemented with smallest possible payment fees, Condition (4.37) must hold in order for the team to make non-negative profits for both patient types.

$$\begin{aligned} u_T^l(PS^{11}) \geq u_T^h(PS^{11}) = 0 \iff \\ \frac{\gamma_P - c_P^l - c_e}{1-\beta} + \frac{[1-\beta(p_P^{ll}+\Delta_p)(\gamma_S - \gamma_P + c_P^l - c_S^h)]}{(1-\beta)[1+\beta(1-p_P^{ll}-p_S^{hh})]} = 0 \end{aligned} \tag{4.37}$$

Solving for γ_P and inserting the binding Condition (4.37) into the binding Incentive Constraint (4.22), delivers

$$\gamma_S^* = c_S^h + c_e - \frac{\beta\Delta_{c_P}(1-p_S^{hh}+\Delta_p)}{1+\beta(1-p_P^{hh}-p_P^{ll})}.$$

Inserting γ_S^* into Condition (4.37) delivers

$$\gamma_P^* = c_P^l + c_e + \frac{\beta\Delta_{c_P}(1-p_P^{ll}-\Delta_p)}{1+\beta(1-p_P^{hh}-p_P^{ll})}.$$

Because Condition (4.22) is the only lower boundary on γ_S, whereas Conditions (4.20) and (4.21) are upper boundaries, this contract implements PS^{11} whenever this is possible.

Necessary conditions:

Condition (4.20) and (4.21) are both upper boundaries on γ_S. If $c_e \leq \tilde{c_e}^S$, Condition (4.21) is implied by Condition (4.20), otherwise Condition (4.20) is implied by Condition (4.21).

Let $c_e \leq \tilde{c_e}^S$. In this case FB^{11} can be implemented if and only if Condition (4.20) and Condition (4.22) can be fulfilled together, i.e.

$$\begin{aligned} -\frac{[1+\beta(1-p_P^{ll}-p_S^{hh})]\Delta_{c_S}}{1+\beta(1-p_S^{ll}-p_S^{hh})} \geq \frac{-[1+\beta(1-p_P^{ll}-p_S^{hh})]\Delta_{c_P}}{1+\beta(1-p_P^{ll}-p_P^{hh})} \iff \\ \Delta_{c_P} \geq \frac{\Delta_{c_S}[1+\beta(1-p_P^{ll}-p_P^{hh})]}{1+\beta(1-p_S^{ll}-p_S^{hh})}. \end{aligned}$$

Let $c_e \geq \tilde{c_e}^S$. In this case FB^{11} can be implemented if and only if Condition (4.21) and Condition (4.22) can be fulfilled together, i.e.

$$\begin{aligned} -\frac{c_e[1+\beta(1-p_P^{ll}-p_S^{hh})]}{\beta\Delta_p} \geq \frac{-[1+\beta(1-p_P^{ll}-p_S^{hh})]\Delta_{c_P}}{1+\beta(1-p_P^{ll}-p_P^{hh})} \iff \\ c_e \leq \tilde{c_e}^P. \end{aligned}$$

If these conditions hold strictly, γ_P^* can be decreased marginally without violating Conditions (4.20) and (4.21). In this case all incentive conditions hold strictly, and thus, the outcome is strictly preferred by the team.

□

4.A.6

Proposition 9. *Let $\Delta_{c_P} < \tilde{\Delta}_{c_P}$. For the team, the following statements hold.*

If $c_e < \tilde{c_e}^P$, PP^{1e_S} or SS^{e_P1} is second-best.

If $c_e > \tilde{c_e}^S$, PP^{0e_S} or SS^{e_P0} is second-best.

If $\tilde{c_e}^P < c_e < \tilde{c_e}^S$, PP^{0e_S}, SS^{e_P1}, or SP^{11} is second-best.

Proof. If the specialist has the relative cost-advantage for l-types ($\Delta_{c_P} < \tilde{\Delta}_{c_P}$), the treatment path PS can not be implemented in the team. In this case, PP^{1e_S} can be implemented for $c_e < \tilde{c_e}^P$, whereas SS^{e_P1} and SP^{11} can be implemented for $c_e < \tilde{c_e}^S$. If the conditions hold with flipped inequality, the respective outcomes can be implemented without effort provision. The proofs for these statements regarding the blind paths are identical to the solo practice case. The proof for path SP follows by exchanging indices S and P for the proof for PS. Outcomes are unique if the conditions hold strictly.

If $c_e < \tilde{c_e}^P$ or $c_e > \tilde{c_e}^S$, one of the blind treatment paths PP or SS is second-best optimal as the remaining path SP is weakly dominated by them for equal effort levels. However, if $\tilde{c_e}^P < c_e < \tilde{c_e}^S$, physicians provide more effort in path SP than in path PP. Thus, SP^{11} can be second-best optimal in this case. □

4.A.7

Proof. If PS^{11} can be implemented, Condition (4.23) provides a lower bound for Δ_{c_P}. If PS^{11} is first-best, Condition (4.5) also provides a lower bound on Δ_{c_P}. The latter condition is more strict if and only if

$$L \leq L^1.$$

Thus, if $L \leq L^1$, Condition (4.5) implies Condition (4.23). Conversely, if $L > L^1$, a λ that fulfills Condition (4.5) with equality as well as $c_e = 0$ and $p_P^{ll} = p_S^{ll}$, fulfills the first-best conditions but PS^{11} can not be implemented in the team because Condition (4.23) is not fulfilled.

Let us now consider the optimal effort. Effort can be incentivized in the team if Condition (4.24) is fulfilled. In terms of effort provision, PS^{11} is first-best if Condition (4.4) is fulfilled. Both conditions are upper boundaries on the cost of effort provision c_e. The latter condition is more strict if and only if

$$L \leq L^2.$$

Thus, if $L \leq L^2$, Condition (4.4) implies Condition (4.24). Conversely, if $L > L^2$, a λ that fulfills Condition (4.4) and (4.5) with equality as well as $p_P^{ll} = p_S^{ll}$, fulfills the first-best conditions but PS^{11} can not be implemented in the team because Condition (4.24) is not fulfilled.

□

4.A.8

Proposition 10. *Let* $\Delta_{c_P} < \tilde{\Delta}_{c_P}$.

(I) Let $c_e \leq \min(\tilde{c_e}^S, \tilde{c_e}^P)$. *There exist parameters such that* MS^{11} *is strictly superior to all other outcomes in both solo practices and the team.*

(II) Let $c_e \geq \max(\tilde{c_e}^S, \tilde{c_e}^P)$. *There exist parameters such that* MS^{00} *is strictly superior to all other outcomes in both solo practices and the team.*

(III) Let $\tilde{c_e}^P < c_e < \tilde{c_e}^S$. There exist parameters such that MS^{01} is strictly superior to all other outcomes in both solo practices and the team.

In order for either statement to be true, c_P^h must be sufficiently large.

(I) For $c_e \leq \min(\tilde{c_e}^S, \tilde{c_e}^P)$, there are, considering only unique equilibria, three possible second-best outcomes: MS^{11} (only solo practices), $SS^{e_P 1}$, and PP^{1e_S}. Given fixed efforts, MS always yields larger welfare than SS (see Lemma 3). Hence, for $c_e \leq \min(\tilde{c_e}^S, \tilde{c_e}^P)$ implementing MS^{11} is always preferred to $SS^{e_P 1}$.

Conditions for which MS^{11} is strictly superior to PP^{1e_S}:

$$\begin{aligned} & W^{l_P}_{MS^{11}} \geq W^{l_P}_{PP^{1e_S}} \iff W^{h_P}_{MS^{11}} \geq W^{h_P}_{PP^{1e_S}} \iff W^{h_S}_{MS^{11}} \geq W^{h_S}_{PP^{1e_S}} \iff \\ & c_P^h \geq \frac{\beta(1-p_P^{hh}+\Delta_p)([L+c_S^h][-1+\beta(p_S^{ll}+\Delta_p)] + c_P^l[1+\beta(1-p_S^{hh}-p_S^{ll})] + c_S^l r[-1+p_S^{hh}-\Delta_p])}{[-1+\beta(p_P^{ll}+\Delta_p)][1+\beta(1-p_S^{hh}-p_S^{ll})]} \\ & \quad - \frac{(L+c_S^l)\beta(1-p_S^{hh}+\Delta_p) + c_S^h(-1+\beta(p_S^{ll}+\Delta_p))}{[1+\beta(1-p_S^{hh}-p_S^{ll})]} \end{aligned} \tag{4.38}$$

$$\begin{aligned} & W^{l_S}_{MS^{11}} \geq W^{l_S}_{PP^{1e_S}} \iff \\ & c_P^h \geq \frac{L\beta[1-p_P^{ll}-\Delta_p] + c_P^l[1+\beta(\Delta_p - p_P^{hh})]}{\beta[-1+p_P^{ll}+\Delta_p]} \\ & \quad - \frac{\{1+\beta(1-p_P^{hh}-p_P^{ll})\}\{(L+c_S^h)\beta(-1+p_S^{ll}+\Delta_p) + c_S^l[-1+\beta(p_S^{hh}-\Delta_p)]\}}{[\beta(-1+p_P^{ll}+\Delta_p)][1+\beta(1-p_S^{ll}-p_S^{hh})]} \end{aligned} \tag{4.39}$$

Given Condition (4.5), Condition (4.39) is stricter than Condition (4.38). Thus, if Condition (4.39) is fulfilled, MS^{11} is strictly superior to PP^{1e_S}. Let us consider now whether parameters exist, such that this is true.

Inserting the supremum c_P^h implied by $\Delta_{c_P} < \tilde{\Delta}_{c_P}$ and the maximum c_P^l implied by Condition (4.6) delivers

$$\begin{aligned} W^{l_S}_{MS^{11}} &\geq W^{l_S}_{PP^{1e_S}} \iff \\ p_P^{hh} - p_S^{hh} &\geq p_S^{ll} - p_P^{ll}, \end{aligned} \tag{4.40}$$

which is strictly true by assumption. Thus, there exists a set of parameters such that MS^{11} is strictly superior to PP^{1e_S} and $SS^{e_P 1}$.

(II) For MS^{00}, an analogue argument applies. The condition for MS^{00} to be strictly superior to PP^{0e_S} is:

$$\begin{aligned} W^{l_S}_{MS^{00}} \geq W^{l_S}_{PP^{0e_S}} \iff & \\ c^h_P \geq & \frac{L\beta[1-p^{ll}_P] + c^l_P[1+\beta(-p^{hh}_P)]}{\beta[-1+p^{ll}_P]} \\ & - \frac{\{1+\beta(1-p^{hh}_P-p^{ll}_P)\}\{(L+c^h_S)\beta(-1+p^{ll}_S+) + c^l_S[-1+\beta(p^{hh}_S)]\}}{[\beta(-1+p^{ll}_P)][1+\beta(1-p^{ll}_S-p^{hh}_S)]} \end{aligned}$$

To illustrate the result, Figure 4.A.1 depicts a simulation that shows which outcome is strictly superior to the other outcome for $c_e < \tilde{c_e}^P$ and $\tilde{c_e}^S < c_e$ depending on feasible parameters c_e and c^h_P. Clearly, there exist feasible parameters such that either outcome can be optimal.

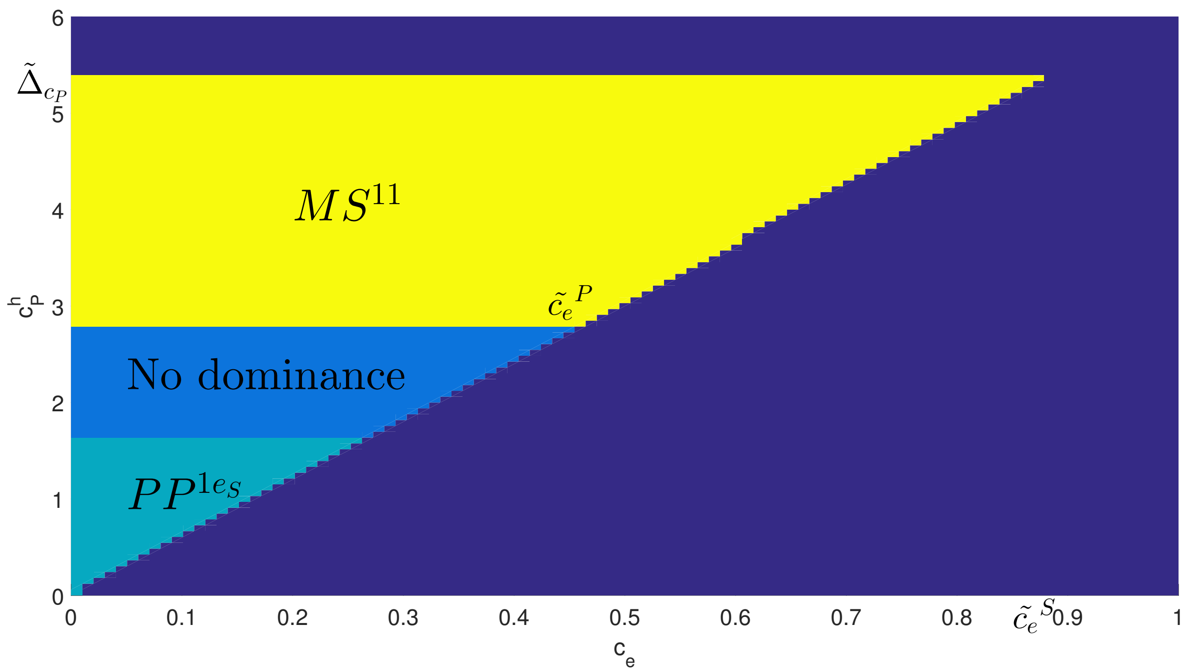
(III) Let us now turn to the case $\tilde{c_e}^P < c_e < \tilde{c_e}^S$. In this case, there are, four possible second-best outcomes: MS^{01} (only solo practices), SS^{e_P1}, SP^{11} (only team) and PP^{1e_S}

There is a trade-off between always receiving specialist effort in outcome SS^{e_P1} and better referral efficiency in outcome MS^{01}. The payer prefers MS^{01} over SS^{e_P1} if and only if the additional costs of specialist treatment for l-types is larger than its expected benefit:

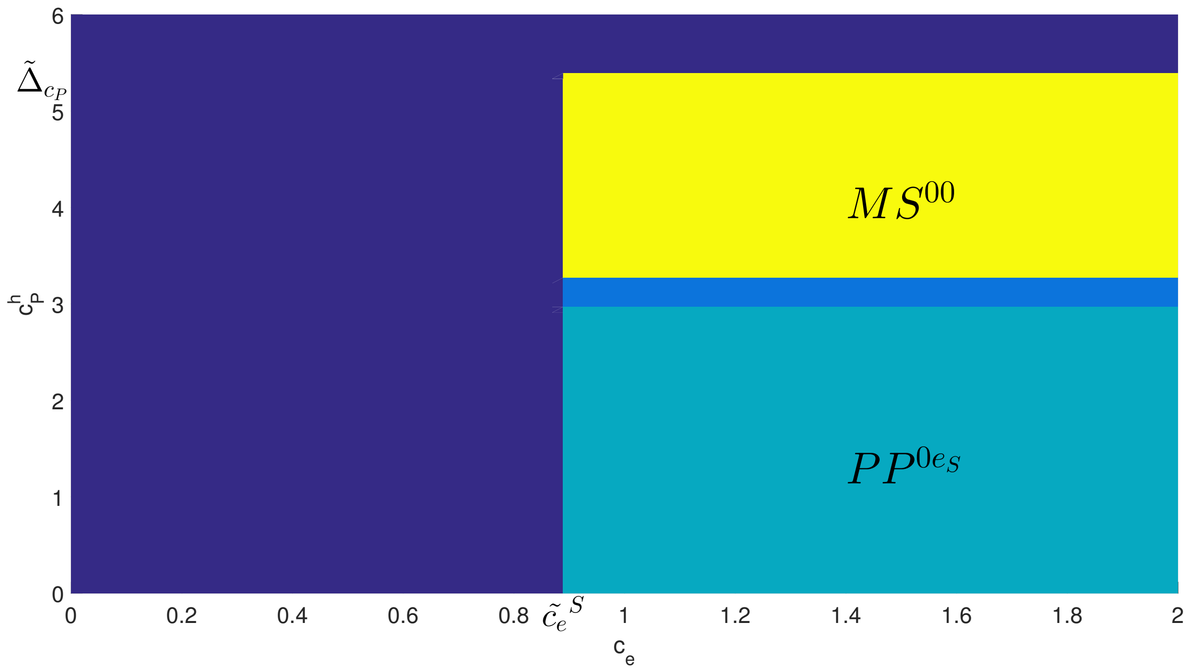
$$c_e - c^l_P + c^l_S \geq \frac{\beta(L+c^h_S-c^l_S)(\Delta_p+p^{ll}_S-p^{ll}_P)}{1+r(1-p^{ll}_S-p^{hh}_S)} \tag{4.41}$$

Let us now turn to the comparison of MS^{01} and PP^{0e_S}.

$$\begin{aligned} W^{l_S}_{MS^{01}} \geq W^{l_S}_{PP^{0e_S}} \iff & \\ c^h_P \geq & \frac{[1+\beta(1-p^{ll}_P-p^{hh}_P)](L+c^h_S)\beta[-1+\Delta_p+p^{ll}_S] + c_e[-1+\beta(-1+p^{ll}_S+p^{hh}_S)] + c^l_S[-1+r(-\Delta_p+p^{hh}_S)]}{[1+\beta(1-p^{ll}_S-p^{hh}_S)]\beta(p^{ll}_P-1)} \\ & + \frac{L\beta(1-p^{ll}_P) + c^l_P(1-\beta p^{hh}_P)}{\beta(p^{ll}_P-1)} \\ W^{l_P}_{MS^{01}} \geq W^{l_P}_{PP^{0e_S}} \iff & W^{h_P}_{MS^{01}} \geq W^{h_P}_{PP^{0e_S}} \iff W^{h_S}_{MS^{01}} \geq W^{h_S}_{PP^{0e_S}} \iff \\ c^h_P \geq & -\frac{(L-c^l_S)\beta[1+\Delta_p-p^{hh}_S] + c_e[-1+\beta(-1+p^{hh}_S+p^{ll}_S)] + c^h_S[-1+\beta(\Delta_p+p^{ll}_S)]}{1+\beta(1-p^{ll}_S-p^{hh}_S)} \\ & + \frac{\beta(p^{hh}_P-1)\{(L+c^h_S)[-1+\beta(\Delta_p+p^{ll}_S)] + c^l_P[1+\beta(1-p^{ll}_S-p^{hh}_S)] + c_e[-1+\beta(-1+p^{ll}_S+p^{hh}_S)]\}}{\beta(1-p^{ll}_P)[1+\beta(1-p^{ll}_S-p^{hh}_S)]} \end{aligned}$$



(a) $c_e < \tilde{c_e}^P, L = 45$



(b) $\tilde{c_e}^S < c_e, L = 25$. There exists no dominant outcome in the blue boundary region in the middle between optimal outcomes.

Figure 4.A.1: Simulation of the second-best outcome with $c_P^l = 0, p_P^{ll} = 0.8, p_P^{hh} = 0.6, c_S^l = 2, c_S^h = 10, p_S^{ll} = 0.8, p_S^{hh} = 0.3, \Delta_p = 0.1, r = 0.99$.

Let us now turn to the comparison of MS^{01} and SP^{11}.

$$W^{l_P}_{MS^{01}} \geq W^{l_P}_{SP^{11}} \iff$$
$$c^h_P \geq \frac{(L - c^l_P + c_e)(1-\beta) + c^h_S \beta(-1 + p^{ll}_P)}{p^{ll}_P - 1}$$
$$+ \frac{(\beta - 1)[1 + \beta(\Delta_p - p^{hh}_P)]\{(L + c^h_S)\beta(-1 + p^{ll}_P) + c^l_S[1 + \beta(1 + \Delta_p - p^{ll}_P - p^{hh}_S)] + (c_e - c^l_P)[1 + \beta(\Delta_p - p^{hh}_S)]\}}{\beta(p^{ll}_P - 1)(\Delta_p + p^{ll}_S - 1)[1 + \beta(\Delta_p - p^{hh}_S)]}$$
$$+ \frac{\beta^3(p^{hh}_P - p^{hh}_S)(p^{ll}_P - 1)(\Delta_p - p^{hh}_S + 1)(L + c^h_S - c^l_S)}{(1 - \beta p^{ll}_P)[1 + \beta(\Delta_p - p^{hh}_S)][1 + \beta(1 - p^{ll}_S - p^{hh}_S)]}$$
$$W^{l_S}_{MS^{01}} \geq W^{l_S}_{SP^{11}} \iff W^{h_P}_{MS^{01}} \geq W^{h_P}_{SP^{11}} \iff W^{h_S}_{MS^{01}} \geq W^{h_S}_{SP^{11}} \iff$$
$$c^h_P \geq c^h_S - \frac{\beta(p^{hh}_P - p^{hh}_S)(L + c^h_S - c^l_S)}{1 + \beta(1 - p^{ll}_S - p^{hh}_S)}$$

Figure 4.A.2 depicts a simulation that shows which outcome is strictly superior to all other outcomes for $\tilde{c_e}^P < c_e < \tilde{c_e}^S$ depending on feasible parameters c_e and c^h_P. Clearly, there exist feasible parameters such that either outcome can be optimal.

4.A.9

To show: If Condition (4.26) does not hold, treatment path PP dominates SP.

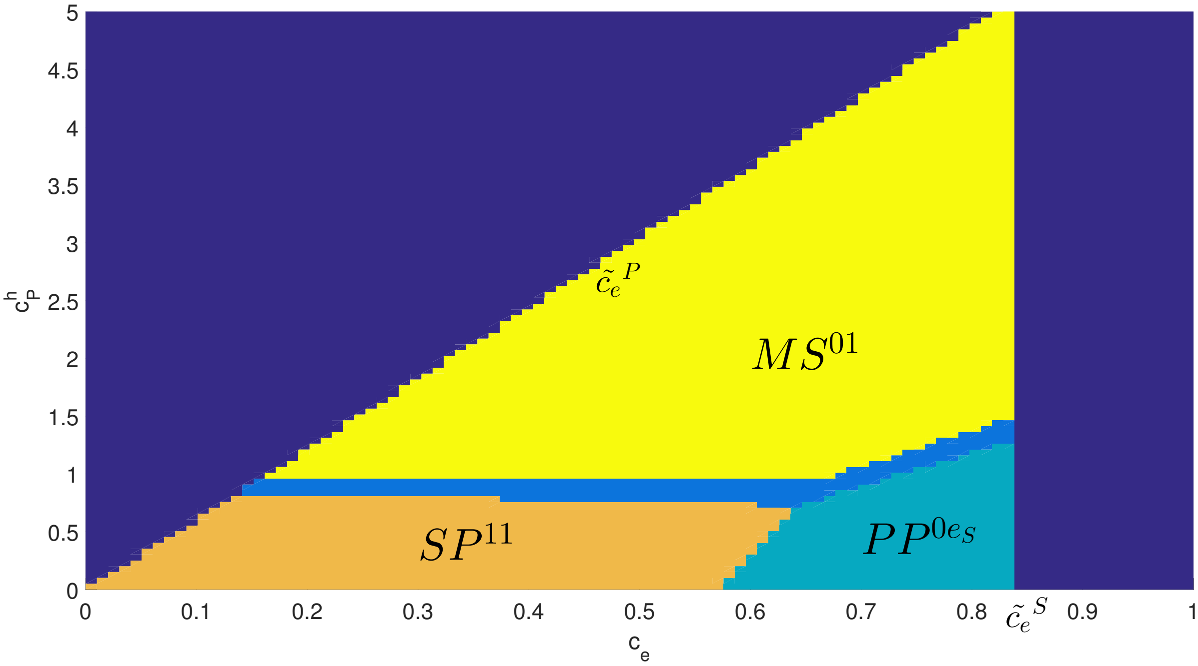
Proof.

$$\begin{aligned} U_T(PP^{1e_S}) \geq U_T(SP^{11}) \iff U_T(PP^{0e_S}) \geq U_T(SP^{00}) \iff \\ c^l_S \geq c^l_P - \frac{\beta(c^h_P - c^l_P)(p^{ll}_P - p^{ll}_S)}{1 + \beta(1 - p^{ll}_P - p^{hh}_P)} \end{aligned} \tag{4.42}$$

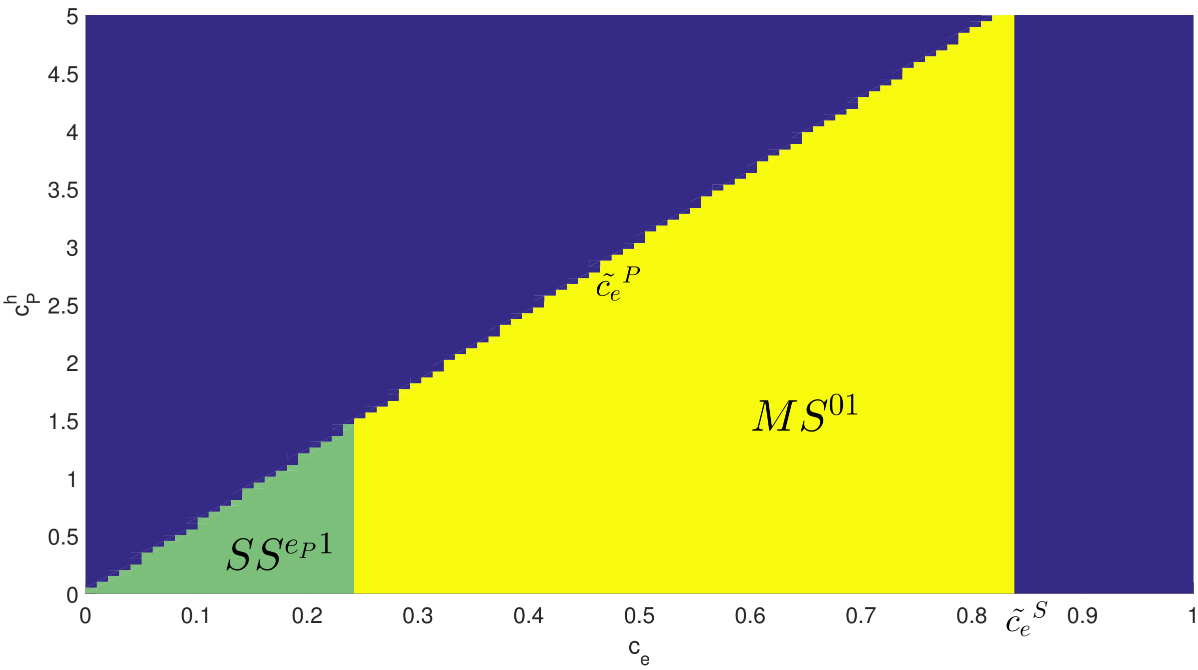
Clearly, this condition holds for $p^{ll}_P \geq p^{ll}_S$. Assume $p^{ll}_P < p^{ll}_S$. Then Condition (4.27) implies Condition (4.42) if and only if Condition (4.26) does not hold:

$$c^l_P - \frac{\beta(c^h_P - c^l_P)(p^{ll}_P - p^{ll}_S)}{1 + \beta(1 - p^{ll}_P - p^{hh}_P)} \leq c^l_P - \frac{\beta(c^h_S - c^l_P)(p^{ll}_P - p^{ll}_S)}{1 + \beta(1 - p^{ll}_P - p^{hh}_S)} \iff$$
$$c^h_S - c^h_P \geq \frac{(p^{hh}_P - p^{hh}_S)\beta(c^h_P - c^l_P)}{1 + \beta(1 - p^{ll}_P - p^{hh}_P)} \iff$$
$$U_T(PS^{11}) \leq U_T(PP^{1e_S})$$

Because Condition (4.27) is already implied by first-best Condition (4.6), PP dominates SP. □



(a) $L = 21$. There exists no dominant outcome in the blue boundary region in the middle between optimal outcomes.



(b) $L = 35$

Figure 4.A.2: Simulation of the second-best outcome for $\tilde{c}_e{}^P < c_e < \tilde{c}_e{}^S$ with $c_P^l = 0, p_P^{ll} = 0.8, p_P^{hh} = 0.6, c_S^l = 2, c_S^h = 10, p_S^{ll} = 0.75, p_S^{hh} = 0.3, \Delta_p = 0.1, r = 0.99$.

4.A.10

Proof that effort in solo practices does not exceed effort in the team, given that physicians are altruistic and PS is implementable in the team.

Proof. I will show the proof for the specialist. The proof for the PCP is analogous. For each possible treatment path (excluding paths with hospital treatment) it needs to hold that effort provision is impossible to implement for $c_e > \tilde{c}_e^{P\alpha}$ given that $\Delta_{c_P} \geq \tilde{\Delta}^{\alpha}_{c_P}$.

$PS : s_P = (T, R)^{e^l_P e^h_P}$

In order to implement PS, the following conditions need to hold.

$e^l_P = 1:$

$$U^{\alpha}_S[(R,T)^{e^l_S 1}] \geq U^{\alpha}_S[(R,T)^{e^l_S 0}] \iff$$
$$u^{x\alpha}_S[(R,T)^{e^l_S 1}] \geq u^{x\alpha}_S[(R,T)^{e^l_S 0}] \forall x \in \mathcal{X} \iff$$
$$\gamma_S \leq c^h_S + \alpha L - \frac{c_e[1+\beta(1-p^{ll}_P - p^{hh}_S - \Delta_p)]}{\Delta_p \beta}$$
$$U^{\alpha}_S[(R,T)^{e^l_S 1}] \geq U^{\alpha}_S[(R,R)^{e^l_S e^h_S}] \iff$$
$$u^{x\alpha}_S[(R,T)^{e^l_S 1}] \geq u^{x\alpha}_S[(R,R)^{e^l_S e^h_S}] \forall x \in \mathcal{X} \iff$$
$$\gamma_S \geq c^h_S - c_e - \frac{\alpha L \beta (p^{hh}_P - p^{hh}_S)}{1+\beta(1-p^{ll}_P - p^{hh}_P)}$$

$e^l_P = 0:$

$$U^{\alpha}_S[(R,T)^{e^l_S 1}] \geq U^{\alpha}_S[(R,T)^{e^l_S 0}] \iff$$
$$u^{x\alpha}_S[(R,T)^{e^l_S 1}] \geq u^{x\alpha}_S[(R,T)^{e^l_S 0}] \forall x \in \mathcal{X} \iff$$
$$\gamma_S \leq c^h_S + \alpha L - \frac{c_e[1+\beta(1-p^{ll}_P - p^{hh}_S)]}{\Delta_p \beta}$$
$$U^{\alpha}_S[(R,T)^{e^l_S 1}] \geq U^{\alpha}_S[(R,R)^{e^l_S e^h_S}] \iff$$
$$u^{x\alpha}_S[(R,T)^{e^l_S 1}] \geq u^{x\alpha}_S[(R,R)^{e^l_S e^h_S}] \forall x \in \mathcal{X} \iff$$
$$\gamma_S \geq c^h_S - c_e - \frac{\alpha L \beta (p^{hh}_P - p^{hh}_S + \Delta_p)}{1+\beta(1-p^{ll}_P - p^{hh}_P)}$$

Either set of conditions can be fulfilled if and only if

$$c_e \leq \frac{\Delta_p \alpha L \beta}{1+\beta(1-p_P^{ll}-p_P^{hh})} < \tilde{c}_e^{\,P\alpha}.$$

$SS/MS/SM$:

In all three paths, the specialist treats all received patients indefinitely. Effort is provided if and only if

$$\begin{aligned} U_S^\alpha[(T,T)^{11}] &\geq U_S^\alpha[(T,T)^{01}], U_S^\alpha[(T,T)^{10}], U_S^\alpha[(T,T)^{00}] \iff \\ u_S^{l_S\alpha}[(T,T)^{11}] &\geq u_S^{l_S\alpha}[(T,T)^{01}], u_S^{l_S\alpha}[(T,T)^{10}], u_S^{l_S\alpha}[(T,T)^{00}] \iff \\ u_S^{h_S\alpha}[(T,T)^{11}] &\geq u_S^{h_S\alpha}[(T,T)^{01}], u_S^{h_S\alpha}[(T,T)^{10}], u_S^{h_S\alpha}[(T,T)^{00}] \iff \\ c_e &\leq \tilde{c}_e^{\,S\alpha}. \end{aligned}$$

Furthermore,

$$\tilde{c}_e^{\,S\alpha} \leq \tilde{c}_e^{\,P\alpha} \iff \Delta_{c_P} \geq \tilde{\Delta}_{c_P}^\alpha.$$

$SP: s_P = (R,T)^{e_P^l e_P^h}$

$e_P^h = 1:$

$$\begin{aligned} U_S^\alpha[(T,R)^{1e_S^h}] &\geq U_S^\alpha[(T,R)^{0e_S^h}] \iff \\ u_S^{x\alpha}[(T,R)^{1e_S^h}] &\geq u_S^{x\alpha}[(T,R)^{0e_S^h}] \forall x \in \mathcal{X} \iff \\ \gamma_S &\geq c_S^l + c_e - \alpha L + \frac{c_e[1+\beta(1-p_S^{ll}-p_P^{hh}+\Delta_p)]}{\Delta_p \beta} \\ U_S^\alpha[(T,R)^{1e_S^h}] &\geq U_S^\alpha[(T,T)^{11}] \iff \\ u_S^{x\alpha}[(T,R)^{1e_S^h}] &\geq u_S^{x\alpha}[(T,T)^{11}] \forall x \in \mathcal{X} \iff \\ \gamma_S &\leq c_S^l + c_e - \alpha L + \frac{(\alpha L + c_S^h - c_S^l)1+\beta(1-p_S^{ll}-p_P^{hh})}{1+\beta(1-p_S^{ll}-p_S^{hh})} \end{aligned}$$

$e_P^h = 0:$

$$
\begin{aligned}
U_S^\alpha[(T,R)^{1e_S^h}] \geq U_S^\alpha[(T,R)^{0e_S^h}] &\iff \\
u_S^{x\alpha}[(T,R)^{1e_S^h}] \geq u_S^{x\alpha}[(T,R)^{0e_S^h}] \forall x \in \mathcal{X} &\iff \\
\gamma_S \geq c_S^l - \alpha L + \frac{c_e[1+\beta(1-p_S^{ll}-p_P^{hh}+\Delta_p)]}{\Delta_p \beta} & \\
U_S^\alpha[(T,R)^{1e_S^h}] \geq U_S^\alpha[(T,T)^{11}] &\iff \\
u_S^{x\alpha}[(T,R)^{1e_S^h}] \geq u_S^{x\alpha}[(T,T)^{11}] \forall x \in \mathcal{X} &\iff \\
\gamma_S \leq c_S^l + c_e - \alpha L + \frac{(\alpha L + c_S^h - c_S^l)1 + \beta(1-p_S^{ll}-p_P^{hh}-\Delta_p)}{1+\beta(1-p_S^{ll}-p_S^{hh})} &
\end{aligned}
$$

Either set of conditions can be fulfilled if and only if

$$c_e \leq \tilde{c}_e^{S\alpha}.$$

$PM : s_P = (T,T)^{e_P^l e_P^h}$

$e_P^l = e_P^h = 1:$

$$
\begin{aligned}
U_S^\alpha[(R,T)^{e_S^l 1}] \geq U_S^\alpha[(R,T)^{e_S^l 0}] &\iff \\
u_S^{h_S\alpha}[(R,T)^{e_S^l 1}] \geq u_S^{h_S\alpha}[(R,T)^{e_S^l 1}] &\iff \\
\gamma_S \leq c_S^h + \frac{\alpha L[1+\beta(\Delta_p - p_P^{hh})]}{1+\beta(1-p_P^{ll}-p_P^{hh})} - \frac{c_e(1-\beta p_S^{hh})}{\Delta_p \beta} & \\
U_S^\alpha[(R,T)^{e_S^l 1}] \geq U_S^\alpha[(R,R)^{e_S^l e_S^h}] &\iff \\
u_S^{h_S\alpha}[(R,T)^{e_S^l 1}] \geq u_S^{h_S\alpha}[(R,R)^{e_S^l e_S^h}] &\iff \\
u_S^{l_S\alpha}[(R,T)^{e_S^l 1}] \geq u_S^{l_S\alpha}[(R,R)^{e_S^l e_S^h}] &\iff \\
\gamma_S \geq c_S^h + c_e - \frac{\alpha L \beta(p_P^{hh}-p_S^{hh})}{1+\beta(1-p_P^{ll}-p_P^{hh})} &
\end{aligned}
$$

$e_P^l = e_P^h = 0:$

$$
\begin{aligned}
U_S^{\alpha}[(R,T)^{e_S^l 1}] \geq U_S^{\alpha}[(R,T)^{e_S^l 0}] &\iff \\
u_S^{h_S\alpha}[(R,T)^{e_S^l 1}] \geq u_S^{h_S\alpha}[(R,T)^{e_S^l 1}] &\iff \\
\gamma_S \leq c_S^h + \frac{\alpha L[1+\beta(p_P^{hh})]}{1+\beta(1-p_P^{ll}-p_P^{hh})} - \frac{c_e(1-\beta p_S^{hh})}{\Delta_p \beta} & \\
U_S^{\alpha}[(R,T)^{e_S^l 1}] \geq U_S^{\alpha}[(R,R)^{e_S^l e_S^h}] &\iff \\
u_S^{h_S\alpha}[(R,T)^{e_S^l 1}] \geq u_S^{h_S\alpha}[(R,R)^{e_S^l e_S^h}] &\iff \\
u_S^{l_S\alpha}[(R,T)^{e_S^l 1}] \geq u_S^{l_S\alpha}[(R,R)^{e_S^l e_S^h}] &\iff \\
\gamma_S \geq c_S^h + c_e - \frac{\alpha L\beta(\Delta_p + p_P^{hh} - p_S^{hh})}{1+\beta(1-p_P^{ll}-p_P^{hh})} &
\end{aligned}
$$

Either set of conditions can be fulfilled if and only if

$$
c_e \leq \frac{\Delta_p \alpha L \beta}{1+\beta(1-p_P^{ll}-p_P^{hh})} < \tilde{c}_e^{\,P\alpha}.
$$

$MP : s_P = (T,T)^{e_P^l e_P^h}$

$e_P^l = e_P^h = 1:$

$$
\begin{aligned}
U_S^{\alpha}[(T,R)^{1 e_S^h}] \geq U_S^{\alpha}[(T,R)^{0 e_S^h}] &\iff \\
u_S^{l_S\alpha}[(T,R)^{1 e_S^h}] \geq u_S^{l_S\alpha}[(T,R)^{0 e_S^h}] &\iff \\
\gamma_S \geq c_S^l + \frac{c_e(1-\beta p_S^{ll})}{\Delta_p \beta} - \frac{\alpha L[1-\beta(p_P^{ll}+\Delta_p)]}{1+\beta(1-p_P^{ll}-p_P^{hh})} & \\
U_S^{\alpha}[(T,R)^{1 e_S^h}] \geq U_S^{\alpha}[(T,T)^{11}] &\iff \\
u_S^{l_S\alpha}[(T,R)^{1 e_S^h}] \geq u_S^{l_S\alpha}[(T,T)^{11}] &\iff \\
u_S^{h_S\alpha}[(T,R)^{1 e_S^h}] \geq u_S^{h_S\alpha}[(T,T)^{11}] &\iff \\
\gamma_S \leq \frac{(c_S^h + c_e)[1+\beta(1-p_P^{hh}-p_P^{ll})] + \alpha L\beta(1+\Delta_p - p_P^{hh})}{1+\beta(1-p_P^{ll}-p_P^{hh})} & \\
- \frac{\beta(1+\Delta_p - p_S^{hh})(c_S^h - c_S^l + \alpha L)}{1+\beta(1-p_S^{ll}-p_S^{hh})} &
\end{aligned}
$$

$e^l_P = e^h_P = 0:$

$$
\begin{aligned}
U^{\alpha}_S[(T,R)^{1e^h_S}] \geq U^{\alpha}_S[(T,R)^{0e^h_S}] &\iff \\
u^{l_S\alpha}_S[(T,R)^{1e^h_S}] \geq u^{l_S\alpha}_S[(T,R)^{0e^h_S}] &\iff \\
\gamma_S &\geq c^l_S + \frac{c_e(1-\beta p^{ll}_S)}{\Delta_p\beta} - \frac{\alpha L[1-\beta(p^{ll}_P)]}{1+\beta(1-p^{ll}_P-p^{hh}_P)} \\
U^{\alpha}_S[(T,R)^{1e^h_S}] \geq U^{\alpha}_S[(T,T)^{11}] &\iff \\
u^{l_S\alpha}_S[(T,R)^{1e^h_S}] \geq u^{l_S\alpha}_S[(T,T)^{11}] &\iff \\
u^{h_S\alpha}_S[(T,R)^{1e^h_S}] \geq u^{h_S\alpha}_S[(T,T)^{11}] &\iff \\
\gamma_S &\leq \frac{(c^h_S+c_e)[1+\beta(1-p^{hh}_P-p^{ll}_P)]+\alpha L\beta(1-p^{hh}_P)}{1+\beta(1-p^{ll}_P-p^{hh}_P)} \\
&\quad - \frac{\beta(1+\Delta_p-p^{hh}_S)(c^h_S-c^l_S+\alpha L)}{1+\beta(1-p^{ll}_S-p^{hh}_S)}
\end{aligned}
$$

Either set of conditions can be fulfilled if and only if

$$c_e \leq \tilde{c}_e^{\,S\alpha}.$$

□

5 Rewards for Information Provision in Patient Referrals: A Theoretical Model and an Experimental Test

Jeannette Brosig-Koch, Malte Griebenow, Mathias Kifmann, and Franziska Then

Abstract

We study whether bonus payments for information provision can improve the information flow between physicians. A primary care physician (PCP) decides on the provision of information of varying qualities to a specialist while referring a patient. Our theoretical model, which includes altruism and loss aversion, predicts that bonus payments increase the provision of both high- and low-quality information. Running a controlled laboratory experiment we find support for this prediction. If the beneficiary of information provision receives a higher payoff than the PCP, we observe that PCPs more often pass on high-quality information when the beneficiary is a patient. If the beneficiary receives a lower payoff than the PCP, the type of the beneficiary (specialist or patient) does not affect the provision of high-quality information.

5.1 Introduction

When patients are referred from one physician to another, the provision of information by the referring physician is important for the optimal care of the patient. The referring physician has already made an assessment of the patient's health and this information can help to treat the patient faster or to reduce the costs of the receiving physician. For example, the referring physician may know the patient's relevant medical history, which is useful in diagnosing the patient more accurately and making faster treatment decisions.

The information flow between physicians, however, does not seem to be optimal. In a literature review, Mehrotra et al., 2011 find that many referrals do not include a transfer of information and when they do, the data is often insufficient for medical decision making (see also Bodenheimer, 2008). As a potential solution to this problem, Bodenheimer proposes that care coordination tasks should be financially rewarded. For example, Medicare has started paying physicians for care coordination services of chronically ill patients in 2015 (Centers for Medicare & Medicaid Services, 2014).

In this paper, we present a model that considers both whether information is provided and, if information is provided, the quality of information. Information may be of low or high quality. High-quality information corresponds to a clear and complete report about the patient's medical history and diagnostic results. Low-quality information corresponds to unclear or incomplete information. Usually, it is prohibitively costly for the payment authority in charge of remunerating health care services (the payer) to verify the quality of information provided by the physicians. Therefore, agency problems exist as the quality of information is private information of the physicians.

Our model allows physicians to be partially altruistic towards their patients and fellow physicians. We further consider that some patients may benefit more from information provision than others. For example, for some patients treatment may be more urgent than for others and the information provided may speed up the treatment process.

We consider a bonus payment for information provision as a remedy for the under-provision of information. According to our model, the payment required to maximize the amount of provided high-quality information is a fraction of the costs of information provision. Increasing payments beyond this threshold only leads to the provision of more low-quality information. Finally, increasing the payment beyond the costs for low-quality information does not additionally change physician behavior.

Besides the standard approach described before, we also include a non-linear version of our model that uses an S-shaped value function in the physician's utility function. This version is inspired by previous behavioral research suggesting that individuals evaluate monetary losses and gains from a reference point differently (see, e.g. Kuehberger, 1998; Bleichrodt et al., 2001; Abdellaoui et al., 2007). Although prospect theory has been initially suggested for risky contexts to explain such reference-dependent preferences (Kahneman and Tversky, 1979), Tversky and Kahneman (1991) apply its properties also to riskless decisions like those in our experiment (see, e.g., Gaechter et al., 2010, for experimental evidence). Here, physicians may also decide depending on a reference point because they can be faced with a decision between a monetary gain (providing low-quality information) and a monetary loss (providing high-quality information) for some levels of bonus payments. In contrast to the linear version, our non-linear model suggests that the amount of high-quality information provision can decrease with an increase in the payment for information provision. Furthermore, if the payments are increased beyond the costs of high-quality information provision, the effect reverses and more high-quality information is provided. This latter effect is caused primarily by reference-dependent loss aversion.

We test the theoretical predictions derived from our two models in a laboratory experiment. Using a controlled lab environment that allows ceteris paribus variations of bonus payments and benefits from information provision, we are able to establish the causal link between payments and the provision of information (for a general discussion of the experimental method see, e.g., Frechette and Schotter, 2015, and for a discussion of health economics experiments see Galizzi and Wiesen, 2018).[1] In our experiment, subjects in the role of primary care physicians (PCPs) decide on passing on information to subjects in the role of specialists while referring a patient. In all experimental conditions, we use a within-subject design and systematically

[1]For a first laboratory experiment on patient referrals see Waibel and Wiesen (2021). Waibel and Wiesen (2021) do not focus on the provision of information in patient referrals, though.

vary the bonus payment as well as the benefit from information provision at the subject level to test the influence of both parameters. In our baseline condition, we focus on a situation in which the patient benefits from information provision and in which the PCP always earns less money than the specialist. The latter is typically the case in most OECD countries (OECD, 2015). Note that, in this baseline condition, bonus payments for information provision are implemented in addition to a fixed capitation payment. This additional payment may not be cost-effective from the payer's perspective, though. Accordingly, we implement a second experimental condition to test the behavioral effects of a cost-neutral version of the bonus payments.

With our remaining three experimental conditions we check the robustness of our behavioral results. First, we change the beneficiary of information provision (specialist instead of patient). Second, we alter the payoff relation between the PCP and the beneficiary. While the PCP always earns less than the beneficiary from information provision in the previous three conditions, we implement two additional conditions in which the PCP always earns more than the beneficiary (patient or specialist). Note that, without additional assumptions, the predictions derived from our two models do not imply a behavioral impact of these changes (cost neutrality, beneficiary, and payoff relation). Still, there is behavioral evidence suggesting such an impact (see the finding by Kesternich et al., 2015, on the role of the receiver type in medically framed distribution and cost dispersion games or the research on inequality aversion by, e.g., Fehr and Schmidt, 1999; G. E. Bolton and Ockenfels, 2000).

As predicted by our model, PCPs in the experiment pass on more low- and high-quality information as the bonus payment increases. If the bonus payment is at least as high as the costs for the provision of high-quality information, PCPs provide less low-quality information and more high-quality information than in decision tasks with lower bonus payments. This behavioral pattern is in line with the non-linear version of our model considering an aversion to lose profit relative to a reference profit in addition to altruism. Moreover, we observe that PCPs' reactions to increases in the bonus payment are similar regardless of whether the bonus payment is introduced cost-neutrally or not. Accordingly, taking the perspective of a payer who trades off the benefit of information provision against his payments, we find that bonus payments exceeding the cost of low-quality information are only efficient if the bonus payment is introduced cost neutrally.

Checking the robustness of our results with regard to the type of the receiver and the payoff relation between the PCP and the beneficiary, we find similar behavioral responses to the bonus payment and the benefit of information provision across all conditions. Still, PCPs tend to make more selfish decisions and provide less high-quality information when two conditions are met – first, the specialist instead of the patient benefits from information provision and second, the PCP earns less than the specialist.

In sum, our study is the first to provide a model that allows to describe the effect of different bonus payments on information provision between physicians. In addition, using the controlled environment of a laboratory, we are able to test the predictions of this model and to check the robustness of results. Our findings contribute to the understanding of the conditions under which information provision between physicians can be effective.

5.2 Model

5.2.1 Benefits and costs of information provision

Two physicians, a primary care physician (PCP, "he") and a specialist ("she"), are tasked by a health care payment authority (payer) with the treatment of a patient. The PCP has determined that he is unable to help the patient and, therefore, refers the patient to the specialist for receiving treatment. In order to ease the transition process, the PCP can provide information on the patient's medical history and the result of his diagnostic testing to the specialist. He can pass on no information ($I = 0$), information of low quality ($I = l$), or information of high quality ($I = h$). The benefits from information provision may accrue to the patient (extra benefit b) or to the specialist (cost savings s). Specifically, the patient receives benefit $B = \bar{b} + b$ including a base benefit of $\bar{b}$ and the specialist accrues costs of $c_S = \bar{c}_S - s$ with maximum costs $\bar{c}_S$ from specialist treatment. In the former case, the patient can experience a better diagnosis and treatment procedure which increases her benefit. For example, the patient may not have to reproduce her medical history and therefore may save

Table 5.2.1: Benefits or savings and costs of information provision

	0	l	h
b or s	0	$\kappa/3$	κ
c	0	$\mu/2$	μ

time. In the latter case, the specialist can provide the same diagnosis and treatment as without the PCP's information but at lower costs. For example, the specialist may not have to repeat diagnostic procedures.

To what extent information provision yields a benefit is likely to differ between cases. For example, in some cases fast treatment is necessary and information provision speeds up the treatment process. Also, diagnostic procedures can differ in time and costs. Furthermore, the benefit of the information depends on its quality. A carefully worded, compact, and complete report of all the relevant information is of greater use to the treating specialist but consumes more of the PCP's time and effort. The benefit from high-quality information provision is measured by $\kappa \geq 0$. If low quality is provided, we assume that the benefit is reduced to one third, i.e., to $\kappa/3$. The cost of information is the same for all patients and corresponds to μ in case of high-quality information and $\mu/2$ for low-quality information (see Table 5.2.1). These assumptions ensure that the provision of high-quality information is more cost-efficient than low-quality information provision. This reflects the empirical finding that information quality is often insufficient for medical decision making (Mehrotra et al., 2011).

Based on this general setting, we derive propositions for physician behavior in the following Subsections. We consider a linear utility function for the PCP in Subsection 5.2.2. For an alternative version of the model, we take aversion to losses from a reference profit into account and allow for a non-linear utility function in Subsection 5.2.3.

5.2.2 Provider behavior – Linear PCP utility

We first formulate a simple, linear model about physician behavior based on the standard model of altruistic physicians (Ellis and T. G. McGuire, 1986) and the assumption that the degree of altruism does not depend on the beneficiary of information provision. The PCP considers the benefit κ in the decision to provide information $I \in \{0, l, h\}$ with the altruism factor $\beta \geq 0$.

A PCP receives a capitation payment F_{PCP} upon seeing a patient. On top, a bonus payment γ is made to the PCP if he sends information. That is, overall payment for the PCP is

$$p(I) := \begin{cases} F_{PCP}, & I = 0 \\ F_{PCP} + \gamma, & I \neq 0 \end{cases}. \tag{5.1}$$

Capitations are set to ensure that the physicians' costs are always covered. We assume that the quality of information is not verifiable. Therefore, the PCP receives the bonus payment regardless of provided quality. If the patient benefits from information provision, the PCP's utility is given by

$$U(I) = p(I) - c(I) + \beta b(I). \tag{5.2}$$

If the specialist benefits from information provision, we have

$$U(I) = p(I) - c(I) + \beta s(I). \tag{5.3}$$

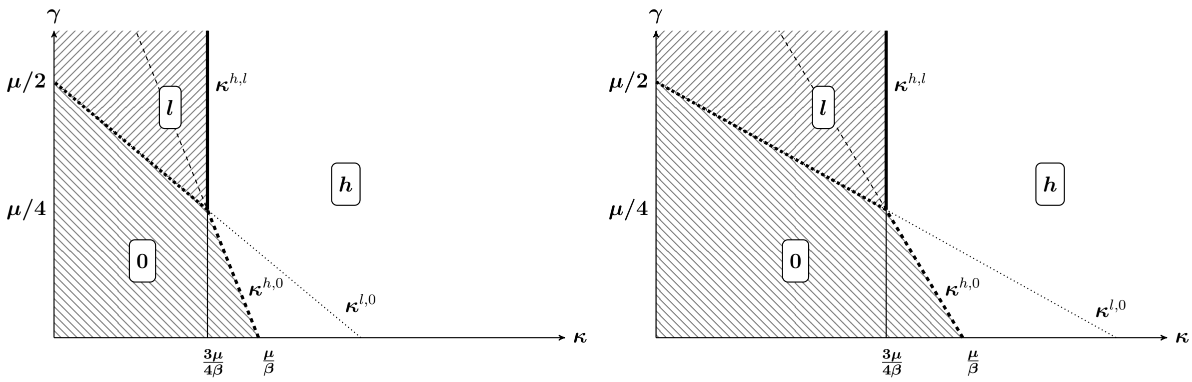
Lemma 5 provides behavioral predictions for the PCP.

Lemma 5. *Let $\beta = 0$. PCP information provision is given by*

$$I^* = \begin{cases} 0, & \textit{if } \gamma \leq \mu/2 \\ l, & \textit{if } \gamma \geq \mu/2 \end{cases}$$

If $\beta > 0$:

$$I^* = \begin{cases} 0, & \textit{if } \kappa \leq \min\left(\kappa^{l,0}, \kappa^{h,0}\right) \\ l, & \textit{if } \kappa^{h,l} \geq \kappa \geq \kappa^{l,0} \\ h, & \textit{if } \kappa \geq \max\left(\kappa^{h,l}, \kappa^{h,0}\right) \end{cases}$$



(a) Large β (b) Small β

Figure 5.2.1: Information provided by the PCP given payment γ and patient type κ, fixed β.

where $\kappa^{l,0} = \dfrac{3(\mu/2-\gamma)}{\beta}$, $\kappa^{h,0} = \dfrac{\mu-\gamma}{\beta}$ *and* $\kappa^{h,l} = \dfrac{3\mu}{4\beta}$.

Proof. If $\beta > 0$:

$$U(h) = F_{PCP} + \beta\kappa - \mu + \gamma \geq F_{PCP} + \beta\kappa/3 - \mu/2 + \gamma = U(l) \iff \kappa \geq \frac{3\mu}{4\beta}$$

$$U(h) = F_{PCP} + \beta\kappa - \mu + \gamma \geq F_{PCP} = U(0) \iff \kappa \geq \frac{\mu-\gamma}{\beta}$$

$$U(l) = F_{PCP} + \beta\kappa/3 - \mu/2 + \gamma \geq F_{PCP} = U(0) \iff \kappa \geq \frac{3(\mu/2-\gamma)}{\beta}$$

If $\beta = 0$:

$$U(h) = -\mu \leq -\mu/2 = U(l)$$

$$U(h) = -\mu + \gamma \geq 0 = U(0) \iff \gamma \geq \mu$$

$$U(l) = -\mu/2 + \gamma \geq 0 = U(0) \iff \gamma \geq \mu/2$$

□

If the PCP is egoistic ($\beta = 0$), he will never provide high-quality information as this is not rewarded. The PCP then only considers whether the payment exceeds the cost of providing information of low quality ($\mu/2$). He will not provide information if $\gamma < \mu/2$, provide low-quality information if $\gamma > \mu/2$, and be indifferent between these two options if $\gamma = \mu/2$.

In contrast, if the PCP is partially altruistic, i.e. $\beta > 0$, the PCP will provide high-quality information if the value of information κ is high enough and provide low-quality information for intermediate values.

Lemma 5 states the values of κ^{I_1,I_2}, with $I_1, I_2 \in \{0, l, h\}$, for which PCPs are indifferent between I_1 and I_2. These are displayed in Figures 5.2.1a and 5.2.1b which show how partially altruistic PCPs provide information depending on the value of information κ and the payment γ. For cases with a high benefit from information provision ($\kappa \geq \kappa^{h,l}$), the PCP prefers high-quality to low-quality information provision, for cases with small benefit ($\kappa \leq \kappa^{h,l}$), the opposite is true. This critical $\kappa^{h,l}$ is independent of the payment γ since the PCP receives the payment for either information quality. For $\kappa \geq \kappa^{h,l}$, the physician will either provide high quality or no information depending on whether $\kappa \gtreqless \kappa^{h,0}$. $\kappa^{h,0}$ is falling in γ, indicating that the physician is trading off providing high-quality information with monetary rewards. Thus, a larger payment γ can induce the physician to provide high-quality information. For $\kappa < \kappa^{h,l}$, the physician will either provide low-quality or no information depending on whether $\kappa \gtreqless \kappa^{l,0}$. $\kappa^{l,0}$ is also falling in γ. Again, a larger payment γ can induce information provision, but only of low quality.[2]

The difference between Figures 5.2.1a and 5.2.1b is the level of physician altruism β. Decreasing altruism shifts all κ boundaries to the right. Thus, a less altruistic PCP will provide less information and worse information quality for some patient types κ that would have received (high-quality) information from a PCP with stronger altruism.

Overall, three areas of information provision emerge. The thick parts of the lines indicate the boundaries between these areas. The payment values $\gamma = \mu/4$ and $\gamma = \mu/2$ are crucial. For $\gamma < \mu/4$, the PCP never provides low-quality information. With such low level of payment, altruism becomes the driving force for information provision, calling for high-quality information if κ is sufficiently high. Increasing γ beyond $\mu/4$ makes the provision of low-quality information interesting for PCPs for low values of κ. Increasing γ beyond $\mu/4$ up to $\gamma = \mu/2$ increases only the amount of low-quality information and has no effect on high-quality information. For

[2] The effect of γ on $\kappa^{l,0}$ is stronger than the effect on $\kappa^{h,0}$. This is the case because low-quality information provision is less effective and less expensive than high-quality information provision.

a bonus payment $\gamma \geq \mu/2$, information will always be provided simply because the payment covers the cost of low-quality information provision.

Proposition 11 (proof in Appendix 5.A.1) summarizes behavioral predictions for partially altruistic physicians that can be inferred from Lemma 5.

Proposition 11.

Let $\beta > 0$.

a) For $\gamma \leq \mu/4$, only no or high-quality information will be provided. For $\gamma = 0$: High-quality information will be provided if and only if $\kappa \geq \mu/\beta$.

b) Increasing γ from 0 to $\mu/4$ incentivizes the PCP to provide high-quality information for $\kappa \in [\kappa^{h,l}, \mu/\beta]$.

c) Increasing γ from $\mu/4$ to $\mu/2$ incentivizes the PCP to provide low-quality information rather than no information if $\kappa \in [\kappa^{l,0}, \kappa^{h,l}]$. It does not change the amount of high-quality information provision. Increasing γ beyond $\mu/2$ does not change the PCP's information provision.

Based on Lemma 5 and Proposition 11, we can formulate testable predictions for partially altruistic physicians for our specific experimental design. These Hypotheses L for the linear version of our model are presented in Section 5.4 after outlining our experimental design and procedures.

5.2.3 Provider behavior – Non-linear PCP utility

If the PCP provides information, he can either gain additional profit or lose profit relative to the case where he does not pass on information. This depends on the payment for information provision γ. For $\gamma < \mu/2$, the PCP loses profit for both low- and high-quality information provision. For $\mu/2 \leq \gamma < \mu$, the PCP gains additional profit when passing on low-quality information and loses profit when passing on high-quality information. Finally, for $\gamma \geq \mu$, the

PCP gains additional profit for both information qualities. In the linear version of our model, this aspect does not make a difference. Prospect theory, however, suggests that individuals' preferences are reference point dependent, that losses from this reference point are weighted more strongly than gains, and that the marginal value of both gains and losses decreases with their size (Tversky and Kahneman, 1991). In our case, PCPs may consider "not providing information" as a reference point since taking this action does not change their profits from the capitation payment which they receive no matter which action they take. In order to study this possibility, we define the value function

$$V(\pi) := \begin{cases} \pi^{\sigma}, & \text{if } \pi \geq 0 \\ -\lambda(-\pi)^{\sigma}, & \text{if } \pi < 0. \end{cases} \tag{5.4}$$

The value function $V(\pi)$, with $\sigma \in (0, 1]$ (smaller σ implies stronger curvature), loss aversion $\lambda > 0$, and decision profit $\pi(I) = \gamma(I) - c(I)$, is an asymmetrical S-shaped value function that reflects the three conditions above. The decision profit is defined as the extra gains or losses that PCPs make when providing information, i.e., the payment minus the cost of information provision. We assume that PCPs evaluate every decision about which information quality to provide in reference to the alternative that they do not provide any information. In this case, their decision profit equals 0 and their gross profit equals the capitation payment F_{PCP}. Figure 5.2.2 depicts an exemplary value function dependent on the decision profit.

As in the linear version of our model, we assume that PCPs trade off the valuation of their own profit against altruistic benefits. Now, however, we replace the payments to the PCP $(p(I) - c(I))$ in the utility function with the value function $V(\pi)$ defined above. Then, PCP utility is captured by the functions

$$U_{PT}(I) := V(\pi(I)) + \beta b(I) \text{ and} \tag{5.5}$$

$$U_{PT}(I) := V(\pi(I)) + \beta s(I) \tag{5.6}$$

for the two cases in which the patient and the specialist, respectively, benefits from information provision. They replace the linear utility functions from Equations (5.2) and (5.3).

In Figure 5.2.2, we illustrate a key difference in the prediction between the linear and non-linear model by comparing bonus payments that just cover the cost of either low- or high-quality

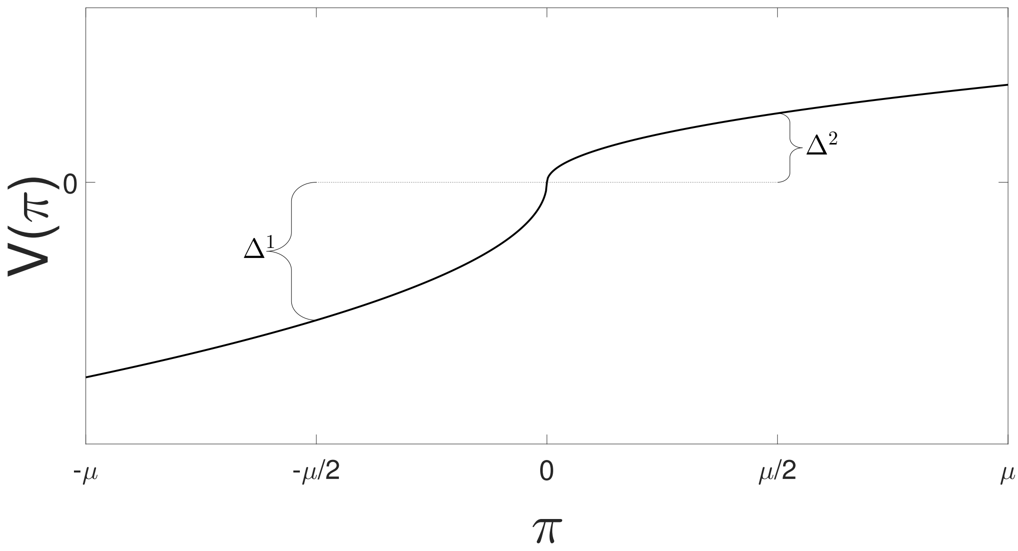


Figure 5.2.2: Example of a Prospect Theory value function

information. Both models predict that for $\gamma \geq \mu/2$ all physicians are willing to provide at least low-quality information quality because its costs are covered. The linear model predicts that increasing γ beyond $\mu/2$ does not impact provided quality because the difference in profits between providing low- and high-quality information is constant.

In the non-linear model, this does not hold. For values of γ between $\mu/2$ and μ, low-quality information results in a gain and high-quality information results in a loss. Under Prospect Theory, if an individual is loss-averse, the difference between a gain and the status quo is less salient than a same-sized difference between a loss and the status quo. This is illustrated in Figure 5.2.2 for an exemplary value function. Given the payment $\gamma = \mu/2$, providing low-quality information yields no decision profit. High-quality information results in a loss of $\mu/2$. The utility difference is given by Δ^1. For $\gamma = \mu$, low-quality information results in a gain of $\mu/2$ and high-quality information results in a zero gain leading to a utility difference Δ^2. Loss aversion implies $\Delta^1 > \Delta^2$. An altruistic PCP weighs his valuation of patient benefit against his valuation of his decision profit. Thus, the non-linear model predicts more high- and less low-quality information provision when the bonus payment is increased from $\mu/2$ to μ.

Focusing on the case with positive altruism, Lemma 6 derives the boundaries for which physicians are indifferent between the different information qualities.

Lemma 6. *Under Prospect Theory, the decision boundaries for the PCP are given by*

$$\kappa^{l,0} = \frac{-3V(\gamma - \mu/2)}{\beta}$$
$$\kappa^{h,0} = \frac{-V(\gamma - \mu)}{\beta} \text{ and}$$
$$\kappa^{h,l} = \frac{3}{2\beta}[V(\gamma - \mu/2) - V(\gamma - \mu)].$$

Lemma 6 follows directly from setting the relevant utilities equal and solving for κ (analogous to Lemma 5). Figure 5.2.3 depicts the predictions for information provision depending on the parameters λ and σ for a fixed altruism parameter β. Note that the value function (5.4) nests the linear version of our model for $\lambda = \sigma = 1$. Therefore, the predictions for linear PCP utility are in the lower right subfigure.

We first consider the upper right subfigure of Figure 5.2.3. Compared to the linear version of our model, only loss aversion λ is introduced. This shifts $\kappa^{h,l}$ to the right for $\gamma \in [0, \mu/2]$ since both quality choices are in the loss-domain. Thus, the difference in profits has a stronger effect on the PCP's value function and he will be less willing to provide information. For $\gamma \in [\mu/2, \mu]$, this effect decreases since providing low-quality information now yields a gain to the PCP. For $\gamma \geq \mu$, the effect is nullified and the same prediction as in the linear utility case emerges.

In the lower left subfigure, only the curvature of the value function is increased compared to the linear version of our model but no loss aversion exists. The result is a point symmetrical S-shaped value function. Several different behavioral predictions emerge. First, the intersection point of $\kappa^{h,l}, \kappa^{h,0}$ and $\kappa^{l,0}$ moves upward from $\mu/4$. It is now in the interval of $[\mu/4, \mu/2)$ (see proof for Proposition 12 in Appendix 5.A.2). Second, increasing the payment γ beyond this point up to $3\mu/4$ increases the amount of low-quality information at the expense of high-quality information. The reason for this is that the marginal utility for profit increases is steepest near the reference point. For low-quality information, this reference point is passed at $\gamma = \mu/2$, whereas the profit from high-quality information is still well in the loss region where the value function is comparatively flat. Increasing γ beyond $3\mu/4$ reverses this effect. Now, more high-quality information is provided at the expense of low-quality information.

Finally, the upper left subfigure depicts the predictions under a value function for which both $\lambda > 1$ and $\sigma < 1$. The effects described above are combined. Proposition 12 (proof in Appendix

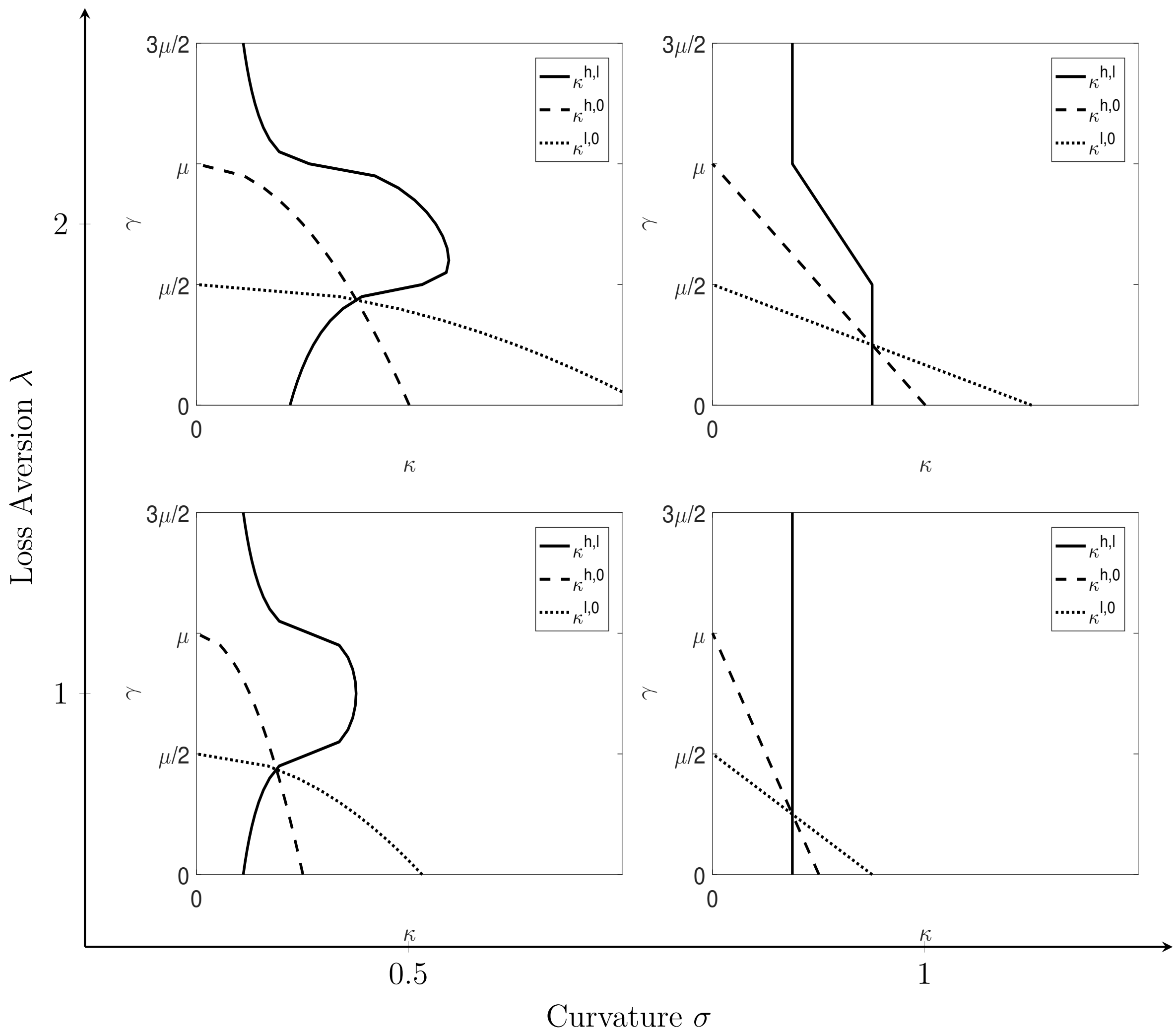


Figure 5.2.3: Predictions for PCP information provision (Prospect Theory)

5.A.2) describes the behavioral predictions of the non-linear version of our model when varying the bonus payment γ for the PCP.

Proposition 12.

Let $\beta > 0, \lambda > 1,$ *and* $\sigma < 1$.

a) Increasing γ *from 0 to* $\mu/4$ *incentivizes the PCP to provide high-quality information rather than no information.*

b) Increasing γ from $\mu/4$ to $\mu/2$ incentivizes the PCP to provide more low- and less high-quality information.

c) Increasing γ from $\mu/2$ to at least μ incentivizes the PCP to provide more high- and less low-quality information.

Based on Lemma 6 and Proposition 12, we can again formulate testable predictions for our specific experimental design. These predictions specifiy differences to the predictions for the linear version of our model (Hypotheses L) and are summarized in Hypotheses NL in Section 5.4, which follows the description of the experimental design and procedures.

5.3 Experiment

5.3.1 Experimental design

In the experiment, subjects take the role of PCPs or specialists. They are matched in groups of two consisting of one PCP and one specialist. Subjects keep their role and partner throughout all decision tasks. In each decision task $t \in \{1, \ldots, T\}$, PCPs decide about the provision of information while referring their patient to the specialist. They can choose between no information (0), low-quality information (l), or high-quality information (h). Based on the information passed on, specialists automatically provide the respective optimal medical treatment in our experiment. By implementing the optimal response, we ensure that PCPs know the exact consequences of their decision and exclude strategic uncertainty as well as risk perceptions as additional influencing factors in our design.[3] As in the seminal model of Ellis and T. G. McGuire (1986), patients are assumed to be fully insured and to accept any medical services provided.

[3]We use this specification in order to investigate pure effects of the bonus payment. Following Tversky and Kahneman (1991), an aversion to lose from their reference profit might still affect PCPs' decisions in our riskless decision context.

We set the cost of providing high-quality information to $\mu = 5$ (and therefore for low-quality information to $\mu/2 = 2.5$) and implement different values for the bonus payment $\gamma \in \{0, 1.25, 2.5, 3.75, 5, 6.25\}$. If $\gamma < 2.5$, PCPs lose profit when passing on low- or high-quality information compared to passing on no information. If $2.5 < \gamma < 5$, they gain additional profit when passing on low-quality information but lose profit when passing on high-quality information. Finally, if $\gamma > 5$, they gain additional profit when passing on information of both qualities.

The benefit from providing high-quality information is κ and the benefit from low-quality information is $\kappa/3$. Our behavioral predictions indicate that PCPs change their choice of information provision depending on β, μ, λ, and σ (see Subsections 5.2.2 and 5.2.3). Since we neither know the individual altruism factor β nor the individual value function parameters λ and σ before the experiment, we choose a broad range of values for κ for the experiment: $\kappa \in \{0, 2, 4, 6, 8, 10, 12, 14, 16, 18\}$. Combining all values for γ and κ results in $T = 60$ decision tasks.

We test the theoretical predictions in five experimental conditions (see Table 5.3.1). We consider the extreme cases that only the patient or only the specialist benefits from information provision. In our baseline condition, P^H, the patient benefits from information provision and the PCP receives a capitation payment of $F_{PCP} = 30$. We set the specialist's capitation payment to $F_S^H = 62.5$ (high) and choose $\bar{c}_S = 20$ as the specialist's maximum costs. Hence, the PCP's profit is always lower than the specialist's profit. We use these specifications for the baseline condition because we assume that usually patients benefit from good communication between their physicians and because PCPs' earnings are lower than specialists' earnings in most OECD countries (OECD, 2015).

In the second condition, P^{CN}, we implement a version of the bonus payment γ that is cost-neutral for the payer. Paying a bonus payment to the physicians may not be cost-effective from the payer's perspective due to the increased costs. Thus, the payer may want to pay the bonus payment in a cost-neutral manner if this does not change the PCP's behavior too much. We implement this by varying the PCP's capitation payment based on the bonus payment. The other experimental parameters are the same as in P^H.

With the third experimental condition, we examine whether PCPs' information provision depends on who benefits from information provision. In condition S^H, only the specialist benefits from information provision in the form of a cost reduction. All other experimental parameters are the same as in P^H. Note that the specialist's maximum costs $\bar{c}_S = 20$ exceed the maximal possible cost reduction from information provision $\kappa^{max} = 18$.

With the fourth and the fifth experimental condition, we investigate whether the relative payments of PCP and specialist or patient, respectively, influence the PCP's willingness to pass on information. Accordingly, we vary the size of the specialist's capitation payment and the patient's base benefit $\bar{b}$ in these conditions. In conditions S^L and P^L, the specialist's capitation payment is reduced to $F_S^L = 27$ (low). Thus, the PCP's profit always exceeds the specialist's (or patient's) profit.[4]

Although no subject takes the role of a patient in our experiment, real patients outside the lab benefit from subjects' decisions. In particular, the monetary value of patient benefit is transferred to the German branch of Doctors of the World (Ärzte der Welt e.V., 80807 Munich, Germany) to support medical treatment for people who have no or only restricted access to the health care system in Germany. The specific procedure is described in the next section.

5.3.2 Experimental protocol

The computerized experiment was programmed with z-Tree (Fischbacher, 2007) and conducted at the Essen Laboratory for Experimental Economics (elfe) at the University of Duisburg-Essen, Germany, from October 2019 to January 2020. We used the online recruiting system ORSEE

[4]In the experimental conditions in which the patient benefits from information provision, the patient's benefit is specified as $B = \bar{b} + b(I)$. In order to keep the decision situation for PCPs comparable between conditions independent of who benefits from information provision, patients receive $\bar{b} = F_S - \bar{c}_S$ as base benefit from medical treatment in all experimental conditions. In conditions in which the specialist benefits from information provision, the patient benefit is reduced to $B = \bar{b}$. Since the fixed patient benefit $\bar{b}$ varies with the specialist's capitation payment F_S, the PCP's payment is always lower than the patient's benefit in conditions with the high payment for the specialist and always higher in conditions with the low payment for the specialist.

Table 5.3.1: Experimental conditions

Experimental condition	Beneficiary from information provision	$F_S \& \bar{b}$	F_{PCP}	Tasks	Groups	Number of subjects
$\mathbf{P^H}$	Patient	High	Fix	60	20	40
$\mathbf{P^{CN}}$	Patient	High	Varying	60	20	40
$\mathbf{S^H}$	Specialist	High	Fix	60	20	40
$\mathbf{S^L}$	Specialist	Low	Fix	60	20	40
$\mathbf{P^L}$	Patient	Low	Fix	60	20	40

(Greiner, 2015) to recruit the 200 participants. As several experimental studies observe no qualitative differences in behavioral responses between medical and non-medical students as well as between medical students and physicians, we chose a conventional subject pool for our study.[5] While this pool of subjects allows identifying the quality of behavioral responses to bonus payments as argued before, we are still cautious with estimating effect sizes based on this pool.

Upon arrival in the laboratory, subjects were randomly assigned to cubicles in which the instructions had been placed before (see Appendix 5.B.1 for the instructions). Once all subjects had finished reading the instructions, they were asked several questions about the decision setting (see Appendix 5.B.2 for the comprehension questions). After answering the questions correctly, subjects were assigned their roles and partner which remained fixed over all 60 decision tasks. The order of the 60 decision tasks was once randomly selected and kept constant for all sessions.[6] In all experimental conditions, subjects were shown a table with the relevant

[5]See, e.g., Brosig-Koch et al. (2016a), Brosig-Koch et al. (2016b), and Brosig-Koch et al. (2017) or Hennig-Schmidt and Wiesen (2014). In particular, Brosig-Koch et al. (2016a), Brosig-Koch et al. (2016b), and Brosig-Koch et al. (2017) include a comparison of medical treatment decisions made by non-medical students and medical students recruited from a similar subject pool as we use for this study. For a short discussion see Galizzi and Wiesen (2018).

[6]We used the same order of tasks in all sessions to keep decisions comparable between different conditions. When we look at the provision of low- and high-quality information over time, we observe no systematic time trend.

information for their decision at the beginning of each task, including the resulting PCP profit, specialist profit, and patient benefit for the three options for information provision. All amounts were given in Euro and one of the 60 tasks was randomly chosen for payment at the end of the experiment. This procedure ensures that income effects that might potentially result from task repetition are excluded.

After the 60 decision tasks, we implemented the three series of pairwise lottery choices from Tanaka et al. (2010) as the second part of the experiment. The three series are designed such that subjects' choices can be used to measure the shape of their value functions according to prospect theory (Kahneman and Tversky, 1979). In particular, series one and two contain paired lotteries with only positive outcomes to determine the concavity of the value function (σ) and the probability weighting function. The paired lotteries in series three contain negative outcomes and can therefore be used to measure subjects' loss aversion parameter (λ). We enforced a monotonic switching point from option A to option B in each of the three series by showing an error message if subjects switched more than once. The amounts from Tanaka et al. (2010) were given in tokens and converted to Euro at an exchange rate of 10 tokens = 1.00 Euro. For each subject, one of the 35 lottery choices was randomly selected for payment and the chosen option was played at the end of the experiment.

As an approximation for subjects' general altruism, we measured their social value orientation (svo) using the decomposed game technique (Liebrand, 1984; McClintock and Liebrand, 1988; see also Brosig, 2002, for a description) in the third part of the experiment. In this part, amounts were again given in Euro and subjects were paid the sum of all amounts they allocated to themselves and all amounts another subject allocated to "other" in the 24 questions. Subjects were aware that there would be a second and third part of the experiment but only shown relevant instructions on screen when a part started. After the third part of the experiment, we asked subjects to complete a short questionnaire which contained questions on risk preferences (questions included in the German Socio Economic Panel, see Dohmen et al., 2011), questions on altruism (based on questions included in the European Values Study, European Values Study, 2015, in the World Values Survey, Inglehart et al., 2014, and in the Global Preference Survey, Falk et al., 2018), and questions on demographics (age, gender, nationality, and field of study).

Subjects in the role of PCPs earned on average 38.08 Euro (min=27.50 Euro, max=206.05 Euro) and subjects in the role of specialists earned on average 37.24 Euro (min=7.60 Euro, max=107.50 Euro) in the experiment.[7] In total, 3098.00 Euro were transferred to Ärzte der Welt e.V. At the end of each experimental session, we randomly selected a subject to monitor the transfer procedure. This subject controlled that a transfer order was correctly filled in and sent to the university's financial department. Therefore, the monitor and one experimenter put the order in an envelope and deposited it in the nearest mailbox. The monitor was paid 5 Euro in addition to his or her earnings from the experiment (see, e.g., Hennig-Schmidt, Selten, et al., 2011, for a similar procedure). Each of the 10 experimental sessions lasted two hours at the most.

5.4 Hypotheses

Based on our theoretical model and our experimental design, we formulate specific hypotheses which we test in the experiment. We start with predictions for partially altruistic physicians derived from the linear version of our model (see Section 5.2.2), i.e. based on Lemma 5 and Proposition 11. All predictions are independent of the specific level of PCPs' altruism.

Hypotheses L.

1. Increasing benefit κ, holding payment γ constant:

a) *The average amount of high-quality information (over all PCPs) increases in κ.*

b) *For $\gamma > \mu/4 = 1.25$, the average amount of low-quality information (over all PCPs) decreases in κ.*

[7]Note that the large maximum payoffs are due to random draws made in the second part of the experiment in which subjects made pairwise lottery choices. Considering the first part of the experiment only, subjects in the role of PCPs earned on average 29.74 Euro (min=25.00 Euro, max=33.75 Euro) and subjects in the role of specialists earned on average 29.60 Euro (min=7.00 Euro, max=52.50 Euro).

2. Increasing payment γ, for an average over all benefits κ:

   a) PCPs do not provide low-quality information for $\gamma \leq \mu/4 = 1.25$.

   b) Increasing γ *from* 0 *to* $\mu/4 = 1.25$ *increases the amount of high-quality information provision and does not affect low-quality information.*

   c) Increasing γ *from* $\mu/4 = 1.25$ *to* $\mu/2 = 2.5$ *increases the amount of low-quality information provision and does not affect high-quality information.*

   d) All PCPs provide some information quality for $\gamma > \mu/2 = 2.5$.

   e) Increasing γ *beyond* $\mu/2 = 2.5$ *does not have an effect on the provision of either information quality.*

Predictions from the non-linear version of our model (see Section 5.2.3) differ only with regard to increases in the bonus payment γ beyond $\mu/4$. Instead of Hypotheses L, 2c) and 2e), we hypothesize based on Lemma 6 and Proposition 12:

Hypotheses NL. Increasing payment γ, for an average over all benefits κ:

a) Increasing γ *from* $\mu/4 = 1.25$ *to* $\mu/2 = 2.5$ *decreases the amount of high-quality information.*

b) Increasing γ *from* $\mu/2 = 2.5$ *to at least* $\mu = 5$ *increases the amount of high-quality information provision and decreases low-quality information provision.*

All other predictions from Hypotheses L remain valid in the non-linear version of our model. In particular, both versions of our model imply that our results should be robust to the implementation of a cost-neutral version of the bonus payment, to a change of the beneficiary of information provision, and to a change of the payoff relation between the PCP and this beneficiary.

5.5 Experimental results

In this section, we analyze the decisions made by subjects in the role of PCPs and test the explanatory power of our theoretical model.[8] Since our model includes altruism, we first look at the general social value orientation of subjects in the role of PCPs before comparing decisions between conditions. We estimate subjects' social value orientation (svo) with their decisions in the social value orientation test which we conducted in the third part of the experiment. The svo angle is measured from the x-axis in a two-dimensional own-other payoff space to a subject's motivational vector which results from his or her choices in the svo test. Based on this angle, subjects can be classified as prosocial (22.5 to 112.5°) or individualistic (–67.5 to 22.5°). We find no significant differences between conditions for the svo angles as well as for the classification in types (see Table 5.C.1 in Appendix 5.C). Moreover, we observe that the svo angle is correlated with altruistic behavior in the main experiment. Over all conditions, the provision of high-quality information in tasks with $\gamma = 0$ is highly positively correlated with the svo angle (Spearman's $\rho = 0.551$, $p < 0.001$). Thus, the svo angle seems to be a valid estimate for subjects' general social value orientations. As the svo angles and types do not significantly differ between conditions, any differences between conditions cannot be explained by differences in subjects' general social value orientations.

In the following, we first test the predictions of our theoretical model in condition P^H, the baseline condition (Subsection 5.5.1). Afterwards, we investigate how a cost-neutral introduction of the bonus payment affects the provision of information (Subsection 5.5.2). In this subsection we also discuss whether a bonus payment is efficient from the payer's perspective. Finally, we check the robustness of our results with regard to a change of the beneficiary from information provision and a change of the payoff relation between the PCP and the beneficiary (patient or specialist; Subsection 5.5.3).

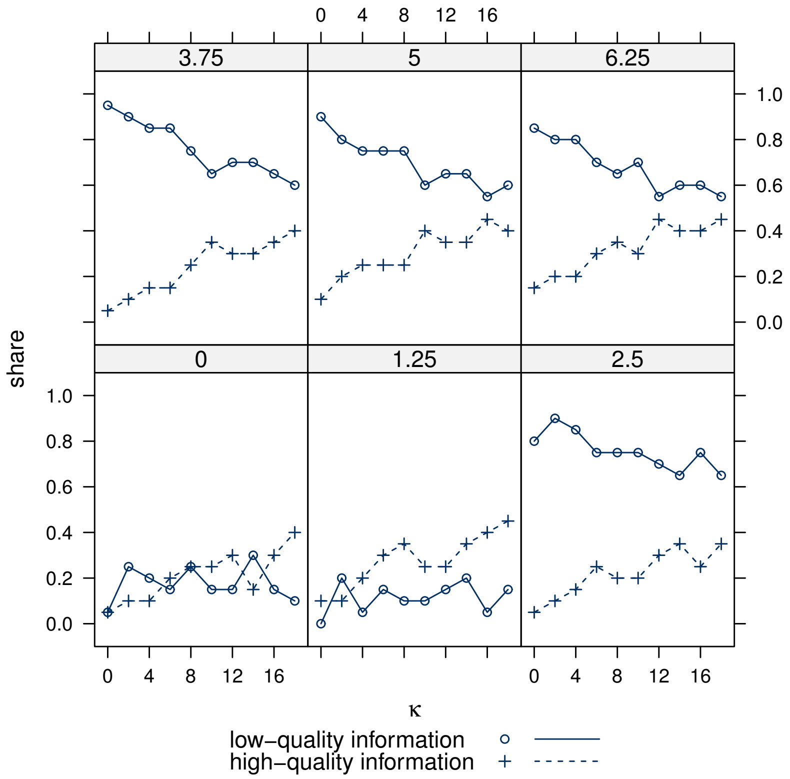


Figure 5.5.1: Share of information passed on by κ (x-axis) and γ (boxes) in condition P^H

5.5.1 Test of model predictions in the baseline condition

In this subsection, we test if decision-making in the baseline condition P^H is in line with our theoretical predictions. Recall that patients benefit from information provision in condition P^H and that the PCPs' profits are always lower in this condition than the specialists' profits (and the patients' benefits). Figure 5.5.1 depicts the share of high- and low-quality information by values of γ (different boxes) and κ (x-axis) in condition P^H.

We can immediately make the following observations. Without a bonus payment ($\gamma = 0$), PCPs do not provide information in most observations. If information is provided, it is often of low quality (17.5% low quality, 21.0% high quality). This is consistent with empirical evidence (Bodenheimer, 2008; Mehrotra et al., 2011). Furthermore, Figure 5.5.1 shows a strong increase

[8] We use R 3.5.3 (R Core Team, 2019) for the data analysis.

of low-quality information provision to a share of around 80% from $\gamma < 2.5$ to $\gamma \geq 2.5$. Note that the costs for passing on low-quality information are $\frac{\mu}{2} = 2.5$. Therefore, PCPs do not lose profit or even gain additional profit if they pass on low-quality information and receive a bonus payment of $\gamma \geq 2.5$.

Let us now consider the Hypotheses from Section 5.4 in more detail. We start with Hypotheses L for the linear version of our model.

Variation in κ (Hypotheses L, part 1):

L, 1a) The average amount of high-quality information (over all PCPs) increases in κ.

In accordance with both versions of our model based on altruism, PCPs are more willing to provide high-quality information if it results in a larger benefit for patients. Over all values for γ, PCPs in condition P^H pass on high-quality information in only 13.3% of observations with $\kappa = 2$. For $\kappa = 18$, the share of high-quality information rises significantly to 40.8% ($p = 0.016$, two-sided Wilcoxon matched-pairs signed-rank test (WMP)).

L, 1b) For $\gamma > 1.25$, the average amount of low-quality information (over all PCPs) decreases in κ.

Again, in line with the theoretical predictions, the share of low-quality information decreases distinctly with κ in tasks with $\gamma > \frac{\mu}{4} = 1.25$ (see Figure 5.5.1). While PCPs pass on low-quality information in 85.0% of observations with $\gamma > 1.25$ and $\kappa = 2$, they do so in only 60.0% of tasks with $\gamma > 1.25$ and $\kappa = 18$ ($p = 0.031$).

These results show that our model can account for the changes in PCPs' information provision that result from a change in the benefit from this provision κ. Furthermore, it confirms that a considerable portion of PCPs are partially altruistic, i.e., they are willing to provide high-quality information only if the benefits are large enough. This is an important requirement for our hypotheses because additional high-quality information could not be incentivized with an increase in γ if all PCPs were fully egoistic or altruistic.

Variation in γ (Hypotheses L, part 2 + Hypotheses NL):

Next, we look at the influence of the bonus payment γ on PCPs' information provision. Table 5.5.1a gives the difference in average low- or high-quality information provision between specific values for γ_1 and γ_2 over all values for κ for all subjects as well as respective theoretical predictions. Let us first consider the predictions of the linear version of our model, that is Hypotheses L part 2.

Table 5.5.1: Effects of γ on information provision in condition P^H

(a) All PCPs ($N = 20$)

	low-quality information				high-quality information			
γ_1 to γ_2	Hyp. L	Hyp. NL	Δ	p-value	Hyp. L	Hyp. NL	Δ	p-value
0 - 1.25	0	0	-0060	0.211	+	+	0.065	0.062
1.25 - 2.5	+	+	0.640	0.000	0	−	-0055	0.250
2.5 - 6.25	0	−	-0075	0.070	0	+	0.100	0.004

(b) PCPs with $\hat{\lambda} > 1$ and $\hat{\sigma} < 1$ ($N = 14$)

	low-quality information				high-quality information			
γ_1 to γ_2	Hyp. L	Hyp. NL	Δ	p-value	Hyp. L	Hyp. NL	Δ	p-value
0 - 1.25	0	0	-0071	0.422	+	+	0.086	0.094
1.25 - 2.5	+	+	0.600	0.000	0	−	-0079	0.250
2.5 - 6.25	0	−	-0107	0.031	0	+	0.121	0.016

Note: Differences in average provision of low and high-quality information between γ_1 and γ_2. Reported p-values result from two-sided Wilcoxon matched-pairs signed-rank tests using average values over all κ for subjects. They test the null hypothesis that information provision does not differ between γ_1 and γ_2. The columns "Hyp. L" and "Hyp. NL" summarize predictions from Hypotheses L and NL, respectively.

L, 2a) PCPs do not provide low-quality information for $\gamma \leq 1.25$.

Against our model's prediction, PCPs provide low-quality information in 14.5% of observations with $\gamma \leq \frac{\mu}{4} = 1.25$. This share differs significantly from zero ($p = 0.004$, two-sided Wilcoxon signed-rank test). This result could potentially be explained by a desire to at least provide some information and, thus, increase patient benefit, even if it is inefficient. In other words, the PCPs may value patient benefits in a non-linear way rather than a linear way as assumed by our model.

L, 2b) Increasing γ *from* 0 *to* 1.25 *increases the amount of high-quality information provision and does not affect low-quality information.*

We observe that PCPs in condition P^H indeed pass on information of low quality similarly often in tasks with $\gamma_2 = 1.25$ as in tasks with $\gamma_1 = 0$ ($\Delta = -0.060$, $p = 0.211$, WMP), but pass on high-quality information somewhat more often ($\Delta = 0.065$, $p = 0.062$).

L, 2c) Increasing γ *from* 1.25 *to* 2.5 *increases the amount of low-quality information provision and does not affect high-quality information.*

In line with the predictions, PCPs more often pass on information of low quality ($\Delta = 0.640$, $p < 0.001$) when the bonus payment increases from $\gamma_1 = 1.25$ to $\gamma_2 = 2.5$ and thus covers the costs for providing low-quality information ($\frac{\mu}{2} = 2.5$). There is no significant difference with regard to the provision of high-quality information ($\Delta = -0.055$, $p = 0.250$).

L, 2d) All PCPs provide some information quality for $\gamma > 2.5$.

For $\gamma > 2.5$, the bonus payment covers the costs for low-quality information provision. PCPs always pass on information. They provide low-quality information in 71.3% and high-quality information in 28.7% of observations with $\gamma > 2.5$.

L, 2e) Increasing γ *beyond* 2.5 *does not have an effect on the provision of either information quality.*

For an increase of the bonus payment beyond $\gamma = \frac{\mu}{2} = 2.5$, the linear version of our model predicts no changes in information provision. However, we observe that the share of low-quality information decreases somewhat ($\Delta = -0.075$, $p = 0.070$), while the share of high-quality information increases ($\Delta = 0.100$, $p = 0.004$) when the bonus payment increases from $\gamma_1 = 2.5$ to $\gamma_2 = 6.25$.

Since our experimental observations cannot be fully explained by the linear version of our model, we further test the non-linear version of our model in which we consider a value function which includes aversion to losses from a reference point in the PCP's utility function (see Subsection 5.2.3). In contrast to the linear version, predictions of the non-linear version of our model differ only in Hypotheses NL which substitute Hypotheses L, 2c) and 2e).

NL, a) Increasing γ from 1.25 to 2.5 decreases the amount of high-quality information.

Although the share of high-quality information tends to decrease when the bonus payment increases from $\gamma_1 = 1.25$ to $\gamma_2 = 2.5$ ($\Delta = -0.055$), the difference is not significant ($p = 0.250$).

NL, b) Increasing γ from 2.5 to at least 5 increases the amount of high-quality information provision and decreases low-quality information provision.

In line with the hypothesis (and as described with regard to Hypothesis L, 2e)), PCPs pass on information of low quality less often ($\Delta = -0.075$, $p = 0.070$) and information of high quality more often ($\Delta = 0.100$, $p = 0.004$) as the bonus payment increases from $\gamma_1 = 2.5$ to $\gamma_2 = 6.25$.[9] This implies that our experimental observations for higher bonus payments (i.e., $\gamma > 1.25$) are more in line with the non-linear version of our model which accounts for an aversion to losses from a reference point in the PCP's utility function than with the linear version.

Since predictions from Hypotheses NL are only valid for loss-averse physicians with an S-shaped value function, we also analyze observations separately for subjects with $\hat{\lambda} > 1$ and $\hat{\sigma} < 1$. We estimate the parameters $\hat{\lambda}$ and $\hat{\sigma}$ from subjects' decisions in the second part of the experiment. Table 5.5.1b gives the difference in average low- or high-quality information provision between γ_1 and γ_2 over all values of κ for subjects with $\hat{\lambda} > 1$ and $\hat{\sigma} < 1$. Again, we find support for Hypotheses NL. Although the decrease in high-quality information provision between $\gamma_1 = 1.25$ and $\gamma_2 = 2.5$ is not significant also for this population, the share of information provided tends

[9]We find qualitatively the same results when we consider a change from $\gamma_1 = 2.5$ to $\gamma_2 = 5$.

to change in the predicted direction.[10] [11]

Result 1. *Without a bonus payment, little information is provided with a considerable share of low-quality information. The introduction of a bonus payment increases the provision of both low- and high-quality information. PCPs provide less low- and more high-quality information as the benefit from information provision increases to at least cover the costs of high-quality information. This observed behavioral pattern is in line with the non-linear version of our model considering aversion to losses from a reference point in addition to altruism in the PCP's utility function.*

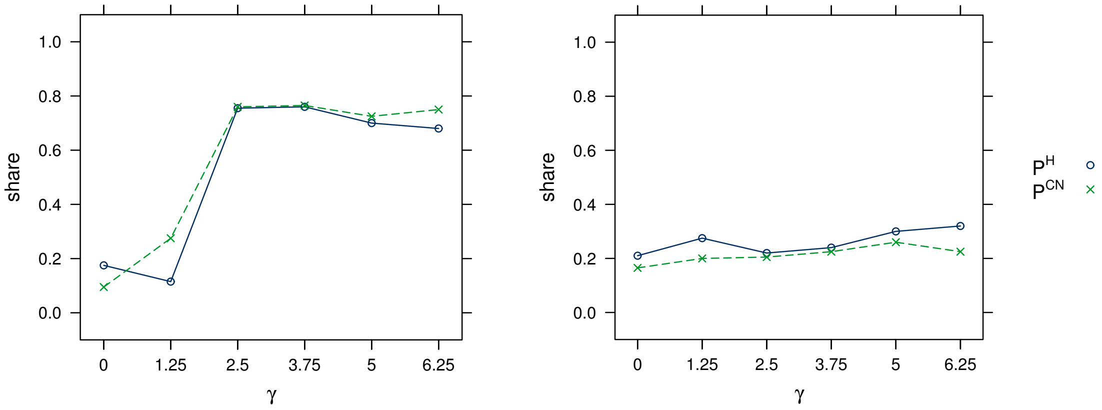
5.5.2 Effects of a cost-neutral version of the bonus payment and welfare implications

In this subsection, we analyze how a cost-neutral version of the bonus payment influences information provision. For this purpose, we reduced the PCP's capitation payment by the costs which incur to the payer in expectation. More specifically, based on our theoretical predictions, PCPs receive $F_{PCP}^{var} = F_{PCP} - \gamma$ in tasks with $\gamma \geq 2.5$ in condition P^{CN}. For $\gamma < 2.5$, , the theoretical predictions for information provision depend on the benefit from information provision κ and on physician altruism β. In order to estimate the costs for information provision

[10]On average, subjects in the role of PCPs have an estimated curvature of $\hat{\sigma} = 0.78$ and an estimated loss aversion $\hat{\lambda} = 3.22$ in the baseline condition P^H. Over all conditions, the values are $\hat{\sigma} = 0.71$ and $\hat{\lambda} = 3.14$. Thus, our PCPs are strongly loss averse but their value functions only show a moderate curvature on average. This may help to explain our weak results for Hypothesis NL, a) and strong results for Hypothesis NL, b) since a) only holds for $\sigma < 1$, whereas b) is mostly driven by loss-aversion (see Figure 5.2.3).

[11]We also analyze decisions of subjects who do not exhibit $\hat{\lambda} > 1$ and $\hat{\sigma} < 1$ separately. The difference in average low- or high-quality information provision between γ_1 and γ_2 for these subjects is given in Table 5.C.2 in Appendix 5.C. The observed reactions to an increase in γ from these subjects are in line with predictions for changes in the share of low-quality information in the linear model. However, these subjects do not significantly change their rate of high-quality information with an increase in γ which is not in line with either version of our model. As only 6 subjects in the role of PCPs do not exhibit $\hat{\lambda} > 1$ and $\hat{\sigma} < 1$, we are cautious drawing conclusions from these observations.

for $\gamma < 2.5$, we conducted experimental sessions for condition P^H before determining the experimental parameters for P^{CN}. In condition P^H, PCPs pass on information of any kind in about 40% of observations with $\gamma < 2.5$. Assuming a similar rate for P^{CN}, we specify $F_{PCP}^{var} = F_{PCP} - 0.4\gamma$ for $\gamma < 2.5$. The actual amounts paid to PCPs are almost identical for all levels of γ, confirming that condition P^{CN} is indeed cost-neutral.[12]



(a) Low-quality information (b) High-quality information

Figure 5.5.2: Effects of γ on information provision in conditions P^H and P^{CN}

Figure 5.5.2 depicts the average low- and high-quality information provision in conditions P^H and P^{CN}. In line with our models that do not imply a difference between the PCPs' decisions made in conditions P^{CN} and P^H, subjects in the role of PCPs provide, on average, a similar share of low- and high-quality information in both conditions in the absence of a bonus payment ($\chi^2 = 0.052$, $p = 0.977$, multivariate Kruskal-Wallis test; see also Table 5.C.3 in Appendix 5.C) as well as over all values for γ ($\chi^2 = 0.035$, $p = 0.983$).[13]

As a result, we observe very similar responses to changes in γ in condition P^{CN} as in P^H. There is only one significant difference with regard to PCPs' responses between the two conditions: While the share of low-quality information significantly increases in P^{CN} between $\gamma_1 = 0$ and

[12]The amount paid to a PCP in tasks without bonus payments in conditions P^H and P^{CN} corresponds to 30 Euro. In condition P^{CN}, the amounts are in the range $[29.84, 30.09]$ over all values for γ.

[13]We perform an approximative version of the multivariate Kruskal-Wallis test (Puri and Sen, 1966; Puri and Sen, 1971) using the R package coin (Hothorn et al., 2008).

$\gamma_2 = 1.25$, there is no significant change in P^H (P^H: $\Delta = -0.060$, P^{CN}: $\Delta = 0.180$, $p < 0.001$, MWU).

Result 2. *Information provision is similar regardless of whether the bonus payment is designed cost neutrally or not.*

Now, we take a look at the welfare implications of our findings. We take the perspective of a payer who considers both the benefit of information provision b and his own payments p. The question we are interested in answering is whether the additional bonus payments are justified by the additional benefits from information provision. Figure 5.5.3 shows, $\Delta(b-p)$, the change in average payer welfare relative to $\gamma = 0$ ($\Delta(b-p) \equiv b(\gamma) - p(\gamma) - [b(\gamma = 0) - p(\gamma = 0)]$), depending on the bonus payment γ for different levels of κ in the baseline condition P^H and the cost-neutral condition P^{CN}. For the cost-neutral condition P^{CN}, an increase in γ is compensated by a decrease in the capitation payment F_{PCP}.

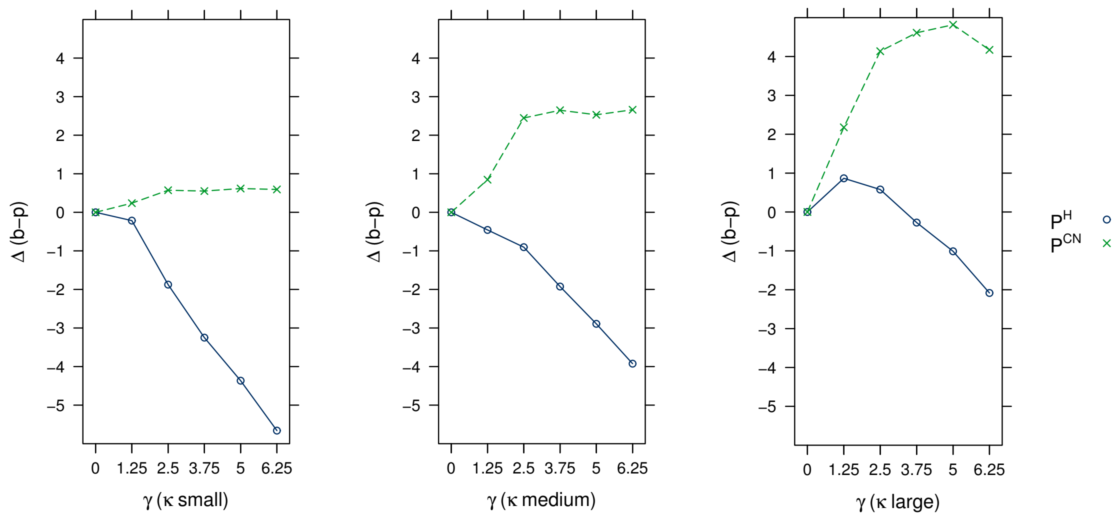


Figure 5.5.3: Average payer welfare (relative to $\gamma = 0$) depending on γ for different levels of κ (small: $\kappa < 5$, medium: $5 \leq \kappa < 14$, large: $14 \leq \kappa$)

For the baseline condition P^H, Figure 5.5.3 shows that a bonus payment is not effective for small and medium values of κ. As shown in Figure 5.5.1, little low- and high-quality information is provided in this range for small values of γ and it generates only few benefits. If γ is increased to 2.5, low-quality is triggered but it only provides little value. Furthermore, as only some PCPs change their information provision between two levels of γ, the payer's welfare additionally deteriorates for an increase in γ due to the extra payments to PCPs who already provide information for smaller bonus payments.

For large κ, however, $\gamma = 1.25$ and $\gamma = 2.5$ enhance payer welfare compared to $\gamma = 0$ with $\gamma = 1.25$ leading to the highest payer welfare. This is because the amount of high-quality information increases for $\gamma = 1.25$. For high values of κ, this enhances welfare. Increasing γ to 2.5 yields more low-quality information which is less efficient from the payer's perspective. Increasing γ from 2.5 to at least 5 improves information quality. However, this effect can not compensate for the increased costs of the bonus payment. Thus, large bonus payments are not efficient if their implementation is not cost-neutral.

In condition P^{CN}, payer welfare is always higher for a positive payment as the payer benefits from any type of information provision without extra costs. Due to the strong increase in low-quality information, payer welfare increases strongly from $\gamma = 0$ to $\gamma = 2.5$ for any positive level of κ. Additionally, if κ is large, a payment that covers the costs of high-quality information ($\gamma = 5$) performs best due to enhanced average information quality.

Result 3. *In the confines of our experiment, we find that the payer who is not able to implement the bonus payment in a cost-neutral manner should only pay a bonus payment if the benefit from information provision κ is sufficiently large. The bonus payment should not exceed half the costs for the provision of low-quality information ($\gamma = 1.25$). If the payer is able to implement the bonus payment in a cost-neutral manner, he should pay a larger bonus payment (e.g., $\gamma = 5$).*

5.5.3 Robustness checks

In this subsection, we test whether our results obtained in the baseline condition are robust to a change of the beneficiary of information provision as well as to a change of the payoff relation between the PCP and the beneficiary. We first test whether PCPs pass on the same information if the specialist benefits from information provision instead of the patient. For this purpose, we compare decisions made in condition S^H to those made in the baseline condition

P^H. Differences between both conditions might particularly occur if PCPs are more or less altruistic towards specialists than towards patients. According to our models, more altruistic physicians will supply more high-quality information for any level of bonus payment. Therefore, we focus on this type of information. As depicted in Figure 5.5.4, PCPs less often pass on information of high quality if specialists benefit from information provision on average over all bonus payments for $\kappa > 0$ (5.8% versus 28.1%; $\chi^2 = 8.251$, $p = 0.011$, multivariate Kruskal-Wallis test; see also Table 5.C.3 in Appendix 5.C for values over all κ). This observation might be explained by a lower degree of altruism of PCPs towards specialists than towards patients, at least as long as the PCP receives a lower profit than the specialist and the patient, respectively.

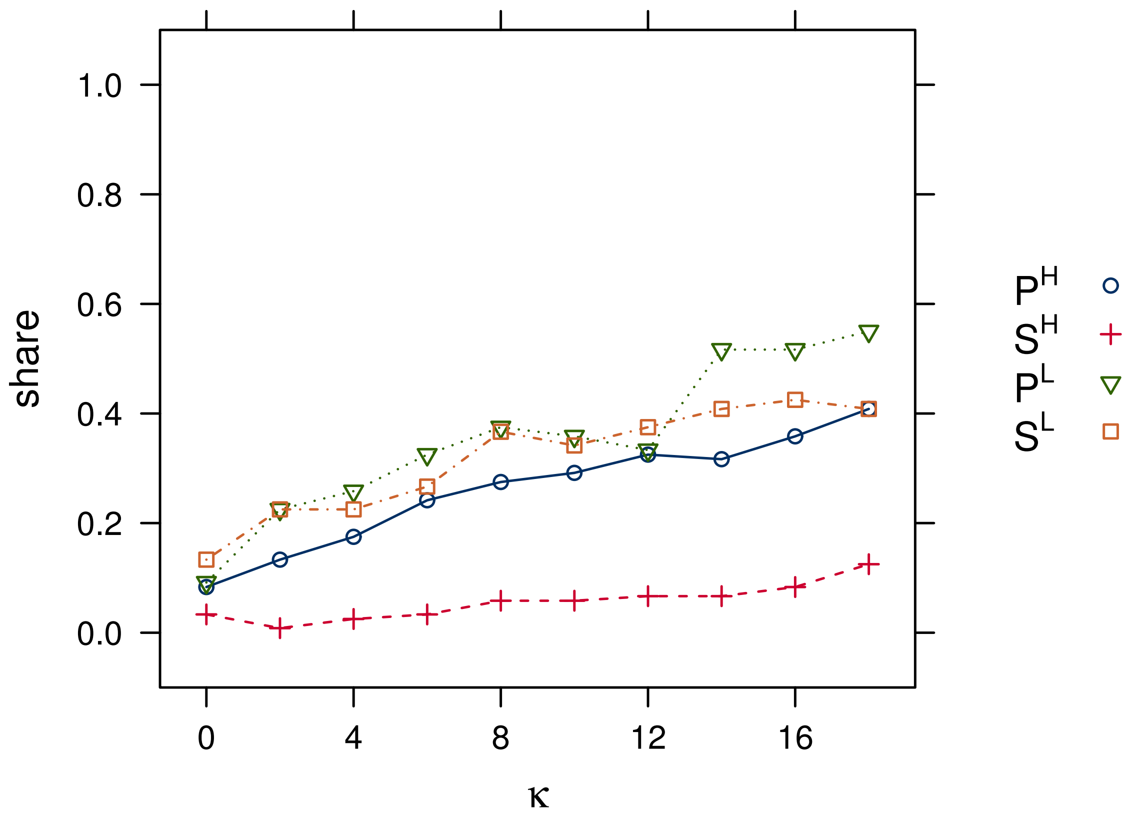


Figure 5.5.4: Average high-quality information provision over all bonus payments γ over values for κ

We also test whether the type of the beneficiary matters for PCPs' information provision for the case that the PCP receives a higher profit than the beneficiary (condition S^L versus condition P^L). As shown in Figure 5.5.4, PCPs pass on information of high quality similarly often if specialists instead of patients benefit from information provision over all bonus payments for $\kappa > 0$ (33.8% versus 38.4%; $\chi^2 = 3.026$, $p = 0.215$, multivariate Kruskal-Wallis test). Hence, we cannot reject the null hypothesis that PCPs are equally altruistic in both conditions.

Result 4. *The beneficiary from information provision only matters for information provision when this beneficiary has a higher profit than the PCP. Then, PCPs pass on less high-quality information on average when specialists benefit from information provision than when patients benefit.*

In conditions S^L and P^L, specialists and patients receive a lower payoff than PCPs compared to conditions S^H and P^H. We can therefore test whether the payoff relation between the PCP and the beneficiary of information provision matters for PCP's behavior. We first compare PCPs' information provision between condition P^L and the baseline condition P^H. In both conditions, patients benefit from information provision. Although the share of high-quality information tends to be slightly larger in P^L than in P^H (see Figure 5.5.4 and Table 5.C.3 in Appendix 5.C), information provision is not significantly different between the two conditions in tasks with $\kappa > 0$ ($\chi^2 = 2.765$, $p = 0.267$, multivariate Kruskal-Wallis test).

Next, we consider the scenario where specialists benefit from information provision, i.e., we compare PCPs' decisions made in condition S^L to those made in S^H. We observe that the share of high-quality information is significantly higher in condition S^L than in S^H when considering tasks with $\kappa > 0$ ($\chi^2 = 7.459$, $p = 0.020$, multivariate Kruskal-Wallis test; see Figure 5.5.4). If $\kappa = 0$, PCPs pass on information of high quality in only a few cases (S^H: 3.3%, S^L: 13.3%).

Result 5. *The payoff relation between the PCP and the beneficiary of information provision only matters for PCPs' information provision when specialists benefit but not when patients do. When specialists benefit from information provision, PCPs pass on less high-quality information when specialists have higher earnings than when specialists have lower earnings than PCPs.*

Overall, our observations regarding the influence of the payoff relation between the PCP and the beneficiary of information provision and regarding the type of the beneficiary might be explained by a lower degree of altruism, which is revealed by PCPs if two conditions hold. First, PCPs earn less than the beneficiary of information provision and second, this beneficiary is a specialist and not a patient.

5.6 Conclusion

In order to study whether bonus payments for information provision can improve the information flow between physicians, we propose a theoretical model according to which PCPs can pass on information to specialists at a cost. Information can be of low or of high quality. We assume high-quality information to be more cost-effective and examine an agency problem in which the payer is not able to discriminate between low- or high-quality information and can only implement a bonus payment for information provision independent of its quality. Regarding the PCP's utility function, we consider both linear preferences and an aversion to losses from a reference point. Both versions of our model predict that partially altruistic PCPs provide more high-quality information when the benefit from information provision increases. Furthermore, a bonus payment for information provision induces the provision of low- and high-quality information. The hypotheses for partially altruistic PCPs regarding changes in the bonus payment depend on the assumed utility function. In contrast to a linear utility function, a non-linear model based on prospect theory predicts that high-quality information provision first decreases and then increases if the bonus payment is raised.

We test the theoretical predictions in a controlled laboratory experiment that allows to observe ceteris paribus variations of parameters. Our results confirm the empirical observation that little information is passed on without a bonus payment and that a considerable share of the information is of low quality. In line with our hypotheses, PCPs provide less low- and more high-quality information as the benefit from information provision increases. PCPs pass on more information on average as the bonus payment increases. Average information quality increases if the bonus payment covers the costs of high-quality information provision instead of covering only the costs of low-quality information. This observed behavioral pattern is in line with the non-linear version of our model considering an aversion to losses from a reference point in addition to altruism. These results hold independently of whether the bonus payment is designed cost neutrally or not. Overall, paying for information provision is cost-effective from a payer's perspective mostly if the benefits are high and the bonus payment at most covers the costs for providing information of low quality. An exception is the cost-neutral introduction of a bonus payment. In this case, a bonus payment that covers the costs for high-quality information performs well in our experiment.

Checking the robustness of our results with regard to changes in the decision-setting, we find that the provision of high-quality information depends on an interaction of who benefits from information provision and the payoff relation between beneficiaries and PCPs. On average, PCPs pass on less high-quality information when specialists benefit from information provision and specialists have higher earnings than PCPs, indicating less concern for the payoff of the high earning specialist.

Our experimental study based on a theoretical model sheds light on a number of key issues pertaining to the information provision between physicians and may therefore be informative for the design of programs which seek to improve information exchange. In particular, our results indicate that bonus payments may lead to a considerable amount of low-quality information provision. Paying for information in form of extra bonuses may therefore generate only little welfare gains. This favors bonus schemes which are financed by the reduction of other payment components. A limitation is that our experimental analysis relied on assumptions on parameters for which no empirical estimates are yet available. This calls for future work aimed at modeling more closely actual information provision bonus programs and the costs and benefits involved.

Appendices

5.A Proof of Propositions

5.A.1 Proof of Proposition 2

Proof. a) $\kappa^{h,l} \leq \kappa^{l,0} \iff \gamma \leq \mu/4 \iff \kappa^{h,0} \leq \kappa^{l,0}$.

If the PCP is willing to provide low-quality information rather than no information at all, this implies $\beta\kappa/3 \geq \mu/2 - \gamma$. For $\gamma \leq \mu/4$, this implies $U(h) = \beta\kappa - \mu \geq \beta\kappa/3 - \mu/2 = U(l)$ and therefore the PCP prefers providing high-quality information rather than low-quality information. This implies that, if $\gamma = 0$, the PCP will provide high-quality information for $\kappa \geq \kappa^{h,0}(\gamma = 0) = \mu/\beta$ and will provide no information otherwise.

b) $\partial\kappa^{h,0}/\partial\gamma < 0$ and $\mu/\beta > \kappa^{h,l}$.

Hence, as long as $\kappa^{h,0} \geq \kappa^{h,l} \iff \gamma \leq \mu/4$, high-quality information provision increases. For $\gamma = \mu/4, \kappa^{h,0} = \kappa^{h,l} = \kappa^{l,0}$ and for any $\kappa \geq \kappa^{h,l}, I^* = (1, h)$.

c) $\kappa^{h,l} \geq \kappa^{l,0} \iff \gamma \geq \mu/4$ and $\partial\kappa^{l,0}/\partial\gamma < 0$.

Therefore, more low-quality information is provided for $\kappa \in [\kappa^{l,0}, \kappa^{h,l}]$ with rising γ. Since $\kappa^{h,0} \geq \kappa^{l,0} \iff \gamma \geq \mu/4$, no further high-quality information is provided. Changing γ does not impact whether a PCP prefers low- or high-quality information for a given κ. Hence, the provision of high-quality information remains unchanged. For $\gamma = \mu/2$, $\kappa^{l,0} = 0$. Therefore, there is no increased information provision beyond this point. □

5.A.2 Proof of Proposition 3

Proof. Let $\beta > 0$ be fixed. First, note that $\kappa^{h,l} = \kappa^{h,0} = \kappa^{l,0} \iff \gamma = \frac{\mu(1-3^{1/\sigma}/2)}{1-3^{1/\sigma}} \in [\mu/4, \mu/2)$. Thus, for $\gamma \leq \mu/4$, PCPs provide no information if $\kappa \leq \kappa^{h,0}$ and high-quality information if $\kappa \geq \kappa^{h,0}$. For $\gamma \geq \mu/2$, PCPs provide low-quality information if $\kappa \leq \kappa^{h,l}$ and high-quality information if $\kappa \geq \kappa^{h,l}$.

a) $\kappa^{h,0}(\gamma = \mu/4) - \kappa^{h,0}(\gamma = 0) = \lambda/\beta[(3\mu/4)^\sigma - \mu^\sigma] < 0$

This implies that the minimum κ for which the PCP is indifferent between supplying h and 0 shrinks. Therefore, the PCP is incentivized to play $I^* = h$ rather than $I^* = 0$.

b) $\kappa^{h,l}(\gamma = \mu/2) - \kappa^{h,0}(\gamma = \mu/4) = \lambda/\beta[3/2(\mu/2)^\sigma - (3\mu/4)^\sigma] > 0$

c) Let $w \geq 0$, then $\kappa^{h,l}(\gamma = w+\mu) - \kappa^{h,l}(\gamma = \mu/2) = 3/(2\beta)[(w+\mu/2)^\sigma - w^\sigma - \lambda(\mu/2)^\sigma] < 0$. □

5.B Information on the experiment

5.B.1 Instructions of the experiment (translated from German)

Note that the instructions refer to condition P^H. The text in squared brackets [] denotes changes for other conditions.

Welcome to the experiment!

You are participating in a study on decision-making in experimental economics research. During the experiment, you are asked to make decisions. By doing so, you can earn money. How much money that is depends on your decisions and on the decisions of other participants.

Immediately after the experiment, all participants are paid individually in cash. Please stay therefore seated after the experiment until your cubicle number is called.

The experiment takes about 120 minutes including the payment procedure and consists of three different parts. Presumably, part 1 takes considerably longer than part 2 and part 3. Before each of the three parts of the experiment, you receive detailed instructions. Please note that your decisions in one part of the experiment do not have any influence on another part of the experiment.

In part 1 and 3 of the experiment, all amounts are given in Euro. In part 2 of the experiment, all amounts are given in tokens. During the experiment, no participant receives any information about the identity of other participants.

Part 1

Please read the following instructions carefully. In case you have questions, you can raise your hand or open the door at any time. We will then come to your cubicle and answer your question.

In the first part of the experiment, you are taking part in 60 decision periods. There are participants in the role of primary care physicians and participants in the role of specialists.

At the beginning of the first decision period, one of the two roles, primary care physician or specialist, is randomly assigned to you. You keep your role during all 60 decision periods. Each primary care physician is matched with one specialist. This matching is also retained during all decision periods.

Description of the decision periods

In each of the 60 decision periods, the primary care physician refers a patient to the specialist. The primary care physician has examined the patient and has information on the diagnosis. The primary care physician cannot treat the patient himself but needs to refer the patient to the specialist for treatment. With the referral, the primary care physician can pass on information on the diagnosis to the specialist. He has three options for this:

i He can pass on no information. This will incur no costs for information provision.

ii He can pass on information of low-quality. This will incur costs of 2.50 Euro.

iii He can pass on information of high-quality. This will incur costs of 5.00 Euro.

The costs for information provision are kept constant for all decision periods and are displayed on screen at the beginning of each period as a reminder. When the primary care physician passes on information to the specialist, there is an additional benefit for the patient [*S:* a cost reduction for the specialist]. The size of the benefit [*S:* the cost reduction] depends on the information quality and may vary between decision periods. The primary care physician is displayed the benefit for the patient [*S:* the cost reduction for the specialist] from information provision on screen for each of the three options for information provision at the beginning of each period. Based on the information passed on by the primary care physician, the examination and treatment decision which is optimal for the patient is made automatically for the specialist. Real patients will benefit from the monetary value of patient benefit (see section "Payment").

Physician income and patient benefit

The primary care physician receives a capitation payment of 30 Euro in each period. [*CN:* The primary care physician receives a capitation payment in each period.] If he passes on information about the diagnosis to the specialist, he also receives an additional bonus payment. The size of the bonus payment [*CN:* The size of the capitation payment depends on the size of the bonus payment. The latter] is independent of the information quality and may differ between decision periods. The size of the [*CN:* capitation payment and the size of the] bonus payment is displayed on screen to the primary care physician at the beginning of each period.

The specialist receives a capitation payment of 62.50 Euro [*L:* 27 Euro] in each period. He incurs costs for the diagnosis of 20 Euro. [*S:* If the primary care physician passes on information on the diagnosis to the specialist, the information provision reduces the specialist's costs.]

Income of the primary care physician =

30.00 (Capitation payment) [*CN:* Capitation payment]
− Costs for information provision (Costs primary care physician)
+ Bonus payment for information provision

Income of the specialist =

62.50 [*L:* 27] (Capitation payment)
− 20.00 (Costs specialist)
[*S:* + Cost reduction from information provision]

The patient receives a benefit of 42.50 Euro [*L:* 7 Euro] from treatment in each period. If the primary care physician passes on information for the diagnosis to the specialist, the patient reveives an additional benefit from information provision.

Patient benefit =

42.50 [*L:* 7] (Benefit from treatment)
+ Benefit from information provision

The income of both physicians and the patient's benefit are displayed to both physicians on screen at the beginning of each period. At the end of each period, both physicians are displayed their own income, the other physician's income and the patient's benefit on screen.

Payment

Upon completion of part 1, one of the 60 decision periods is randomly selected for payment. The payment of physicians or patients are determined from this period. The payment results from the income or benefit in this period. This amount is paid out in cash at the end of the experiment together with the earnings from the other parts of the experiment.

In this part of the experiment, no participants in the role of patients are physically present in the laboratory. Real patients benefit from the patient benefits resulting from the decisions made in the randomly selected period. The monetary value of these benefits in Euro is provided for national projects of Ärzte der Welt e.V., 80807 Munich. These projects provide basic health care for people in Germany who do not have health insurance or for other reasons do not have access to the health system. Ärzte der Welt e.V. bears the DZI Seal-of-Approval which certifies that the donations are used transparently, purposefully and economically.

After the experiment, the amount is transmitted to Ärzte der Welt e.V. by the experimenter together with a control person. The control person enters the amount in Euro, which results from the realized patient benefits of the randomly selected round, into a form for payment order to Ärzte der Welt. The payment of the amount from the funds earmarked for this experiment is then arranged by the financial administration of the University of Duisburg-Essen to Ärzte der Welt. The form is placed in a postpaid envelope addressed to the finance department of the University of Duisburg-Essen. This envelope is then posted in the nearest mailbox by the control person together with the experimenter.

After part 3 of the experiment, one participant is randomly chosen as the control person. The control person receives a compensation of 5 Euros for this task in addition to the payment from the experiment. The control person confirms by signing a statement which can be viewed by all participants in the office of the Chair of Quantitative Economic Policy (Room WST-A.10.08) that he/she has correctly fulfilled the tasks described here. All participants of the experiment may upon request also see a payment receipt from Ärzte der Welt in the office of the above-mentioned chair.

The payment for part 1 is only made at the end of the experiment. Not until then you will be shown the payment information on screen.

Prior to the 60 decision periods, we ask you to complete a comprehension test on screen. The decision periods will start as soon as all participants have answered the comprehension questions correctly. These questions do not affect your payment.

Part 2 and part 3

The instructions for part 2 and part 3 of the experiment will appear on screen when the part starts. In part 2, all amounts are given in tokens, in part 3 all amounts are given in Euro.

At the end of the experiment, your earnings from part 2 are converted to Euro. Subsequently, your earnings from all three parts of the experiment are added up and paid to you in cash.

After part 3 of the experiment, a questionnaire opens. The payment procedure starts after all participants have answered the questionnaire completely.

Part 2 (on screen)

In part 2 of the experiment, you see three lists with pairs of different lotteries on screen. Lists 1 and 2 each contain 14 pairs of different lotteries, list 3 contains 7 pairs of different lotteries. The pairs of lotteries each differ in payments, not in probabilities for the possible payments. For each of the 35 pairs of lotteries in total, you have to choose either option A or option B.

The payments of the lotteries are given in tokens with 10 tokens = 1 Euro. At the end of the experiment, one of the 35 pairs of lotteries is randomly chosen for payment. The option you have chosen for this pair is carried out. The payment from this lottery is converted to Euro at the end of the experiment and paid to you in cash together with your earnings from the other parts of the experiment.

Part 3 (on screen)

In part 3 of the experiment, you are randomly matched with another participant. In the following we name the other participant "the other person". You do not receive any information about the other person's identity. Conversely, the other person receives no information about your identity.

You and the other person see 24 questions. In each question, you can both choose independently between option A and option B. Each option allocates a certain amount of money in Euro to you and the other person. Your payment is determined by the sum of monetary amounts assigned to you in all 24 questions. You receive the sum of monetary amounts you have allocated to yourself. In addition, you receive the sum of monetary amounts the other person has allocated to you. The other person receives the sum of monetary amounts you allocated to the other person and the sum of monetary amounts they allocated to themselves. This means that your decisions affect both your own payment and the payment of the other person. Likewise, the other person's decisions affect your payment and the payment the other person receives.

For each of the 24 questions, please indicate the option you prefer most. Once you have made your decision, mark the appropriate item and click "OK" to proceed to the next question.

5.B.2 Comprehension questions (P^H [S^H])

1. Assuming, the following values are given in a period:

   a. What would be the costs for information provision for the primary care physician if the primary care physician passes on information of low-quality?

   b. What would be the benefit from information provision for the patient [cost reduction from information provision for the specialist] if the primary care physician passes on information of low-quality?

   c. What would be the income of the primary care physician if he passes on information of high-quality to the specialist?

	For passing on no information	For passing on information of low-quality	For passing on information of high-quality
Costs for information provision	0.00	2.50	5.00
Bonus payment for information provision	0.00	3.00	3.00
Benefit for the patient [Cost reduction for the specialist] from information provision	0.00	3.00	9.00
Income primary care physician	30.00	30.50	28.00
Income specialist	42.50	42.50 [45.50]	42.50 [51.50]
Total patient benefit	42.50	45.50 [42.50]	51.50 [42.50]

Please mark the correct answer in each case:

2. The income of physicians or the benefit of the patient from how many decision periods is paid out?

   a. The income of physicians or the benefit of the patient from one randomly selected period.

   b. The sum of the income of physicians or the benefit of the patient from 3 randomly selected periods.

   c. The sum of the income of physicians or the benefit of the patient from all periods.

3. Participants in the role of a specialist ...

   a. ... decide in each period on passing on information.

   b. ... do not make own decisions about the examination and treatment of the patient. The optimal examination and treatment decision is made automatically.

c. ... decide in each period on the examination and treatment of the patient.

5.C Additional figures and tables

Table 5.C.1: Classification from social value orientation test

		Svo type	
	Svo angle	prosocial	individualistic
$\mathbf{P^H}$	20.804	9	11
$\mathbf{P^{CN}}$	16.434	11	9
$\mathbf{S^H}$	15.807	7	13
$\mathbf{S^L}$	13.999	9	11
$\mathbf{P^L}$	22.731	10	10
Null hypothesis	p-value^a	p-value^b	
$P^H = P^{CN}$	0.580	0.752	
$P^H = S^H$	0.680	0.748	
$P^H = P^L$	0.454	0.751	
$S^H = S^L$	0.620	0.748	
$S^L = P^L$	0.093	1.000	

Note: Average angle resulting from decisions in the social value orientation test and number of subjects in the role of primary care physicians who are classified as a specific type based on the angle. For simplicity, we subsume the "altruistic" and "cooperative" type as "prosocial" and the "individualistic" and "competitive" type as "individualistic". Considering all four types separately leads to similar results. a) Reported p-values result from two-sided Mann-Whitney U-tests. They test the null hypothesis that the svo angles are the same in the respective conditions. b) The p-values result from Fisher's exact tests. They test the null hypothesis that the number of PCPs who are categorized as a specific svo type is independent of the condition. ($N = 20$)

Table 5.C.2: Effects of γ on information provision in condition P^H for PCPs without $\hat{\lambda} > 1$ and $\hat{\sigma} < 1$ ($N = 6$)

	low-quality information				high-quality information			
γ_1 to γ_2	Hyp. L	Hyp. NL	Δ	p-value	Hyp. L	Hyp. NL	Δ	p-value
0 - 1.25	0	0	−0033	1.000	+	+	0.017	0.750
1.25 - 2.5	+	+	0.733	0.031	0	-	0.000	–
2.5 - 6.25	0	-	0.000	1.000	0	+	0.050	0.500

Note: Differences in average provision of low and high-quality information between γ_1 and γ_2 for physicians without $\hat{\lambda} > 1$ and $\hat{\sigma} < 1$ ($N = 6$). Reported p-values result from two-sided Wilcoxon matched-pairs signed-rank tests using average values over all κ for subjects. They test the null hypothesis that information provision does not differ between γ_1 and γ_2. The columns "Hyp. L" and "Hyp. NL" summarize predictions from Hypotheses L and NL, respectively.

Table 5.C.3: Information provision

		All		$\gamma = 0$		$\gamma = 1.25$		$\gamma = 2.5$		$\gamma = 6.25$	
	N	Low	High	Low	High	Low	High	Low	High	Low	High
$\mathbf{P^H}$	20	0.531	0.261	0.175	0.210	0.115	0.275	0.755	0.220	0.680	0.320
$\mathbf{P^{CN}}$	20	0.562	0.213	0.095	0.165	0.275	0.200	0.760	0.205	0.750	0.225
$\mathbf{S^H}$	20	0.607	0.056	0.045	0.020	0.120	0.030	0.815	0.040	0.875	0.110
$\mathbf{S^L}$	20	0.484	0.318	0.095	0.280	0.230	0.260	0.600	0.345	0.635	0.365
$\mathbf{P^L}$	20	0.473	0.355	0.235	0.245	0.315	0.230	0.615	0.350	0.485	0.510

Note: Average share of low-quality and high-quality information provision for specific values of γ.

Summary and List of Publications

Summary

As patients' needs get increasingly complex such that multiple physicians are required for patient care, cooperation of physicians is increasingly important for the provision of high-quality and cost-efficient care. However, empirical evidence suggests that physician cooperation is deficient in practice. Specifically, some patients who require specialist care are not referred by their PCP (under-referral) and some patients who do not require specialist care are referred anyways (over-referral). Furthermore, the information exchange between physicians is flawed. Frequently, no information is sent at all, and if it is, it is often of low quality. Additionally, management of chronic diseases has been observed to often be improper, leading to avoidable future treatments.

This thesis analyzes physician cooperation with the help of theoretical models and a laboratory experiment. Based on this analysis, this thesis makes recommendations concerning the design of contracts in order to improve physician cooperation. This thesis is split up into three parts, their main results are outlined below.

The first part studies altruistic physicians who may prefer effective but expensive specialist treatment to more efficient PCP treatment. In order to align the incentives of physicians with the payer's aims, markups for immediate PCP treatment or cost sharing for specialist treatment can be employed. The main result of the first part is that incentives for specialists matter for the referral incentives of PCPs. Due to higher rent-efficiency, it may be preferable for the payer to employ cost sharing for the specialist rather than markups for the PCP in order to align both physicians' incentives simultaneously.

Part two compares two options for the organizational design of physicians—solo practices and a physician team—for the management of chronically ill patients by a PCP and a specialist. The main result is that solo practices can be superior to teams despite the existence of coordination problems between physicians. If the difference in treatment costs between patients of different disease severities is larger for the specialist than

for the PCP, treatment is inefficient in the team. By contrast, in the solo practices a gatekeeping outcome in which only severely ill patients are referred to the specialist can always be implemented. If, on the other hand, treatment cost differences are larger for the PCP, the team yields superior treatment outcomes. In this case, markups should be paid for PCP treatment and cost sharing should be imposed on specialist treatment in order to incentivize physician effort in the team.

Part three studies the information provision of a PCP who refers a patient to a specialist. She can decide whether to provide information and which information quality to provide. This part considers the impact of a bonus payment for information provision with the help of a theoretical model and a laboratory experiment. The bonus payment depends on whether information was provided but not on the quality of information, which is not verifiable by the payer. Without a bonus payment too little information is provided. Paying a bonus payment leads to an increase in both high- and low-quality information. Increasing the bonus payment from just covering the cost of low-quality information to at least covering the costs of high-quality information results in an increase in high-quality information and a decrease in low-quality information. This behavior can be explained by a model that considers both altruism and loss aversion. An additional result of part three is that PCPs act in a significantly more selfish manner if the specialist both benefits from the information provision instead of the patient and earns larger profits than the PCP.

Zusammenfassung

Die Behandlung von Patienten erfordert zunehmend die Anstrengungen mehrerer Ärzte. Daher gewinnt die zwischenärztliche Kooperation zunehmend an Bedeutung. Empirische Evidenz deutet jedoch darauf hin, dass Ärzte häufig in unzureichendem Maße kooperieren. Zum Beispiel werden Patienten, die eigentlich die Behandlung eines Facharztes bräuchten, nicht überwiesen. Zudem werden Patienten, die keine Überweisung bräuchten, trotzdem überwiesen. Außerdem ist der Informationsaustausch zwischen Ärzten mangelhaft. Oft wird keine Information mit der Überweisung gesandt, und wenn doch, ist sie häufig von niedriger Qualität. Zudem werden chronisch Kranke häufig nicht angemessen behandelt. Dies bedroht den Gesundheitszustand der Patienten und kann zu vermeidbaren Behandlungen führen.

Diese Dissertation analysiert die Kooperation von Ärzten mit Hilfe von theoretischen Modellen und einem Laborexperiment. Basierend auf dieser Analyse werden Empfehlungen für die optimale Vertragsgestaltung getroffen, um die zwischenärztliche Kooperation zu verbessern. Diese Dissertation ist in drei Teile aufgeteilt. Die Hauptergebnisse der Teile sind im Folgenden beschrieben.

Im ersten Teil werden altruistische Ärzte analysiert, die eine effektive aber teure Facharztbehandlung einer effizienteren Behandlung durch den Hausarzt vorziehen können. Um die Ärzte so anzureizen, dass sie sich den Präferenzen des Zahlers entsprechend verhalten, müssen entweder Aufschläge für die sofortige Behandlung durch den Hausarzt oder an den Facharzt für die Rücküberweisung zum Hausarzt gezahlt werden. Das zentrale Resultat des ersten Teils ist, dass Anreize für den Facharzt auch für den Hausarzt von Bedeutung sind. Aufgrund der höheren Renteneffizienz, kann es von Vorteil sein, Anreize für den Facharzt zu setzen und damit den Hausarzt indirekt anzureizen, anstatt den Hausarzt direkt anzureizen.

Der zweite Teil vergleicht zwei Optionen, um die Zusammenarbeit von einem Hausarzt und einem Facharzt bei der Behandlung chronisch Kranker zu organisieren – Einzel-

praxen und Gemeinschaftspraxen. Das zentrale Resultat ist, dass Einzelpraxen trotz Koordinationsproblemen den Gemeinschaftspraxen überlegen sein können. Wenn der Unterschied in den Behandlungskosten zwischen Patienten mit unterschiedlichem Schweregrad der Erkrankung für den Facharzt größer ist als für den Hausarzt, ist die Behandlung in der Gemeinschaftspraxis ineffizient. Im Gegensatz dazu kann in den Einzelpraxen ein gate keeping System implementiert werden, in dem nur schwer erkrankte Patienten zum Facharzt überwiesen werden. Wenn allerdings der Unterschied in den Behandlungskosten beim Hausarzt größer ist, liefert die Gemeinschaftspraxis bessere Behandlungsergebnisse. In diesem Fall sollten in der Gemeinschaftspraxis Aufschläge für die Behandlung des Hausarztes gezahlt werden. Zudem sollte der Facharzt unter seinen Kosten bezahlt werden, um Behandlungsanstrengungen anzureizen.

Der dritte Teil studiert die Informationsübertragung eines Hausarztes, der einen Patienten zum Facharzt überweist. Der Hausarzt kann entscheiden, ob und in welcher Qualität Informationen übertragen werden sollen. Die Auswirkungen einer Bonuszahlung für die Übertragung von Informationen wird mit Hilfe eines theoretischen Modells und eines Laborexperiments analysiert. Diese Bonuszahlung bezieht sich nur auf die Übertragung der Information, nicht aber auf deren Qualität, da diese für den Zahler nicht verifizierbar ist. Ohne Bonuszahlung werden zu wenige Informationen übertragen. Die Bonuszahlung führt zu einer höheren Bereitstellung von Information sowohl niedriger als auch hoher Qualität. Wenn die Bonuszahlung von einem Niveau, das gerade die Bereitstellungskosten der niedrigen Informationsqualität deckt auf ein Niveau, das über die Bereitstellungskosten von hochqualitativer Information hinausgeht, erhöht wird, resultiert dies in einer Verbesserung der durchschnittlichen Informationsqualität. Dieses Verhalten kann innerhalb eines Modellrahmens erklärt werden, der sowohl Altruismus als auch Verlustaversion der Ärzte berücksichtigt. Ein weiteres Resultat von Teil drei ist, dass Hausärzte weitaus eigennütziger agieren, wenn der Facharzt sowohl von der Informationsübertragung profitiert als auch höhere Profite erzielt als der Hausarzt.

List of Publications

1. Griebenow, Malte and Kifmann, Mathias (2021). Diagnostics and Treatment: On the Division of Labor between Primary Care Physicians and Specialists, HCHE Research Paper No. 24, Universität Hamburg.

References

Abdellaoui, Mohammed, Bleichrodt, Han, and Paraschiv, Corina (2007). "Loss Aversion Under Prospect Theory: A Parameter-Free Measurement". In: *Management Science* 53.10, pp. 1659–1674. DOI: 10.1287/mnsc.1070.0711.

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.

Allard, Marie, Léger, Pierre Thomas, and Rochaix, Lise (2009). "Provider Competition in a Dynamic Setting". In: *Journal of Economics & Management Strategy* 18.2, pp. 457–486. DOI: 10.1111/j.1530-9134.2009.00220.x.

Arrow, K.J. (1963). "Uncertainty and the Welfare Economics of Medical Care". In: *American Economic Review* 53, pp. 941–973.

Bakker, Dinny H. de, Struijs, Jeroen N., Baan, Caroline A., Raams, Joop, Wildt, Jan-Erik de, Vrijhoef, Hubertus J.M., and Schut, Frederik T. (2012). "Early Results From Adoption Of Bundled Payment For Diabetes Care In The Netherlands Show Improvement In Care Coordination". In: *Health Affairs* 31.2, pp. 426–433. DOI: 10.1377/hlthaff.2011.0912.

Barham, Victoria and Milliken, Olga (2015). "Payment Mechanisms and the Composition of Physician Practices: Balancing Cost-Containment, Access, and Quality of Care". In: *Health Economics* 24.7, pp. 895–906. DOI: 10.1002/hec.3069.

Barnett, Karen, Mercer, Stewart W, Norbury, Michael, Watt, Graham, Wyke, Sally, and Guthrie, Bruce (2012). "Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study". In: *The Lancet* 380.9836, pp. 37–43. DOI: 10.1016/s0140-6736(12)60240-2.

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.

Bleichrodt, Han, Pinto, Jose Luis, and Wakker, Peter P. (2001). "Making Descriptive Use of Prospect Theory to Improve the Prescriptive Use of Expected Utility". In: *Management Science* 47.11, pp. 1498–1514. DOI: 10.1287/mnsc.47.11.1498.10248.

Blümel, Miriam and Busse, Reinhard (2015). *The German Health Care System*. URL: https://international.commonwealthfund.org/countries/germany/.

Bodenheimer, Thomas (2008). "Coordinating Care — A Perilous Journey through the Health Care System". In: *New England Journal of Medicine* 358.10, pp. 1064–1071. DOI: 10.1056/nejmhpr0706165.

Bodenheimer, Thomas, Chen, Ellen, and Bennett, Heather D. (2009). "Confronting The Growing Burden Of Chronic Disease: Can The U.S. Health Care Workforce Do The Job?" In: *Health Affairs* 28.1, pp. 64–74. DOI: 10.1377/hlthaff.28.1.64.

Bodenheimer, Thomas, Wagner, Edward H, and Grumbach, Kevin (2002). "Improving primary care for patients with chronic illness." In: *JAMA* 288 (14), pp. 1775–1779.

Bolton, Gary E and Ockenfels, Axel (2000). "ERC: A Theory of Equity, Reciprocity, and Competition". In: *American Economic Review* 90.1, pp. 166–193. DOI: 10.1257/aer.90.1.166.

Bolton, Patrick and Dewatripont, Mathias (2005). *Contract Theory*. Cambridge, Mass: MIT Press.

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.

Brosig, Jeannette (2002). "Identifying cooperative behavior: some experimental results in a prisoner's dilemma game". In: *Journal of Economic Behavior & Organization* 47.3, pp. 275–290. DOI: 10.1016/S0167-2681(01)00211-6.

Brosig-Koch, Jeannette, Hennig-Schmidt, Heike, Kairies-Schwarz, Nadja, and Wiesen, Daniel (2016a). *Physician performance pay: Evidence from a laboratory experiment.* eng. Ruhr Economic Papers 658. Essen. DOI: 10.4419/86788764.

— (2016b). "Using artefactual field and lab experiments to investigate how fee-for-service and capitation affect medical service provision". In: *Journal of Economic Behavior & Organization* 131, pp. 17–23. DOI: 10.1016/j.jebo.2015.04.011.

— (2017). "The Effects of Introducing Mixed Payment Systems for Physicians: Experimental Evidence". In: *Health Economics* 26, pp. 243–262. DOI: 10.1002/hec.3292.

Bruin, Simone R. de, Heijink, Richard, Lemmens, Lidwien C., Struijs, Jeroen N., and Baan, Caroline A. (2001). "Impact of disease management programs on healthcare expenditures for patients with diabetes, depression, heart failure or chronic obstructive pulmonary disease: A systematic review of the literature". In: *Health Policy* 101.2, pp. 105–121. DOI: 10.1016/j.healthpol.2011.03.006.

Buttorff, Christine (2017). *Multiple chronic conditions in the United States.* Santa Monica, CA: RAND.

Centers for Medicare & Medicaid Services (2014). *Policy and payment changes to the Medicare Physician Fee Schedule for 2015.* URL: https://www.cms.gov/newsroom/fact-sheets/policy-and-payment-changes-medicare-physician-fee-schedule-2015 (visited on 09/30/2021).

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.

Choné, Philippe and Ma, Ching-to A. (2011). "Optimal Health Care Contract under Physician Agency". In: *Annals of Economics and Statistics* 101/102, pp. 229–256. DOI: 10.2307/41615481.

Civaner, M. Murat, Balcioglu, Harun, and Vatansever, Kevser (2016). "Medical Students' Opinions About the Commercialization of Healthcare: A Cross-Sectional Survey". In: *Journal of Bioethical Inquiry* 13.2, pp. 261–270. DOI: 10.1007/s11673-016-9704-6.

Clemens, Jeffrey and Gottlieb, Joshua D. (2014). "Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health?" In: *American Economic Review* 104.4, pp. 1320–1349. DOI: 10.1257/aer.104.4.1320.

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

Dohmen, Thomas, Falk, Armin, Huffman, David, Sunde, Uwe, Schupp, Jürgen, and Wagner, Gert G. (2011). "INDIVIDUAL RISK ATTITUDES: MEASUREMENT, DETERMINANTS, AND BEHAVIORAL CONSEQUENCES". In: *Journal of the European Economic Association* 9.3, pp. 522–550. DOI: 10.1111/j.1542-4774.2011.01015.x.

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. DOI: 10.1016/j.jhealeco.2005.08.001. ppublish.

Dusheiko, Mark, Gravelle, Hugh, Martin, Stephen, Rice, Nigel, and Smith, Peter C. (2011). "Does better disease management in primary care reduce hospital costs? Evidence from English primary care". In: *Journal of Health Economics* 30.5, pp. 919–932. DOI: 10.1016/j.jhealeco.2011.08.001.

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.

Eggleston, K. (2000). "Risk Selection and Optimal Health Insurance-Provider Payment Systems". In: *Journal of Risk and Insurance* 67, pp. 173–196.

— (2005). "Multitasking and Mixed Systems for Provider Payment". In: *Journal of Health Economics* 24, pp. 211–223. DOI: 10.1016/j.jhealeco.2004.09.001.

Einav, Liran, Finkelstein, Amy, and Mahoney, Neale (2018). "Provider Incentives and Healthcare Costs: Evidence From Long-Term Care Hospitals". In: *Econometrica* 86.6, pp. 2161–2219. DOI: 10.3982/ecta15022.

Ellis, Randall P. (1998). "Creaming, Skimping and Dumping: Provider Competition on the Intensive and Extensive Margins". In: *Journal of Health Economics* 17, pp. 537–555.

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.

Epstein, Andrew J. and Johnson, Scott J. (2012). "Physician response to financial incentives when choosing drugs to treat breast cancer". In: *International Journal of Health Care Finance and Economics* 12.4, pp. 285–302. DOI: 10.1007/s10754-012-9117-y.

European Values Study (2015). *European Values Study Longitudinal Data File 1981-2008 (EVS 1981-2008)*. ZA4804 Data file Version 3.0.0. Cologne, Germany: GESIS Data Archive. DOI: 10.4232/1.12253.

Falk, Armin, Becker, Anke, Dohmen, Thomas, Enke, Benjamin, Huffman, David, and Sunde, Uwe (2018). “Global Evidence on Economic Preferences”. In: *The Quarterly Journal of Economics* 133.4, pp. 1645–1692. DOI: 10.1093/qje/qjy013.

Fehr, E. and Schmidt, K. M. (1999). “A Theory of Fairness, Competition, and Cooperation”. In: *The Quarterly Journal of Economics* 114.3, pp. 817–868. DOI: 10.1162/003355399556151.

Filar, Jerzy and Vrieze, Koos (1996). *Competitive Markov Decision Processes*. Springer New York. DOI: 10.1007/978-1-4612-4054-9.

Fischbacher, Urs (2007). “z-Tree: Zurich toolbox for ready-made economic experiments”. In: *Experimental Economics* 10.2, pp. 171–178. DOI: 10.1007/s10683-006-9159-4.

Franks, P., Zwanziger, J., Mooney, C., and Sorbero, M. (1999). “Variations in primary care physician referral rates.” In: *Health services research* 34 (1 Pt 2), pp. 323–329. ppublish.

Frechette, Guillaume and Andrew Schotter, eds. (2015). *Handbook of Experimental Economic Methodology*. Oxford University Press.

Gaechter, Simon, Johnson, Eric, and Herrmann, Andreas (2010). *Individual-level loss aversion in riskless and risky choices*. Discussion Papers 2010-20. The Centre for Decision Research and Experimental Economics, School of Economics, University of Nottingham.

Galizzi, Matteo M. and Wiesen, Daniel (2018). *Behavioral Experiments in Health Economics*. Oxford Research Encyclopedia of Economics and Finance. Oxford University Press. DOI: 10.1093/acrefore/9780190625979.013.244.

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

Gask, Linda (2005). “Role of specialists in common chronic diseases”. In: *BMJ* 330.7492, pp. 651–653. DOI: 10.1136/bmj.330.7492.651.

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.

Greiner, Ben (2015). "Subject pool recruitment procedures: Organizing experiments with ORSEE". In: *Journal of the Economic Science Association* 1.1, pp. 114–125. DOI: 10.1007/s40881-015-0004-4.

Hafsteinsdottir, Elin J.G. and Siciliani, Luigi (2010). "DRG Prospective Payment Systems: Refine or not Refine?" In: *Health Economics* 19, pp. 1226–1239. DOI: 10.1002/hec.1547.

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.

Heider, Ann-Kathrin and Mang, Harald (2020). "Effects of Monetary Incentives in Physician Groups: A Systematic Review of Reviews". In: *Applied Health Economics and Health Policy* 18.5, pp. 655–667. DOI: 10.1007/s40258-020-00572-x.

Hennessy, David A. (2008). "Prevention and cure efforts both substitute and complement". In: *Health Economics* 17.4, pp. 503–511. DOI: 10.1002/hec.1270.

Hennig-Schmidt, Heike, Selten, Reinhard, and Wiesen, Daniel (2011). "How payment systems affect physicians' provision behaviour—An experimental investigation". In: *Journal of Health Economics* 30.4, pp. 637–646. DOI: 10.1016/j.jhealeco.2011.05.001.

Hennig-Schmidt, Heike and Wiesen, Daniel (2014). "Other-regarding behavior and motivation in health care provision: An experiment with medical and non-medical students". In: *Social Science & Medicine* 108, pp. 156–165.

Hey, John D. and Patel, Mahesh S. (1983). "Prevention and cure? : Or: Is an ounce of prevention worth a pound of cure?" In: *Journal of Health Economics* 2.2, pp. 119–138. DOI: 10.1016/0167-6296(83)90002-4.

Hothorn, Torsten, Hornik, Kurt, van de Wiel, Mark A., and Zeileis, Achim (2008). "Implementing a class of permutation tests: The coin package". In: *Journal of Statistical Software* 28.8, pp. 1–23. DOI: 10.18637/jss.v028.i08.

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.

Inglehart, R. et al. (2014). *World Values Survey: Round Six - Country-Pooled Datafile Version.* Tech. rep. Madrid, Spain: JD Systems Institute. URL: www.worldvaluessurvey.org/WVSDocumentationWV6.jsp (visited on 10/12/2021).

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

Jack, W. (2005). "Purchasing Health Care Services from Providers with Unknown Altruism". In: *Journal of Health Economics* 24, pp. 73–93.

Jelovac, Izabela and Macho-Stadler, Inés (2002). "Comparing organizational structures in health services". In: *Journal of Economic Behavior & Organization* 49.4, pp. 501–522. DOI: 10.1016/s0167-2681(02)00008-2.

Jones, R. (2002). "Declining altruism in medicine". In: *BMJ* 324.7338, pp. 624–625. DOI: 10.1136/bmj.324.7338.624.

Kaarbøe, Oddvar and Siciliani, Luigi (2011). "Multi-Tasking, Quality and Pay for Performance". In: *Health Economics* 20, pp. 225–238. DOI: 10.1002/hec.1582.

Kahneman, D. and Tversky, A. (1979). "Prospect Theory: An Analysis of Decision Under Risk". In: *Econometrica* 47, pp. 263–291.

Kash, Bita and Tan, Debra (2016). "Physician Group Practice Trends: A Comprehensive Review". In: *J Hosp Med Manage* 2.1.

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.

Kesternich, Iris, Schumacher, Heiner, and Winter, Joachim (2015). "Professional norms and physician behavior: Homo oeconomicus or homo hippocraticus?" In: *Journal of Public Economics* 131, pp. 1–11. DOI: 10.1016/j.jpubeco.2015.08.009.

King, Dana E., Xiang, Jun, and Pilkerton, Courtney S. (2018). "Multimorbidity Trends in United States Adults, 1988–2014". In: *The Journal of the American Board of Family Medicine* 31.4, pp. 503–513. DOI: 10.3122/jabfm.2018.04.180008.

Kuehberger, Anton (1998). "The Influence of Framing on Risky Decisions: A Meta-analysis". In: *Organizational Behavior and Human Decision Processes* 75.1, pp. 23–55. DOI: 10.1006/obhd.1998.2781.

Lemieux-Charles, Louise and McGuire, Wendy L. (2006). "What Do We Know about Health Care Team Effectiveness? A Review of the Literature". In: *Medical Care Research and Review* 63.3, pp. 263–300. DOI: 10.1177/1077558706287003.

Li, Jing, Dow, William H., and Kariv, Shachar (2017). "Social preferences of future physicians". In: *Proceedings of the National Academy of Sciences* 114.48, E10291–E10300. DOI: 10.1073/pnas.1705451114.

Li, R., Zhang, P., Barker, L. E., Chowdhury, F. M., and Zhang, X. (2010). "Cost-Effectiveness of Interventions to Prevent and Control Diabetes Mellitus: A Systematic Review". In: *Diabetes Care* 33.8, pp. 1872–1894. DOI: 10.2337/dc10-0843.

Liebrand, Wim B. G. (1984). "The effect of social motives, communication and group size on behaviour in an N-person multi-stage mixed-motive game". In: *European Journal of Social Psychology* 14.3, pp. 239–264. DOI: 10.1002/ejsp.2420140302.

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.

Ma, Ching-to A. and Mak, Henry Y. (2019). *Incentives in Healthcare Payment Systems.* DOI: 10.1093/acrefore/9780190625979.013.61.

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

— (2005). "Supplier discretion over provision: theory and an application to medical care". In: *RAND Journal of Economics* 36.2, pp. 412–432. URL: https://www.jstor.org/stable/4135249.

MATLAB (2016). *Version 9.0.0 (R2016b).* Natick, Massachusetts: The MathWorks Inc.

McClintock, Charles G. and Liebrand, Wim B. G. (1988). "Role of Interdependence Structure, Individual Value Orientation, and Another's Strategy in Social Decision Making: A Transformational Analysis". In: *Journal of Personality and Social Psychology* 55.3, pp. 396–409.

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.

Navickas, Rokas, Petric, Vesna-Kerstin, Feigl, Andrea B., and Seychell, Martin (2016). "Multimorbidity: What Do We Know? What Should We Do?" In: *Journal of Comorbidity* 6.1, pp. 4–11. DOI: 10.15256/joc.2016.6.72.

OECD (2015). *Health at a Glance 2015*. OECD Publishing. DOI: 10.1787/health_glance-2015-en.

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.

Pascucci, Domenico, Sassano, Michele, Nurchis, Mario Cesare, Cicconi, Michela, Acampora, Anna, Park, Daejun, Morano, Carmen, and Damiani, Gianfranco (2020). "Impact of interprofessional collaboration on chronic disease management: findings from a systematic review of clinical trial and meta-analysis". In: *Health Policy*. DOI: 10.1016/j.healthpol.2020.12.006.

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.

Perry, Joshua E. (2009). "The Ethical Costs of Commercializing the Professions: First-Person Narratives from the Legal and Medical Trenches". In: *University of Pennsylvania Journal of Law and Social Change* 13, p. 169.

Practice Fusion (2010). *Survey: Patients See 18.7 Different Doctors on Average*. URL: https://www.prnewswire.com/news-releases/survey-patients-see-187-different-doctors-on-average-92171874.html (visited on 09/30/2021).

Puri, Madan Lal and Sen, Pranab Kumar (1966). "On a Class of Multivariate Multisample Rank-Order Tests". In: *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)* 28.4, pp. 353–376.

— (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley publication in mathematical statistics. New York: John Wiley & Sons.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Relman, Arnold S. (2007). “Medical Professionalism in a Commercialized Health Care Market”. In: *JAMA* 298.22, p. 2668. DOI: 10.1001/jama.298.22.2668.

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.

Schubert, Ingrid, Siegel, Achim, Köster, Ingrid, and Ihle, Peter (2016). “Evaluation der populationsbezogenen ‚Integrierten Versorgung Gesundes Kinzigtal‘ (IVGK). Ergebnisse zur Versorgungsqualität auf der Basis von Routinedaten”. In: *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen* 117, pp. 27–37. DOI: 10.1016/j.zefq.2016.06.003.

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.

Song, Zirui, Sequist, Thomas D., and Barnett, Michael L. (2014). “Patient Referrals”. In: *JAMA* 312.6, p. 597. DOI: 10.1001/jama.2014.7878.

Starfield, Barbara and Shi, Leiyu (2002). “Policy relevant determinants of health: an international perspective”. In: *Health Policy* 60.3, pp. 201–218. DOI: 10.1016/s0168-8510(01)00208-1.

Tanaka, Tomomi, Camerer, Colin F., and Nguyen, Quang (2010). “Risk and Time Preferences: Linking Experimental and Household Survey Data from Vietnam”. In: *American Economic Review* 100.1, pp. 557–71. DOI: 10.1257/aer.100.1.557.

Tversky, A. and Kahneman, D. (1991). “Loss Aversion in Riskless Choice: A Reference-Dependent Model”. In: *The Quarterly Journal of Economics* 106.4, pp. 1039–1061. DOI: 10.2307/2937956.

Waibel, Christian and Wiesen, Daniel (2021). “An experiment on referrals in health care”. In: *European Economic Review* 131, p. 103612. DOI: 10.1016/j.euroecorev.2020.103612.

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.

Whynes, David K, Ennew, Christine T, and Feighan, Teresa (1999). “Entrepreneurial attitudes of primary health care physicians in the United Kingdom”. In: *Journal of Economic Behavior & Organization* 38.3, pp. 331–347. DOI: 10.1016/s0167-2681(99)00013-x.

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.