Brosig-Koch, Jeannette; Groß, Mona; Hennig-Schmidt, Heike; Kairies-Schwarz, Nadja; Wiesen, Daniel

Working Paper

Physicians' incentives, patients' characteristics, and quality of care: A systematic experimental comparison of fee-for-service, capitation, and pay for performance

Ruhr Economic Papers, No. 923

Provided in Cooperation with:
RWI – Leibniz-Institut für Wirtschaftsforschung, Essen

Suggested Citation: Brosig-Koch, Jeannette; Groß, Mona; Hennig-Schmidt, Heike; Kairies-Schwarz, Nadja; Wiesen, Daniel (2021) : Physicians' incentives, patients' characteristics, and quality of care: A systematic experimental comparison of fee-for-service, capitation, and pay for performance, Ruhr Economic Papers, No. 923, ISBN 978-3-96973-079-9, RWI - Leibniz-Institut für Wirtschaftsforschung, Essen,
https://doi.org/10.4419/96973079

This Version is available at:
https://hdl.handle.net/10419/243148

RUHR
ECONOMIC PAPERS

Jeannette Brosig-Koch
Mona Groß
Heike Hennig-Schmidt
Nadja Kairies-Schwarz
Daniel Wiesen

Physicians' Incentives, Patients' Characteristics, and Quality of Care: A Systematic Experimental Comparison of Fee-for-Service, Capitation, and Pay For Performance

UNIVERSITÄT
DUISBURG
ESSEN
Offen im Denken

#923

Imprint

Ruhr Economic Papers

Published by

RWI – Leibniz-Institut für Wirtschaftsforschung
Hohenzollernstr. 1-3, 45128 Essen, Germany

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Ruhr Economic Papers #923

Responsible Editor: Volker Clausen

All rights reserved. Essen, Germany, 2021

ISSN 1864-4872 (online) – ISBN 978-3-96973-079-9

The working papers published in the series constitute work in progress circulated to stimulate discussion and critical comments. Views expressed represent exclusively the authors' own opinions and do not necessarily reflect those of the editors.

Ruhr Economic Papers #923

Jeannette Brosig-Koch, Mona Groß, Heike Hennig-Schmidt,
Nadja Kairies-Schwarz, and Daniel Wiesen

Physicians' Incentives, Patients' Characteristics, and Quality of Care: A Systematic Experimental Comparison of Fee-for-Service, Capitation, and Pay For Performance

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

Bibliografische Informationen der Deutschen Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at **http://dnb.dnb.de**

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.

http://dx.doi.org/10.4419/96973079
ISSN 1864-4872 (online)
ISBN 978-3-96973-079-9

Jeannette Brosig-Koch, Mona Groß, Heike Hennig-Schmidt,
Nadja Kairies-Schwarz, and Daniel Wiesen[1]

Physicians' Incentives, Patients' Characteristics, and Quality of Care: A Systematic Experimental Comparison of Fee-for-Service, Capitation, and Pay For Performance

Abstract

This paper systematically studies how performance pay, complementing either baseline fee-for-service or capitation, affects physicians' medical service provision and the quality of care. Using a series of controlled experiments with physicians and students, we test the incentive effect of performance pay at a within-subject level. A discrete bonus is granted if a quality threshold is reached, which varies with the patients' severity of illness. We find that performance pay significantly reduces non-optimal service provision and enhances the quality of care. Effect sizes depend on the patients' severity of illness and whether the baseline is fee-for-service or capitation. Health policy implications, including a cost benefit analysis of introducing performance pay, are discussed.

JEL-Code: C91, I11

Keywords: Fee-for-service; capitation; pay for performance; heterogeneous patients; artefactual field experiment; laboratory experiments

September 2021

[1] Jeannette Brosig-Koch: Otto von Guericke University Magdeburg and Health Economics Research Center (CINCH) Essen; Mona Groß: University of Cologne; Heike Hennig-Schmidt: University of Bonn; Nadja Kairies-Schwarz: Heinrich Heine University Duesseldorf and CINCH Essen; Daniel Wiesen: University of Cologne. - We thank Dr. Harald Herholz (Association of Statutory Health Insurance Physicians in Hesse, Germany), Rainer Diehl, and Aline Zetsche for supporting the field experiments at the Akademie für Ärztliche Fort- und Weiterbildung, Bad Nauheim, Germany. We are grateful for valuable comments and suggestions from Josse Delfgaauw, Robert Dur, Glenn Harrison, Tor Iversen, Albert Ma, and Luigi Siciliani, and seminar participants at Berlin, Hamburg, Rotterdam, Mainz, Oslo, Tutzingen, Wuppertal, the 3rd BEH-Net Workshop in Essen, COPE Vienna, ESA Cologne and New York, GfeW Hamburg, EuHEA Hamburg, and CAS Workshop Munich. We also thank Andrew Kinder, Johanna Kokot, Emanuel Castillo Ruiz, and Bastian Zöller for their excellent research assistance. Financial support by the German Research Foundation (DFG grant: BR 2346/2-1/2) and the German Federal Ministry of Education and Research (BMBF, grant: 01EH1602A) is gratefully acknowledged. Hennig-Schmidt acknowledges a Käte Hamburger Kolleg fellowship. Hennig-Schmidt and Wiesen are grateful for financial support from the Research Council of Norway (grant: 231776 IRECOHEX) during their employment at the University of Oslo, Norway. This paper is a completely revised version of the manuscript "Physician performance pay: Evidence from a laboratory experiment" (Ruhr Economic Paper No. 658). – All correspondence to: Mona Groß, University of Cologne, Department of Health Care Management, e-mail: mona.gross@wiso.uni-koeln.de

1 Introduction

Paying physicians for performance has become prominent among health policy-makers, for example in the USA (e.g., Rosenthal et al., 2006; Stokes et al., 2018; Song et al., 2019) and in the UK (see, e.g., Roland, 2004; Doran et al., 2006; Roland and Campbell, 2014). Performance pay (P4P) is usually granted if a quality threshold is reached. Traditional physician payment systems are lump-sum capitation (CAP) or fee-for-service (FFS), in which physicians receive a fee for each service provided, with FFS typically being used in specialty care and CAP being prevalent in primary care (for Germany, see Brosig-Koch et al. 2020). These systems, generally, are not tied to the quality of care provided. FFS incentivizes physicians to overserve patients, whereas CAP embeds an incentive to underserve them. Thus, paying physicians on the basis of direct performance measures has attracted particular attention.

In health care, P4P typically complements either FFS or CAP. From a theoretical point of view, blending P4P with FFS (FFS+P4P) is likely to affect physicians' medical service provision differently compared to P4P blended with capitation (CAP+P4P), due to the different incentives of the baseline payment systems. A systematic comparison of the effectiveness of P4P between CAP and FFS based on comparable designs is lacking. Also, it is not well understood how patients with different severities of illness are affected by incentives of the P4P systems. The heterogeneous impact of payment incentives on different patient types has been indicated in recent empirical (e.g., Clemens and Gottlieb, 2014) and experimental studies (e.g., Hennig-Schmidt et al., 2011; Brosig-Koch et al., 2017a).

The empirical evidence on whether, and, if so, how P4P affects physicians' medical service provision and quality of care, is rather mixed (e.g., Scott et al., 2011; Emmert et al., 2012; Eijkenaar et al., 2013; Milstein and Schreyögg, 2016). Moreover, it has been argued that the *design* of a P4P system is key to effectively changing physician behavior (Epstein, 2012; Maynard, 2012; Kristensen et al., 2016; Anselmi et al., 2020). Potential reasons for the difficulty in establishing a causal link between performance pay and physicians' provision behavior comprise the likely endogeneity of institutions (e.g., Baicker and Goldman, 2011), the biased and incomplete performance measures (e.g., Mullen et al., 2010), measurement errors (e.g., Campbell et al., 2009), the limited availability of data (e.g., Gravelle et al., 2010; Maynard, 2012), and introduction of P4P in parallel to other interventions (e.g., Lindenauer et al., 2007).

Our study contributes to better understanding the effects of different P4P systems on the quantity and quality of care. To this end, we designed a controlled behavioral experiment, in which the physicians' financial incentives in baseline FFS and CAP are mirror images of each other. We complement FFS and CAP with a discrete bonus that is kept constant across both payment systems FFS+P4P and CAP+P4P. The bonus is paid when a quality threshold tied to a patient's optimal health outcome is reached. Meeting the quality threshold still allows for over- and underprovision, as we assume asymmetric information between the physician and the payer. Service provision according to the threshold thus might increase the physician's profit while still not providing the optimal care. This mirror design allows us to systematically compare the two blended payment schemes (FFS+P4P and CAP+P4P) – an analysis that is currently missing in the literature. In addition, we keep the patient population constant. Physicians are confronted with identical patients regarding their severities of illness and their marginal health benefit from each medical service provided. This feature allows us to investigate systematically whether the effects of P4P are specific to patients' illnesses and severities of illness despite the mirror design of the payment systems. Finally, we also consider health policy implications, including cost-benefit analyses, for our experimental design of performance incentives.

Our experimental design is well grounded in theory. Behavioral predictions are derived from an illustrative model and are tested with physicians in lab-in-the field conditions and with medical and non-medical students in lab experiments. To establish the causal link between P4P and the quantity and quality of medical service provision, we exogenously vary physicians' remuneration at a within-subject level from the baseline non-blended payment schemes to the blended performance-pay systems. In a medically framed decision situation, subjects decide on the quantity of medical services for abstracted patients with different severities of illness (mild, intermediate, severe) and marginal health benefits (low, high). Quantity choices determine the physicians' own profit and the patients' health benefits measured in monetary terms. Participants are informed that their decisions affect the health of real-world patients, as the money corresponding to the aggregated health benefits is transferred to a charity and is used exclusively for surgery of cataract patients. For an analogous procedure, see, for example, Hennig-Schmidt et al. (2011), Brosig-Koch et al. (2016, 2017a, 2020), and Waibel and Wiesen (2020).

With our parsimonious experiment, we address the following research questions. First, we analyze how the effect of introducing P4P affects medical service provision and the quality of care when complementing FFS. Second, we study whether such an effect is specific to the patients' characteristics such as the severity of illness and the marginal health benefit. Third, we explore the effect of P4P blended with baseline CAP, and fourth, we investigate the effect differences that are due to the patients' health characteristics. Finally, in a joint analysis, we examine whether potential differences in subjects' reactions to FFS and CAP exist and whether the P4P effect varies between FFS+P4P and CAP+P4P, despite the mirror-image design of financial incentives.

Our behavioral results indicate that the introduction of P4P reduces non-optimal service provision, enhances the quality of care, and patients' health benefit under FFS and under CAP. We find that the effects of P4P are specific to the patients' severities of illness. Under FFS, the marginal benefit of P4P on medical service provision, the quality of care as well as the patients' health benefit decreases in patients' severity of illness. Under CAP, we observe the reverse pattern: the marginal benefit of P4P increases with increasing severity. In other words, our behavioral results indicate that the introduction of pay for performance is most beneficial for mildly ill patients under FFS, whereas it is most beneficial for highly ill patients under CAP. Patients of intermediate severity of illness are almost equally treated under both performance-pay systems.

While our results suggest that P4P serves as a means to counteract misaligned financial incentives for overprovision under FFS and underprovision under CAP, they also emphasize the importance of its design elements. Utilizing the symmetric design components across baseline payment conditions, we are also able to analyze the cost and benefits of introducing P4P under both payment conditions and derive health-policy implications. In sum, health policy-makers need to take into account that the effectiveness of P4P is specific to a patient's severity of illness and the underlying baseline payment condition when designing P4P systems.

We contribute to several streams in the health economics literature. First, we complement the empirical literature that analyzes the effects of P4P on physicians' treatment decisions. Quite often, P4P programs are evaluated using administrative longitudinal data. Empirical evidence is rather mixed, showing only modest positive effects (if at all) on the quality of medical service provision; for extensive literature reviews, see Scott et al. (2011) for primary care, Jia et al. (2021) for general outpatient care, and Mathes et al. (2019) for inpatient care. Mullen et al. (2010), for example, using longitudinally data from quarterly performance reports, find only little empirical support for a positive effect of introducing P4P on process quality of multi-specialty medical groups in the US. Studies mostly evidence some increase in a few clinical processes; yet, the P4P effects on outcome quality are not clear (e.g., Peckham and Wallace, 2010; Li et al., 2014). While empirical studies typically rely on aggregated data, we add insights on a causal effect of P4P at the individual subject level. The highly controlled environment in our experiment allows us to implement 'clean' measures for the quality of medical service provision of the individual physicians. It also enables us to analyze systematically how variations in patients' health characteristics and payment systems (CAP versus FFS) relate to the effect of P4P.

Second, our study contributes to the scarce experimental literature analyzing performance pay for physicians by means of controlled behavioral experiments. These studies provide first evidence for a positive effect of P4P on physicians' treatment behavior (e.g., Brosig-Koch et al., 2020; Oxholm et al., 2021). Our study differs, however, from this literature by systematically analyzing the effects of FFS *and* CAP as well as of the respective blended P4P systems. The study by Brosig-Koch et al. (2020), with a representative primary-care physician sample, investigates the effect of a threshold-based P4P system with a discrete bonus blended with CAP, analogously to our CAP+P4P condition. A lab experiment with Danish medical students by Oxholm et al. (2021) shows that P4P affects the allocation of medical care across patients with low and high responsiveness to treatment compared to lump-sum CAP payments. Considering FFS and P4P, Keser et al. (2014) report from a laboratory experiment with German medical students that a bonus tied to the share of optimally treated patients leads to some increase in the quality of care. In a lab experiment with US medical students, Cox et al. (2016b) find that utilizing P4P mechanisms incentivizes cost-effective reductions in hospital re-admissions.

Third, we add to the literature on behavioral experiments in health (Galizzi and Wiesen, 2017, 2018), focusing on incentives and physician behavior. In particular, our study complements experiments in a medical framework analyzing the effects of financial incentives on physician behavior, such as FFS or CAP (Hennig-Schmidt et al., 2011; Hennig-Schmidt and Wiesen, 2014; Green, 2014; Brosig-Koch et al., 2016; Lagarde and Blaauw, 2017; Di Guida et al., 2019; Martinsson and Persson, 2019; Reif et al., 2020) and blended payment systems (Brosig-Koch et al., 2017a, 2020). In a broader sense, we also relate to the experimental literature on credence goods markets, which typically apply neutral framings but for which health care characterized by high information asymmetries is a key example (e.g., Dulleck and Kerschbamer, 2006; Dulleck et al., 2011).[1] Our study adds causal evidence on how, in P4P systems, physician behavior is affected by the baseline payment (FFS versus CAP), how patient characteristics influence treatment decisions, and which design features of a payment system could potentially be implemented to enhance the quality of care for different patient types.

This paper proceeds as follows. Section 2 describes our experimental design and behavioral hypotheses. Section 3 presents the behavioral results and Section 4 discusses implications for cost and benefits

[1] Medical services are considered as credence goods due to high informational advantages of physicians towards their patients. This enables physicians to exploit their patients, for example through overtreatment under FFS. In our experimental design, we incorporate the 'credence goods' problematic by assuming that our patients are passive and accept each quantity of medical services provided by the physician. Typically applying neutral framings, experiments in the credence goods literature showed that overtreatment can be reduced by costly second opinions (Mimra et al., 2016), competition (Huck et al., 2016), and separating treatment from diagnosis and prescription decisions (Greiner et al., 2017). Recent experiments show that monitoring mechanisms with financial consequences reduce overtreatment and the overcharging of patients (Angerer et al., 2021; Hennig-Schmidt et al., 2019; and Groß et al., 2021). We complement these experiments by investigating whether performance-based financial incentives which implicitly rely on monitoring a physician's performance are capable of coping with non-optimal medical service provision such as overtreatment under FFS.

within the confines of our experiment. Section 5 concludes.

2 Experimental design, protocol, and hypotheses

2.1 Decision situation

Our experiment employs a medical frame. All subjects decide in the role of physicians on the provision of medical services. We employ a within-subject design to analyze the effect of P4P on physicians' provision of medical services. To this end, each subject makes his or her decisions under non-blended and blended payments. First, subjects are incentivized either by FFS or by CAP, which serve as baseline payments. Second, we introduce physicians' P4P in addition to the respective baseline payments (FFS+P4P or CAP+P4P). We randomly assign subjects to one of the two experimental conditions.[2]

In all payment systems, physician i decides on the quantity of medical services $q \in [0, 10]$ for nine different patients ($j = 1, \ldots, 9$). Patients differ in illnesses $k \in \{A, B, C\}$ and in the severity of illnesses $l \in \{x, y, z\}$. Patients are assumed to be passive and fully insured, accepting each quantity of medical services provided by the physician. This is a common assumption in the theoretical health economics literature (for a comprehensive review, for example, see McGuire, 2000), corresponding to the assumption of information asymmetry between expert (physician) and customer (patient) in the credence goods literature (e.g., Dulleck and Kerschbamer, 2006). In our experiment, patients' characteristics are the same in all payment conditions. The patient population for which a physician chooses services thus remains constant.

Physician i's payment is $R(q) = L + pq + b_l I_{b_l}$, with L being the lump-sum payment, p the fee per service rendered to a patient, and b_l the bonus payment; I_{b_l} denotes an indicator variable which equals 1, if the physician's chosen quantity does not differ by more than one unit from the patient's optimal treatment, and 0 otherwise. In FFS, $L = 0$ and $b_l^{\text{FFS}} = 0$ and in CAP $p = 0$, and $b_l^{\text{CAP}} = 0$.

Physician i's profit is given as

$$\pi(q) = L + pq + b_l I_{b_l} - c(q), \tag{1}$$

with $L, p, b_l \geq 0$, $c'(q) > 0$ and $c''(q) > 0$. In the experiment, $c(q) = q^2/10$ for all payment systems.

When deciding on q, physician i simultaneously determines her own profit $\pi(q)$ and the patient's health benefit $H(q)$ for patient j. Common to all patients' health-benefit functions is a global optimum at q^* on $q \in (0, 10)$. The patient health-benefit function employed in our experiment is

$$H(q) = \begin{cases} H_0 + \theta q & \text{if } q \leq q^* \\ H_1 - \theta q & \text{if } q \geq q^*, \end{cases} \tag{2}$$

with $H_0, H_1 \geq 0$ and $\theta > 0$.[3] In particular, for illnesses A and B $\theta = 1$, and for illness C $\theta = 2$. For illnesses A, B, and C, the maximum health benefit is $H_A(q^*) = 7$, $H_B(q^*) = 10$, and $H_C(q^*) = 14$, respectively. Figure A.2 in Appendix A illustrates the patient health benefits in our experiment, which

[2] Notice that the general decision situation of our experiment is similar to Hennig-Schmidt et al. (2011), Hennig-Schmidt and Wiesen (2014), Brosig-Koch et al. (2016, 2017a), and Brosig-Koch et al. (2020). In the latter three studies, incentives under FFS (CAP) are the same as in the present paper.

[3] Note that $H_1 = H_0 + 2\theta q^*$. H_0 and H_1 are allowed to be different, which reflects the patient health benefit parameters in the experiment. For example, for illness A (with $\theta = 1$) and severity x (with $q^* = 3$), $H_0 = 4$ and $H_1 = 10$, as $H_1 = 4 + 2 \cdot 1 \cdot 3 = 10$.

is varied systematically for the patients' illness k, which determines the patients' marginal health benefit and and the severity of the patient's illness l.[4]

The patient-optimal quantity q^* depends on a patient's severity of illness l. For mild (x), intermediate (y), and high (z) severe illnesses, the patient-optimal quantities are $q_x^* = 3$, $q_y^* = 5$, and $q_z^* = 7$, respectively. Varying patients' characteristics in our lab experiment are motivated by the recent theoretical literature (see, e.g., Allard et al., 2011), which assumes that patient characteristics affect the physicians' behavior. Figure A.2 illustrates the patient health benefits in our experiment, which is varied systematically for the patients' illness k and severity of illness l. The differences in optimal quantities and marginal health benefits by patients' characteristics are motivated by recent claims for more value-based health care which focuses on patients' needs. In the experiment, the patient-optimal quantity q^* for all patients is common knowledge, so are all parameters of the experiment. Thus, when making their quantity choices, physicians are aware of cost, payment, profit, and the patient's health benefit for each quantity; for an illustration of the decision situation, see the instructions in Subsection A.3 in Appendix A.[5]

Our experimental design enables us to investigate how different payment schemes and performance-based payment components, which are linked to the generated health benefit (health outcome), affect treatment decisions. We are able, first, to analyze the quantity of medical services and, second, to introduce a 'clean' quality measure related to the patient-optimal treatment (see Section 2.2). Moreover, the symmetric design of patient health benefits implies that the marginal effects (i.e., the absolute value of $H'(q)$) of over- and underprovision of medical services are equivalent. This parsimonious design with mirror-image incentives allows for a systematic comparison of incentives from P4P on the quantity and quality of care and a systematic cost-benefits analysis.

2.2 Payment systems

Recall that each subject decides in the role of a physician on the provision of medical services under non-blended and blended payment systems. Table 1 provides an overview of payment systems employed in our experiment. In part I of the experiment, subjects decide either under FFS or CAP. Subjects paid by FFS (CAP) in part I decide under the associated P4P system (FFS+P4P or CAP+P4P) in part II. The profit functions of FFS and FFS+P4P systems mirror those of the respective CAP and CAP+P4P systems. While varying the components of the payment systems, we keep maximum profit levels and marginal profits constant. The profit parameters are illustrated in Figure A.2, and the complete set of parameter values is shown in Table A.2 in Appendix A.

In FFS, physicians are paid a fee of $p = 2$ per service. Accordingly, profit is $\pi(q) = 2q - c(q)$. In CAP, physicians receive a lump-sum payment of $L = 10$ per patient, independently of the quantity of medical

[4] Patients' health benefits are measured in monetary terms. The accumulated benefits are then transferred to a charity that supports surgical treatment of real cataract patients. Note that this "mechanism" implies that a monetary amount deriving from subjects' decisions in the lab is applied to the treatment of real patients, which makes it different from the kinds of donations analyzed in the charitable-giving literature; see, for example, Andreoni (1989) or DellaVigna et al. (2012). This procedure, which was introduced by Hennig-Schmidt et al. (2011), has been used in several experiments in health economics, as it embeds an incentive for subjects in the experiment that relates to real patients' health in the real world. Equivalent mechanisms have been employed in recent behavioral experiments in the field of health care which have analyzed physician behavior (Hennig-Schmidt and Wiesen, 2014; Godager et al., 2016; Brosig-Koch et al., 2016, 2017a, 2020; Byambadalai et al., 2019; Di Guida et al., 2019; Martinsson and Persson, 2019; Huesmann et al., 2020; Waibel and Wiesen, 2020; Wang et al., 2020). In Kesternich et al. (2015) and Lagarde and Blaauw (2017), subjects could choose from several (medical) charities to which a donation could be transferred.

[5] This allows a clean analysis of the extent to which patient-regarding concerns guide physicians' medical service provision, while excluding potential additional influences like risk preferences.

Table 1: Experimental parameters

First part of the experiment (Non-blended payment systems)				Second part of the experiment (Blended payment systems)						Subjects (physicians, medical students, non-medical students)
Payment	L	p	R	Payment	Severity l	L	p	b_l	R	
FFS	–	2	$2q$	FFS+P4P	x	–	2	5.6	$2q + 5.6$	52 (10, 22, 20)
					y	–	2	3.6	$2q + 3.6$	
					z	–	2	2.4	$2q + 2.4$	
CAP	10	–	10	CAP+P4P	x	10	–	2.4	$10 + 2.4$	55 (10, 22, 23)
					y	10	–	3.6	$10 + 3.6$	
					z	10	–	5.6	$10 + 5.6$	

Notes. This table shows the parameters and the number of participants in each experimental part. Note that the performance pay b_l is only granted to subjects if their quantity choice fulfills the quality requirement $|q - q^*| \leq 1$; otherwise the performance pay equals zero. Data for the non-blended payment systems correspond to a part of the data analyzed in Brosig-Koch et al. (2016).

services. Physicians' profit per patient is thus $\pi(q) = 10 - c(q)$ with the maximum attainable profit being 10 in both payment systems FFS and CAP. The profit-maximizing quantity of medical services for each of the nine patients is $\hat{q}_j^{\text{FFS}} = 10$ and $\hat{q}_j^{\text{CAP}} = 0$ in FFS and CAP, respectively. This reflects the prevalent financial incentives for overprovision under FFS and underprovision under CAP.

Our performance measure is linked to a patient's health outcome – namely, the optimal patient health benefit. P4P is granted if the quantity chosen by a physician does not deviate by more than one unit from the patient-optimal quantity q^*; i.e., whenever $|q - q^*| \leq 1$. We thereby assume that the quality is not fully contractible due to information asymmetry. P4P systems, thus, mitigate inherent incentives to provide too many services under FFS and too few under CAP, respectively. In our experiment, we determine the profit-maximizing quantities under P4P such that they are 'closer' to the patient-optimal quantities than in non-blended FFS or CAP, but do not coincide with them. Since the design of performance-based bonus payments incentivizes the smallest possible deviation from q^* instead of q^* itself, we are also able to differentiate between profit maximization and optimal patient care in our P4P conditions.

We set bonus rates such that incentives are comparable across payment systems. For severities x, y, and z, $b_x^{\text{FFS}} = 5.6$, $b_y^{\text{FFS}} = 3.6$, $b_z^{\text{FFS}} = 2.4$ in FFS+P4P, and $b_x^{\text{CAP}} = 2.4$, $b_y^{\text{CAP}} = 3.6$, $b_z^{\text{CAP}} = 5.6$ in CAP+P4P, respectively. The bonus implies an increase in the maximum attainable profit $\pi(\hat{q}_j)$ by 20 percent. For each severity, choosing $\hat{q}_j$ equal to 4, 6, or 8 (2, 4, or 6) in FFS+P4P (CAP+P4P) thus yields a profit of 12 for the physician.

2.3 Experimental protocol

Overall, 107 subjects participated in our experiment. Among these were 44 medical and 43 non-medical students who took part in the lab experiments and 20 physicians who took part in artefactual field experiments. Each subject was randomly assigned to only one of the two baseline payment systems. In particular, 55 subjects took part in CAP/CAP+P4P and 52 in FFS/FFS+P4P; with 22 medical students and 10 physicians under each payment system; see Table 1.

The computerized experiment was programmed with z-Tree (Fischbacher, 2007). Physicians and students were presented with identical computer screens, instructions, and comprehension questions. The only differences were a higher exchange factor from the experimental currency to Euro for the physicians' payoffs compared to the students' payoffs and minor deviations in the experimental procedure.[6]

[6] Before the experiments, physicians were briefly introduced to the experimental economics method, the universities involved in running the experiment, and the funding institution of our research project (DFG, German Research Foundation).

The artefactual field experiments were conducted in 2012 and 2013 using the mobile lab of the Essen Laboratory for Experimental Economics (elfe) at the Academy for Training and Education of Physicians (Akademie für Ärztliche Fort- und Weiterbildung) in Bad Nauheim, Germany. At the Academy German physicians contracting with the statutory health insurers take mandatory annual education and training courses. The physicians were recruited by announcements in their courses. They voluntarily participated before or after their courses. The lab experiments were conducted between 2011 and 2013 at elfe at the University of Duisburg-Essen.[7] Student subjects were recruited via the online recruiting system ORSEE (Greiner, 2015).

The experimental procedure was as follows: Upon arrival, subjects were randomly assigned to workstations separated by panels to ensure that decisions could be made in full anonymity. They then were given ample time to read the instructions for part I. Subjects were informed that the experiment consisted of two parts, but received detailed instructions for part II only after having finished part I of the experiment. To check for the subjects' understanding of the decision task, they had to answer a set of control questions. The experiment did not start unless all subjects had answered the control questions correctly. Instructions are to be found in Appendices A.3. In each of the two parts of the experiment, subjects then subsequently decided on the quantity of medical services for each of the nine patients, i.e., for each possible combination of illnesses and severities. The order of patients was randomly determined and kept constant for all subjects and all conditions: $Bx, Cx, Az, By, Bz, Ay, Cz, Ax, Cy$.

Before making their decision for a specific patient, subjects were informed about their payment, their cost and profit, as well as about the patient benefit for each quantity from 0 to 10. All monetary amounts are given in Taler, our experimental currency. The exchange rate is 1 Taler = EUR 0.80 in the lab experiment and 1 Taler = EUR 3.40 in the artefactual field experiment. Compared to the lab, the payment in the field experiment was increased by a factor of 4.25 to provide adequate incentives for the physicians.[8] The procedure was exactly the same in part II of the experiment. After finishing part II, we asked the subjects to complete a questionnaire on social demographics (e.g., age and gender) and on personality traits elicited by a ten-item personality inventory, which comprises five personality dimensions: extraversion, agreeableness, conscientiousness, neuroticism, and openness (Rammstedt and John, 2007). An overview on summary statistics on social demographics and personality traits can be found in Table A.1 in Appendix A.

At the end of the experiment, when all subjects had made their decisions, we randomly determined one decision in each part of the experiment to be relevant for a subject's actual payoff and the patient benefit. This was done to rule out income effects. Subjects were paid in private according to these two randomly determined decisions.

To verify that the money corresponding to the sum of patient benefits in a session was actually transferred to the charity, we applied a procedure similar to Hennig-Schmidt et al. (2011) and Brosig-Koch et al. (2016, 2017a). One of the participants was randomly chosen to be the monitor. After the experiment, the monitor verified that a payment order on the aggregated benefit in the respective session was

After the experiment, physicians were debriefed and informed about results of previous health-related economic experiments.

[7] For a picture of the setup of the mobile lab in Bad Nauheim and the typical setup of the computer laboratory at elfe, see Figure A.1.

[8] The amount physicians could earn in the experiment was set such that it reflects the average net hourly wage of a physician in Germany, bearing in mind potential differences, for example across the physicians' specialization and seniority. We set this factor after consultation with Dr. Harald Herholz of the Association of Statutory Health Insurance Physicians in Hesse (Germany), who has been involved in budget negotiations for physicians' remuneration.

written to the financial department of the University of Duisburg-Essen to transfer the money to the Christoffel Blindenmission, which used the monetary transfers exclusively to support surgical treatments of cataract patients in a hospital in Masvingo (Zimbabwe) staffed by ophthalmologists from the charity.[9] The order was sealed in an envelope and the monitor and experimenter then walked together to the nearest mailbox and deposited the envelope. The monitor was paid an additional 5 EUR.

Laboratory sessions lasted for about 60 minutes. Subjects earned, on average, EUR 16.37. The average benefit per patient was EUR 13.25. In total, EUR 1,152.80 were transferred to the Christoffel Blindenmission. The average cost for a cataract operation amounts, according to the Christoffel Blindenmission, to about EUR 30. Thus, our experiment allowed 38 cataract patients to be treated. The sessions of the artefactual field experiment lasted for about 50 minutes. Physicians earned, on average, EUR 62.73. The average benefit per patient was EUR 67.83. In total, EUR 1,356.60 were transferred to the Christoffel Blindenmission, allowing the treatment of 45 cataract patients.

2.4 Behavioral hypotheses

To organize our thoughts, we now describe the physicians' behavior more formally and derive behavioral predictions for our experiment. To this end, we follow the intuition of an illustrative model by Brosig-Koch et al. (2017a). The formal model is relegated to Appendix B. We assume that a physician derives utility from her own profit and from a patient's health benefit. The weight the physician attaches to the patient's health benefit is interpreted as a measure for physician altruism. The assumption of physicians being altruistic has become common in the health economics literature, since Arrow (1963) coined the importance of physicians' patient-regarding motivation in the delivery of medical services.[10]

First, we consider a physician's behavior under the baseline payment systems FFS and CAP. For the profit and patient benefit parameters in our experiment and the given altruism of a physician, we conjecture that FFS induces overprovision of medical services, which decreases in the severity of a patient's illness and in the patient's marginal health benefit.

On the contrary, we expect that CAP induces underprovision of medical services, which increases in the severity of illness, and decreases in the marginal health benefit. Ample evidence for these conjectures on effects of FFS and CAP exists from related experiments (e.g., Hennig-Schmidt et al., 2011; Brosig-Koch et al., 2016, 2017a; Martinsson and Persson, 2019; Brosig-Koch et al., 2020). With a higher severity of illness, more medical services are provided (Hennig-Schmidt et al., 2011; Brosig-Koch et al., 2016, 2017a; Brosig-Koch et al., 2020). These behavioral effects related to the severity of illness are particularly relevant in our experiment, as the levels of P4P are tied to the patients' severity of illness; for an illustration, see Figure A.3 in Appendix A.2.[11]

Our main focus is on the effect of P4P. When introducing P4P the bonus b_l is granted if and only if a physician's treatment decision meets the quality threshold $|q - q^*| \leq 1$ for a patient with a severity

[9] Notice that we did not inform the subjects that the money was assigned to a developing country. We wanted to avoid motives like compassion for people in developing countries that are independent of being in need of ophthalmic surgery. Feedback from the subjects in a pre-experimental pilot session in Hennig-Schmidt et al. (2011) actually raised this issue.

[10] In addition to the importance for designing optimal payment schemes (e.g., Ellis and McGuire, 1986, 1990; Ma, 1994; Chalkley and Malcomson, 1998; Jack, 2005; Choné and Ma, 2011; Olivella and Siciliani, 2017), a physician's altruism is essential, for example in analyzing referral decisions (Allard et al., 2011; Waibel and Wiesen, 2020), responses to transparency (Kolstad, 2013), prescription of generic drugs (e.g., Hellerstein, 1998; Crea et al., 2019), and the delegation of treatment decisions (Liu and Ma, 2013).

[11] Performance pay for mild-severity patients, b_x^{FFS}, is highest with $b_x^{\mathrm{FFS}} > b_y^{\mathrm{FFS}} > b_z^{\mathrm{FFS}}$, as the 'risk' of a mild-severity patient being overserved and therefore suffering disutility is highest under FFS. In CAP, the incentive to undertreat patients increases in the severity of illness; therefore, $b_z^{\mathrm{CAP}} > b_y^{\mathrm{CAP}} > b_x^{\mathrm{CAP}}$ (see Table 1).

of illness l. Recall that we thus assume that the quality is not fully contractible due to information asymmetry between physician and payer. By linking performance pay to the optimal health benefit, the interests of the physician and the patient become (more) aligned. While P4P incentivizes less altruistic physicians to provide medical services 'close' to the patient-optimal quantity, baseline financial incentives for underprovision under CAP and overprovision under FFS are still inherent (albeit to a substantially lower extent). Hence, we conjecture that P4P reduces overprovision of medical services in FFS and underprovision in CAP. For the proof, see Appendix B.

Intuitively, whether a physician under baseline FFS and CAP chooses a quantity of medical services corresponding to the quality threshold depends on physician i's degree of altruism towards the patient, which counterbalances the incentive effects of FFS and CAP. Given a physician's altruism $\alpha_i \in [0, 1)$, we therefore expect P4P to reduce non-optimal service provision under FFS+P4P and CAP+P4P. Previous experimental evidence (Hennig-Schmidt et al., 2011; Brosig-Koch et al., 2016, 2017a; Brosig-Koch et al., 2020) has shown that, in the basic payment schemes FFS and CAP, non-optimal service provision is highest for patients where $\hat{q}$ and q^* are most misaligned. This is the case for mild-severity patients under FFS and high-severity patients under CAP. We therefore expect the effect sizes of P4P to vary with patients' severities of illness. In sum, we state the following hypotheses about physicians' medical service provision and the quality of care.

Hypothesis 1. P4P blended with FFS.
A threshold-based performance pay system with a discrete bonus reduces the overprovision of medical services under fee-for-service and increases the quality of care.

Hypothesis 2. FFS+P4P and patient's health characteristics.
Under performance pay and fee-for-service, the effect of performance pay on medical service provision and the quality of care decreases in the patient's severity of illness and the patient's marginal health benefit.

Hypothesis 3. P4P blended with CAP.
Introducing performance pay reduces the underprovision of medical services under capitation and enhances the quality of care.

Hypothesis 4. CAP+P4P and patient's health characteristics.
Under performance pay and capitation, the effect of performance pay increases in the patient's severity of illness and the patient's marginal health benefit.

Following directly from Hypotheses 2 and 4, we state:

Hypothesis 5. Comparison of FFS+P4P and CAP+P4P.
FFS+P4P leads to a larger improvement in the quality of care for mildly ill patients compared to CAP+P4P. For severely ill patients, the increase in quality of care is larger for CAP+P4P, while for intermediately ill patients, the quality of care does not differ between the two pay-for-performance systems.

3 Behavioral results

In this section, we first provide an introductory, mostly descriptive analysis of medical service provision and the quality of care under the baseline payment systems, FFS and CAP, and the blended performance pay systems, FFS+P4P and CAP+P4P (Subsection 3.1). Second, we test our Hypotheses 1 and 2 on the effects of performance pay when blended with FFS (Subsection 3.2) and, third, analogously, for a blended CAP and performance pay, we test Hypotheses 3 and 4 (Subsection 3.3). Finally, we compare the effects of blended payment systems FFS+P4P and CAP+P4P according to Hypotheses 5 (Subsection 3.4).

In our analyses, we consider the quantity of medical services q and capture the quality of care considering two quality measures. First, our choice-based measure is the absolute deviation from the patient-optimal quantity $\rho = |q - q^*|$. Second, our outcome-based measure is the proportional health benefit $\hat{H}$. It comprises the patient's health benefit realized by the physician's actual quantity choice (H_{kl}) as a proportion of the highest achievable health benefit (H_l^*). To facilitate the comparisons across patients (kl) who also vary in terms of their minimal health benefit $H^{\min}$, we normalize our measure accounting for the lower bounds of achievable health benefit, and define $\hat{H}_{kl} = \frac{H_{kl}^{\min} - H_{kl}}{H_{kl}^{\min} - H_l^*}$. When physician i provides the patient-optimal quantity q^*, the proportional health benefit is highest and $\hat{H}_{kl} = 1$. $\hat{H}_{kl}$ implies a measurable health outcome which allows us to compare actual with optimal quality of care across the different payment systems.

3.1 Introductory analyses

Figure 1 illustrates the average quantity under the four payment systems for each of the nine patients who differ by illness k and severity of illnesses l.[12] We find that subjects provide significantly more services under FFS (Mean 6.69, s.d. 2.07) than under CAP (Mean 3.32, s.d. 2.13), aggregated over all patients ($p < 0.001$, two-sided Mann-Whitney U-test, MWU in the following), see Panel A of Table C.1 in Appendix C. This finding is in line with earlier experimental studies (recall Section 2.4). In FFS+P4P, the quantities of medical services decrease by about 16.4 percentage points (Mean 5.59, s.d. 1.71), and in CAP+P4P, they increase by about 32.5 percentage points (Mean 4.40, s.d. 1.66).

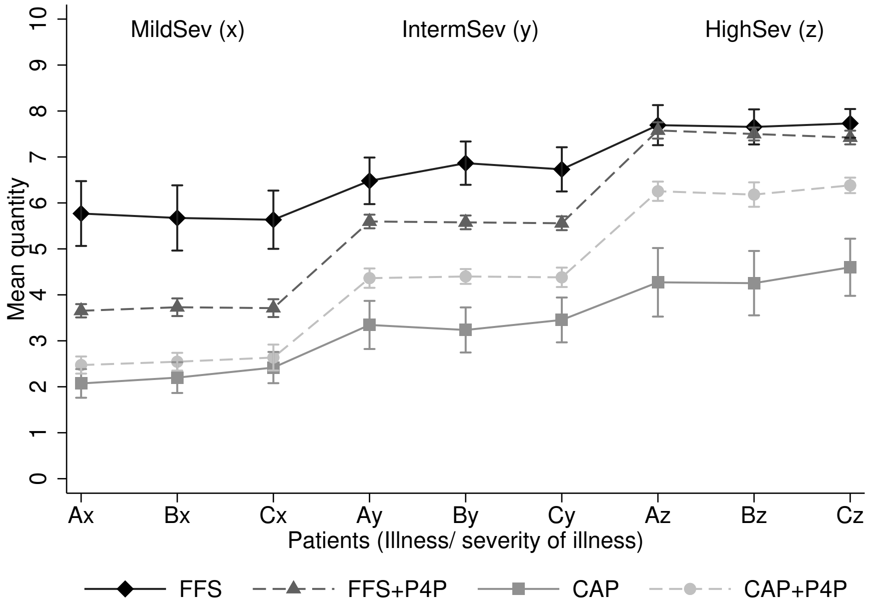
Under FFS, the absolute deviation from the patient-optimal quantity $\rho = 1.82$ (s.d. 1.95), aggregated over all patients. Introducing P4P reduces the average non-optimal service provision ρ to 0.63 (s.d. 0.55), which is a reduction by 65.4 percentage points. Under CAP, $\rho = 1.77$ (s.d. 2.01), while under CAP+P4P ρ declines to 0.65 (s.d. 0.75), a decrease by 63.3 percentage points. See Panel B of Table C.1 for detailed descriptive statistics on our choice-based measure ρ.

For the proportional health benefit $\hat{H}$, we find that, on average, around 71% of the maximum health benefit is realized in the two basic payment systems and around 90% in the two blended payment systems. The effect of P4P thus corresponds to an overall increase in the proportional health benefit by 19 percentage points under CAP+P4P and FFS+P4P. Detailed descriptive statistics on $\hat{H}$ are provided in Panel C of Table C.1 in Appendix C. On the aggregate, the introduction of P4P leads to a significant increase in the quantity, and in both the choice-based and the outcome-based quality measures ($p < 0.001$, Wilcoxon signed-rank test, two-sided).

We further observe that the patients' severity of illness substantially affects the subjects' behavior in all payment systems; see Table C.1 in Appendix C. Overprovision of medical services is highest for

[12] See the distributions of medical services differentiated by payment schemes and patients in Figure C.1, Appendix C.

Figure 1: Mean quantity by patients' health characteristics



Notes. This figure shows the mean quantity with 95% confidence interval under the four payment systems for each of the nine patients kl. Patients vary by their illness $k = A, B, C$ and severity of illnesses l with mild (x), intermediate (y), and high (z) severe illnesses.

mildly ill patients in both FFS conditions, and underprovison is highest for severely ill patients in both CAP conditions. The behavioral effect is rather less pronounced for the marginal health benefit. For a detailed overview on descriptive statistics and non-parametric tests for all payment systems and the patients' characteristics, see Table C.2 in Appendix C.

We now briefly analyze whether responses to the baseline payment systems FFS and CAP are different. To this end, we run a regression on the quantity and quality of care at a between-subject level, see Table C.3 in Appendix C. In line with the inherent incentives, we find that the treatment quantity is, on average, 3.44 medical services lower under CAP than under FFS. Due to the opposing inherent incentives, effects on the quality of care are more meaningful when comparing both payment systems. On average, neither non-optimal care nor proportional health benefit differs significantly between both baseline payment systems; see Panel B and C of Table C.3 in Appendix C. Our regression results indicate that the incentives to underprovide in CAP are equally strong as the incentives to overprovide in FFS. For a patient's severity of illness, we find no significant differences in non-optimal care, except in the proportional patient's health benefit, which is on average about 10.0 percentage points lower for an intermediately ill than for a mildly ill patient. Patients with a high marginal health benefit receive on average a significantly higher quality of care than patients with a low marginal health benefit.

3.2 The effect of blending fee-for-service with performance pay

We next analyze how introducing P4P+FFS affects the quantity and quality of medical service provision on the individual level. These analyses are particularly important, as a detailed experimental investigation of these effects was lacking up to now.

To estimate the P4P effect, we use OLS regressions for the independent variables q_{ij} (quantity chosen), and $\rho_{ij} = |q_{ij} - q_j^*|$ and a fractional probit response model for the proportional health benefit $\hat{H}_{ij}$, scaled between 0 and 1. Our base econometric specification is as follows:

$$y_{ij} \quad = \quad \alpha + \beta_1 \text{P4P} + \beta_2 \textsc{IntermSev} + \beta_3 \textsc{HighSev} + \beta_4 \textsc{High}\text{MHB} + \beta_5 \mathbf{X}_i + \epsilon_{ij}. \tag{3}$$

InTERMSEV and HIGHSEV are dummy variables for intermediate and high severities of illness, respectively. HIGHMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). P4P is a dummy variable indicating the introduction of P4P. X_i is a vector of subject i's characteristics comprising gender, personality traits and medical background (non-medical students, medical students or physicians). We account for potentially confounding effects by medical background as previous experimental evidence (e.g., Brosig-Koch et al., 2016) indicates that behavioral responses to financial incentives might differ (merely in size not qualitatively) by subject pool. Our estimated effects of P4P remain stable when we control for subjects' medical background and other characteristics.[13]

According to Hypothesis 1, we expect that P4P reduces the quantity of medical services, induces less overprovision, and induces a higher proportional health benefit. Models (1), (4), and (7) in Table 2 show that Hypothesis 1 is supported. Introducing FFS-based P4P leads to a highly significant reduction in treatment quantity by, on average, 1.10 medical services. Non-optimal care also declines highly significantly, by 1.20 medical services on average. The proportional patient's health benefit rises by about 18.9 percentage points when P4P is introduced. We summarize the regression results as follows:

Result 1 (P4P blended with FFS). *Complementing fee-for-service with a threshold-based performance-pay system leads to a decrease in overprovision of medical services, which corresponds to an increase in the quality of medical care and in the proportional health benefit.*

Hypothesis 2 considers the effect the severities of illness and the marginal health benefit have on physicians' responses to P4P. Before testing the hypothesis, we analyze the respective impacts on the physicians' treatment decisions under FFS payment conditions in general. Compared to mildly ill patients, treatment quantities increase significantly for intermediately and severely ill patients by, on average, 1.44 and 2.90 medical services, respectively; see Model (1) of Table 2. Considering treatment quality, non-optimal care significantly decreases with increasing severity (Model (4)). The proportional health-benefit increase for severely ill patients is significantly higher than for mildly ill patients (by 13.3 percentage points) but does not significantly differ between intermediately and mildly ill patients (Model (7)). These findings are in line with results reported by, e.g., Brosig-Koch et al. (2017a) and Martinsson and Persson (2019).

Hypothesis 2 states that the effect of P4P on medical service provision and on the quality of care decreases in the severity of illness and in the marginal health benefit. To estimate the moderating effects that patients' severities of illness have on responses to P4P, we consider the following model:

$$\begin{aligned} y_{ij} \quad &= \quad \alpha + \beta_1 \textsc{IntermSev} + \beta_2 \textsc{HighSev} + \beta_3 \textsc{High}\text{MHB} + \beta_4 \text{P4P}\mathbf{x}\textsc{MildSev} \\ &+ \quad \beta_5 \text{P4P}\mathbf{x}\textsc{IntermSev} + \beta_6 \text{P4P}\mathbf{x}\textsc{HighSev} + \beta_7 \mathbf{X}_i + \epsilon_{ij}. \end{aligned} \tag{4}$$

Following an econometric approach by Clark and Huckman (2012), we include the terms β_4P4P**x**MILDSEV

[13] For a comparisons of regression estimates without individual controls, see Tables C.4 and C.5 in Appendix C. For regression estimates with the full list of individual controls, see Tables C.6 and C.7 in Appendix C.

Table 2: Regression models on the effect on quantity and quality under FFS conditions

	A. Quantity of medical services q			**B. Absolute deviation from optimal care ρ**			**C. Proportional health benefit $\hat{H}$**		
Method:	OLS	OLS	OLS	OLS	OLS	OLS	Frac. Probit	Frac. Probit	Frac. Probit
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
P4P	-1.100***			-1.199***			0.189***		
	(0.185)			(0.172)			(0.025)		
IntermSev	1.439***	1.000***	1.439***	-0.529***	-0.936***	-0.529***	0.004	0.019	0.004
	(0.086)	(0.147)	(0.086)	(0.086)	(0.149)	(0.086)	(0.010)	(0.012)	(0.010)
HighSev	2.901***	2.000***	2.901***	-0.997***	-1.782***	0.997***	0.133***	0.184***	0.133***
	(0.129)	(0.230)	(0.129)	(0.141)	(0.256)	(0.141)	(0.016)	(0.023)	(0.016)
HighMHB	-0.016	-0.016	0.010	-0.054	-0.054	-0.074	0.009	0.009	0.008
	(0.053)	(0.053)	(0.089)	(0.051)	(0.051)	(0.087)	(0.008)	(0.008)	(0.011)
P4P×MildSev		-1.994***			-1.994***			0.187***	
		(0.277)			(0.277)			(0.021)	
P4P×IntermSev		-1.115***			-1.179***			0.162***	
		(0.192)			(0.183)			(0.021)	
P4P×HighSev		-0.192			-0.423***			0.079***	
		(0.132)			(0.111)			(0.017)	
P4P×LowMHB			-1.083***			-1.212***			0.171***
			(0.195)			(0.179)			(0.022)
P4P×HighMHB			-1.135***			-1.173***			0.153***
			(0.177)			(0.172)			(0.018)
Constant	5.623***	6.070***	5.615***	2.621***	3.019***	2.627***			
	(0.315)	(0.350)	(0.318)	(0.315)	(0.354)	(0.317)			
Wald test (p-value)									
H_0: P4P×MildSev =P4P×IntermSev		<0.001			<0.001			0.010	
H_0: P4P×MildSev= P4P×HighSev		<0.001			<0.001			<0.001	
H_0: P4P×IntermSev=P4P×HighSev		<0.001			<0.001			<0.001	
H_0: P4P×LowMHB=P4P×HighMHB			0.556			0.658			0.062
Observations	936	936	936	936	936	936	936	936	936
Subjects	52	52	52	52	52	52	52	52	52
(Pseudo) R^2	0.563	0.599	0.563	0.336	0.379	0.336	0.150	0.157	0.150

Notes. This table shows parameter estimates from OLS regressions (Panel A and B) and average marginal effects from fractional probit response regressions (Panel C). Robust standard errors clustered for subjects are shown in parentheses. P4P is a dummy variable indicating the introduction of P4P. IntermSev and HighSev are dummy variables for intermediate and high severities of illness. HighMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). All models control for individual characteristics which comprise gender, medical background (non-medical student, medical student, physician), and personality traits; for the respective estimates, see Table C.6 in Appendix C. * $p < 0.10$, ** $p < 0.05$, and, *** $p < 0.01$.

β_5P4PxINTERMSEV, and β_6P4PxHIGHSEV, which interact P4P with each severity level of illness to determine the extent to which the effect (marginal benefit) of P4P depends on the patient's severity of illness. By construction, the estimates of β_4, β_5, and β_6 represent the total effect of P4P for patients with either a mild severity of illness, an intermediate severity of illness or a high severity of illness, respectively.

Regression results provide support for Hypothesis 2 (see Models (2), (5), and (8), as well as Wald test results). First, P4P positively affects the quantity and quality of care, as all coefficients on the effects are significantly different from zero, except the effect of P4P on the quantity of medical services for severely ill patients (see Model (2)). Second, we find the anticipated relation between severity of illness and P4P such that coefficients are significantly higher for less severely ill patients.

For a patient's level of marginal health benefit, we neither find a significant effect on the quantity of medical service provision nor on the quality of care, see Models (1), (4), and (7). To estimate whether the positive effect of P4P differs between patients with high and low marginal benefits, we consider a model similar to Equation (4), in which we interact P4P with each marginal health-benefit level. When comparing the effect of P4P for patients with a low marginal health benefit (P4PxLOWMHB) to the effect for patients with a high marginal health benefit (P4PxHIGHMHB), we observe no significant differences; see Models (3), (6), and (9) of Table 2 and the Wald tests. We summarize our findings as follows:

Result 2 (**FFS+P4P and patients' health characteristics**). *While introducing performance pay improves the quality of medical service provision across all severity types, the effect of performance pay significantly decreases with increasing severity of illness. For patients' marginal health benefit, the effect of performance pay is less systematic.*

3.3 The effect of blending capitation with performance pay

We now analyze the effects of introducing P4P to CAP on the quantity and quality of care. An earlier study by Brosig-Koch et al. (2020) used the same design to investigate the P4P effect with general practitioners and medical students when CAP is the baseline payment. We repeat the analyses with our data according to our econometric specifications in Equations (3) and (4). We thus provide the basis for jointly comparing the payment systems FFS, CAP, FFS+P4P, and CAP+P4P in Section 3.4.

According to Hypothesis 3, we expect that introducing P4P to CAP increases the treatment quantity, reduces the underprovision of medical services, and enhances the quality of care. Our data support Hypothesis 3. Models (1), (4), and (7) of Table 3 show that CAP+P4P leads to a highly significant increase in the treatment quantity by on average 1.09 services, a reduction of non-optimal care by on average 1.12 medical services, and an increase in the proportional health benefit by about 17.5 percentage points.[14] In sum, we state:

Result 3 (**P4P blended with CAP**). *Complementing capitation with performance pay leads to an increase in medical services, a decrease in underprovision, and a rise in patients' proportional health benefit.*

To analyze the effect severities of illness and the marginal health benefit have on the physicians' responses to CAP+P4P (Hypothesis 4), we again first study the impact the severities of illness have on the physicians' treatment decisions, as indicated in Equation (3). We find that treatment quantities for

[14] Regression estimates for models without individual controls yield very similar results; see Table C.5 in Appendix C.

Table 3: Regression models on the effect on quantity and quality under CAP conditions

	A. Quantity of medical services q			**B. Absolute deviation from optimal care ρ**			**C. Proportional health benefit $\hat{H}$**		
Method:	OLS	OLS	OLS	OLS	OLS	OLS	Frac. Probit	Frac. Probit	Frac. Probit
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
P4P	1.085***			-1.117***			0.175***		
	(0.189)			(0.180)			(0.026)		
IntermSev	1.473***	1.115***	1.473***	0.436***	0.848***	0.436***	-0.143***	-0.201***	-0.143***
	(0.100)	(0.139)	(0.100)	(0.074)	(0.138)	(0.074)	(0.016)	(0.024)	(0.016)
HighSev	2.933***	2.145***	2.933***	0.958***	1.758***	0.958***	-0.149***	-0.227***	-0.149***
	(0.151)	(0.245)	(0.151)	(0.134)	(0.250)	(0.134)	(0.019)	(0.028)	(0.019)
HighMHB	0.179***	0.179***	0.261***	-0.115**	0.115**	-0.194**	0.017**	0.017**	0.024***
	(0.048)	(0.048)	(0.069)	(0.044)	(0.044)	(0.073)	(0.007)	(0.007)	(0.009)
P4P×MildSev		0.321**			-0.309***			0.055***	
		(0.140)			(0.108)			(0.017)	
P4P×IntermSev		1.036***			-1.133***			0.157***	
		(0.201)			(0.189)			(0.021)	
P4P×HighSev		1.897***			-1.909***			0.180***	
		(0.296)			(0.292)			(0.023)	
P4P×LowMHB			1.139***			-1.170***			0.165***
			(0.195)			(0.188)			(0.024)
P4P×HighMHB			0.976***			-1.012***			0.135***
			(0.193)			(0.173)			(0.018)
Constant	1.725***	2.107***	1.698***	1.440***	1.036***	1.466***			
	(0.250)	(0.231)	(0.254)	(0.242)	(0.227)	(0.247)			
Wald test (p-value)									
H_0: P4P×MildSev=P4P×IntermSev		<0.001			<0.001			<0.001	
H_0: P4P×MildSev=P4P×HighSev		<0.001			<0.001			<0.001	
H_0: P4P×IntermSev=P4P×HighSev		<0.001			<0.001			0.015	
H_0: P4P×LowMHB=P4P×HighMHB			0.097			0.049			0.004
Observations	990	990	990	990	990	990	990	990	990
Subjects	55	55	55	55	55	55	55	55	55
(Pseudo) R^2	0.509	0.534	0.509	0.287	0.328	0.287	0.131	0.140	0.131

Notes. This table shows parameter estimates from OLS regressions (Panel A and B) and average marginal effects from fractional probit response regressions (Panel C). Robust standard errors clustered for subjects are shown in parentheses. P4P is a dummy variable indicating the introduction of P4P. IntermSev and HighSev are dummy variables for intermediate and high severities of illness. HighMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). All models control for individual characteristics which comprise gender, medical background (non-medical student, medical student, physician), and personality traits; for the respective estimates, see Table C.7 in Appendix C. * $p < 0.10$, ** $p < 0.05$, and, *** $p < 0.01$.

intermediately and severely ill patients are significantly higher than for mildly ill patients by, on average, 1.47 and 2.93 medical services, respectively; see Model (1) of Table 3. Nevertheless, the quality of care is significantly lower for intermediately ill (severely ill) patients by on average 0.44 (0.96) services below the patient-optimal quantity. Correspondingly, the proportional health benefit is on average 14.3 (14.9) percentage points lower for these patients; see Models (4) and (7) of Table 3.

Hypothesis 4 states that under CAP+P4P, the P4P effect increases in the patient's severity of illness and in the marginal health benefit. Our estimations based on Equation (4) provide the following results: While the average effect of P4P on the quantity and the quality of medical services is positive and significant for all severity levels, we find substantial heterogeneity when splitting the P4P effect by severities. P4P enhances q by, on average, 0.32, 1.04, and 1.90 medical services for mildly, intermediately, and severely ill patients. This corresponds to a reduction in ρ by, on average, 0.31, 1.31, and 1.91 medical services, and to an increase in $\hat{H}$ by, on average, 5.5, 15.7, and 18.0 percentage points, respectively; Wald tests show that effect sizes are significantly different from each other, see Models (2), (5), and (8) of Table 3. Under capitation, the quantity of medical services for severely ill patients deviates the most from the patient-optimal quantity, resulting in the lowest proportional health benefit (Table C.1). As physicians respond to P4P, these patients are those who benefit the most from introducing CAP+P4P, which is in line with Hypothesis 4.

We also find that patients with a high marginal health benefit receive significantly more medical services and quality of care compared to patients with a low marginal health benefit. Moreover, while both patient types benefit from CAP+P4P, the patients with a low marginal benefit gain more from introduction of P4P than those with a high marginal benefit; see Models (6) and (9) of Table 3 and the respective Wald tests. This pattern is not in line with Hypothesis 4. However, differences in effect sizes of P4P for patients with a low and high level of marginal health benefit are rather small, and adding interaction terms of marginal health benefits and P4P does not explain better the variation in our data (comparing Models (1) to (3), (4) to (6), and (7) to (9)). In sum, we state:

Result 4 (CAP+P4P and patients' health characteristics). *The effect of performance pay significantly increases in patients' severities of illness. Patients with a low as well as a high level of marginal health benefit gain from performance pay; yet the effect on quality is smaller for patients with a higher marginal benefit.*

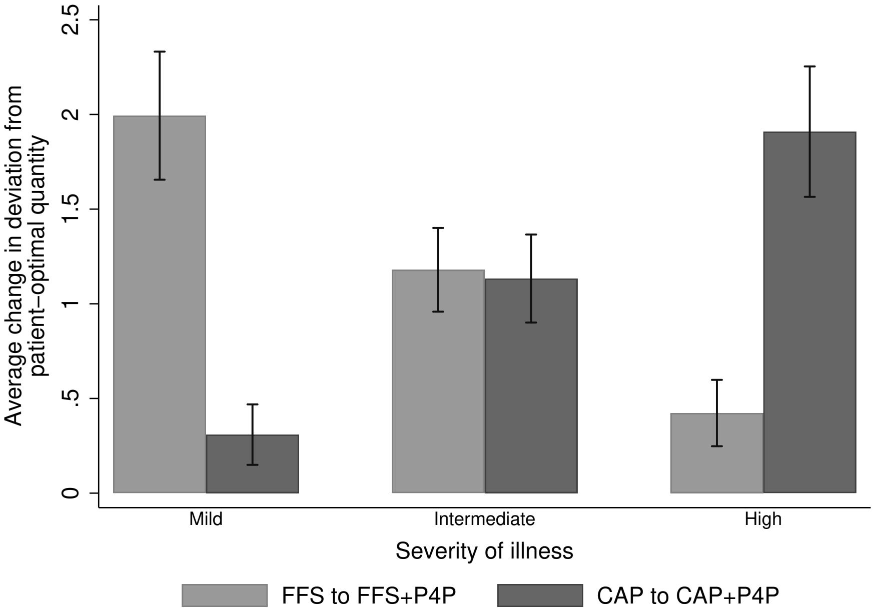
Results 3 and 4 are in line with findings by Brosig-Koch et al. (2020) for the effect of P4P on treatment quantity and the quality of care. In their study, the effects for the marginal health benefit go in the same direction, but they are statistically not significant.

3.4 Comparison of pay for performance effects in capitation and fee-for-service payment systems

In this section, we investigate how subjects' responses to performance pay differ between FFS and CAP conditions. Although the payment systems are structurally symmetric due to our mirror-image design, the effect sizes may be different for the following reasons.FFS, a piece-rate system with fees higher than marginal costs, incentivizes the provision of care to be more than patient-optimal. Under CAP, however, physicians have an incentive for underprovision as each medical service provided is costly, thus reducing the physician's profit. Depending on the baseline payment condition, introducing P4P provides incentives that go in opposite directions: to reduce services under FFS and to expand treatment under CAP.

To address Hypothesis 5, we investigate whether the severity-specific effects of P4P on the quality of care differ between FFS+P4P and CAP+P4P.[15] Figure 2 shows that the effects of blended P4P systems on ρ seem to depend on the patient's severity of illness. Table 4 provides more detailed descriptive statistics on the two quality measures ρ and $\hat{H}$ differentiated by patients' severities of illness and marginal health benefits. For mildly ill patients, the improvement in the quality of care is significantly higher under FFS+P4P than under CAP+P4P ($p < 0.001$, two-sided Mann-Whitney U-tests for both quality measures). We observe the reverse pattern for severely ill patients ($p < 0.001$) and no significant differences for patients with an intermediate severity of illness ($p \geq 0.598$). When differentiating patients by their marginal health benefit, no significant differences in the P4P effect across payment conditions are found; neither for patients with a low nor with a high level ($p \geq 0.308$). [16]

Figure 2: Reduction in the absolute deviation from optimal care by payment system and severity of illness



Notes. This figure shows the reduction in ρ achieved by introducing performance pay, differentiated by FFS and CAP conditions and severities of illness.

In order to investigate Hypothesis 5 further, we use regression analyses, extending our basic econo-

[15] Note that a comparison of the effect differences in quantity appears unreasonable, since P4P leads to opposite responses, i.e., a decrease under FFS or an increase under CAP. Thus, a joint comparison of effects on quantity of care for both payment systems does not allow us to draw any meaningful inferences.

[16] For the effect of patients' marginal health benefits, see the estimation results in Table C.8 in Appendix C.

Table 4: Comparison of effects of performance pay blended with fee-for-service and capitation on the quality of care

	FFS to FFS+P4P	CAP to CAP+P4P	Diff.	p-value
A. Change in absolute deviation from optimal care ρ				
Aggregate	-1.20 (1.73)	-1.12 (1.79)	-0.08	0.278
Mild severity	-1.99 (2.14)	-0.31 (1.04)	-1.68	<0.001
Intermediate severity	-1.18 (1.40)	-1.13 (1.51)	-0.05	0.598
High severity	-0.42 (1.11)	-1.91 (2.24)	-1.49	<0.001
Low marginal health benefit	-1.21 (1.78)	-1.17 (1.87)	-0.04	0.519
High marginal health benefit	-1.17 (1.63)	-1.01 (1.63)	-0.16	0.316
B. Change in proportional health benefit $\hat{H}$				
Aggregate	0.19 (0.27)	0.18 (0.29)	0.01	0.286
Mild severity	0.28 (0.31)	0.04 (0.15)	0.24	<0.001
Intermediate severity	0.24 (0.28)	0.23 (0.31)	0.01	0.608
High severity	0.06 (0.16)	0.27 (0.32)	0.21	<0.001
Low marginal health benefit	0.19 (0.28)	0.19 (0.30)	0.00	0.550
High marginal health benefit	0.19 (0.26)	0.16 (0.26)	0.03	0.308
Observations	468	495		
Subjects	52	55		

Notes. The table reports descriptive statistics on the changes in our quality measures ρ and $\hat{H}$ when moving from unblended to pay-for-performance payment schemes (means; standard deviations in parentheses). We differentiate by patients' severities of illness and the marginal health benefit. Column 'Diff' reports average differences in effect sizes between both payment schemes; reported p-values are based on two-sided Mann-Whitney U tests.

metric model by the between-payment system comparison as follows:

$$\begin{aligned} y_{ij} &= \alpha + \beta_1\text{CAP} + \beta_2\textsc{IntermSev} + \beta_3\textsc{HighSev} + \beta_4\textsc{High}\text{MHB} \\ &+ \beta_5\text{CAP} \times \textsc{IntermSev} + \beta_6\text{CAP} \times \textsc{HighSev} \\ &+ \beta_7\text{CAP+P4P} \times \textsc{MildSev} + \beta_8\text{FFS+P4P} \times \textsc{MildSev} \\ &+ \beta_9\text{CAP+P4P} \times \textsc{IntermSev} + \beta_{10}\text{FFS+P4P} \times \textsc{IntermSev} \\ &+ \beta_{11}\text{CAP+P4P} \times \textsc{HighSev} + \beta_{12}\text{FFS+P4P} \times \textsc{HighSev} \\ &+ \beta_{13}\mathbf{X}_i + \epsilon_{ij}, \end{aligned} \tag{5}$$

The variable CAP is a dummy which equals 1 if a physician is remunerated by CAP, and 0 if he or she is remunerated by FFS. $\textsc{IntermSev}\times$CAP and $\textsc{HighSev}\times$CAP show interaction effects between CAP and the respective level of severity. To determine how severity-specific effects of P4P vary by the underlying remuneration condition, we interact the variables CAP+P4P and FFS+P4P, which are dummies for the respective blended payment systems with each level of severity. The estimate for β_7 thus represents the total effect of P4P for mildly ill patients under CAP, while β_8 represents the total effect for mildly ill patients under FFS and, respectively, for β_9 to β_{12}.

Table 5 presents estimates for two versions of Equation 5 for each quality measure which differ in that they include X_i as the vector of subject i's characteristics. Effect differences at a between-subject level may be sensitive to individual characteristics. We find that our estimates on the severity-specific effects

Table 5: Comparison of effects of blended performance pay systems

	A. Absolute deviation from patient-optimal care ρ		**B. Proportional health benefit $\hat{H}$**	
Method:	OLS	OLS	Frac. Probit	Frac. Probit
Model:	(1)	(2)	(3)	(4)
CAP	-1.828***	-1.832***	0.210***	0.209***
	(0.342)	(0.310)	(0.037)	(0.032)
IntermSev	-0.936***	-0.936***	0.020*	0.020
	(0.147)	(0.148)	(0.012)	(0.012)
HighSev	-1.782***	-1.782***	0.182***	0.178***
	(0.253)	(0.254)	(0.021)	(0.020)
HighMHB	-0.086**	-0.086**	0.013**	0.013**
	(0.033)	(0.034)	(0.005)	(0.005)
CAP×IntermSev	1.784***	1.784***	-0.243***	-0.241***
	(0.201)	(0.201)	(0.029)	(0.028)
CAP×HighSev	3.540***	3.540***	-0.492***	-0.484***
	(0.354)	(0.355)	(0.033)	(0.033)
CAP+P4P×MildSev	-0.309***	-0.309***	0.056***	0.054***
	(0.107)	(0.107)	(0.017)	(0.016)
FFS+P4P×MildSev	-1.994***	-1.994***	0.171***	0.170***
	(0.274)	(0.275)	(0.017)	(0.016)
CAP+P4P×IntermSev	-1.133***	-1.133***	0.148***	0.148***
	(0.187)	(0.188)	(0.018)	(0.017)
FFS+P4P×IntermSev	-1.179***	-1.179***	0.151***	0.150***
	(0.181)	(0.182)	(0.017)	(0.016)
CAP+P4P×HighSev	-1.909***	-1.909***	0.168***	0.167***
	(0.289)	(0.290)	(0.018)	(0.017)
FFS+P4P×HighSev	-0.423***	-0.423***	0.075***	0.077***
	(0.110)	(0.111)	(0.015)	(0.015)
Constant	2.759***	2.950***		
	(0.316)	(0.324)		
Individual controls	No	Yes	No	Yes
Wald tests (p-value):				
H_0: CAP+P4P×MildSev = FFS+P4P×MildSev	<0.001	<0.001	<0.001	<0.001
H_0: CAP+P4P×IntermSev = FFS+P4P×IntermSev	0.860	0.860	0.872	0.884
H_0: CAP+P4P×HighSev = FFS+P4P×HighSev	<0.001	<0.001	<0.001	<0.001
Observations	1926	1926	1926	1926
Subjects	107	107	107	107
(Pseudo) R^2	0.240	0.312	0.094	0.129

Notes. For Panel A, OLS estimates are reported with robust standard errors clustered for subjects (in brackets). For Panel B, average marginal effects (AMEs), based on a fractional probit response model, are reported with robust standard errors clustered for subjects (in brackets). CAP = 1 if physicians are remunerated by CAP, and = 0 otherwise (by FFS). P4P is a dummy variable indicating the introduction of P4P. IntermSev and HighSev are dummy variables for intermediate and high severities of illness. HighMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). Controls for subjects' individual characteristics comprise gender, medical background (non-medical student, medical student, physician), and personality traits; for the respective estimates, see Table C.9 in Appendix C. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

of P4P are robust to controlling for individual characteristics (comparing Models (1) to (2) and (3) to (4) of Table 5). For simplicity, we, thus focus on Models (2) and (4) when describing our estimation results.

Our findings that the effects of blended P4P systems are severity-specific support Hypothesis 5. We find that the marginal benefit of P4P on the quality of care is highest for mildly ill patients under FFS+P4P. Models (2) and (4) of Table 5 show that the absolute deviation from the patient-optimal quantity is reduced by on average 1.99 medical services, and the patients' health benefit increases by about 17.0 percentage points. On the contrary, the effect is lowest for mildly ill patients under CAP+P4P. Estimates indicate a reduction in ρ by about 0.31 medical services and an increase in $(\hat{H})$ by 5.4 percentage points. The introduction of P4P is therefore 6.5 times (3.1 times) more effective in terms of ρ $(\hat{H})$ for mildly ill patients under FFS+P4P than under CAP+P4P.

For severely ill patients, the estimates show a reverse pattern in that the P4P effect is significantly higher under CAP+P4P compared to FFS+P4P. P4P leads to a reduction in ρ by about average 1.91 medical services under CAP+P4P and about 0.42 medical services under FFS+P4P. $\hat{H}$ increased by 16.7 (7.7) percentage points under CAP+P4P (FFS+P4P).

For intermediately ill patients, we find no significant difference in P4P effects between payment systems. Put differently, the introduction of P4P yields similar quality improvements for intermediately ill patients, which lead to a reduction of about 1.13 (1.18) in ρ and a higher $\hat{H}$ by about 14.8 (15.0) percentage points under CAP+P4P (FFS+P4P). In sum, we state the following result:

Result 5 (Comparisons of FFS+P4P and CAP+P4P). *The effect of performance pay on the quality of care is specific to the patient's severity of illness for the two blended pay-for-performance systems. While the effect on quality of care is significantly higher for mildly ill patients under FFS+P4P, it is significantly higher for severely ill patients under CAP+P4P. For intermediately ill patients the effect of performance pay on the quality of care does not differ between payment systems.*

4 Benefits and costs of introducing performance pay

To put the behavioral results in context, we now discuss the benefits and costs of introducing performance pay. Most research on the effects of initiating a P4P system focuses on quality measure targets, thereby often neglecting the pertinent issues of individual health outcomes and costs (Meacock et al., 2014). In the following, we will address this issue by analyzing the effects of *bonus payments* in addition to the baseline payments in FFS and CAP, respectively.

Given our experimental parameters, the average patient health benefit $\overline{H}$ is 7.92 in FFS and 8.01 in CAP (see Table 6). $\overline{H}$ significantly increases to 9.47 in FFS+P4P and to 9.51 in CAP+P4P ($p < 0.001$, Wilcoxon signed rank-test).[17] Also, the physicians' remuneration increases significantly($p < 0.001$, Wilcoxon signed rank-test). This finding is in line with, for example, Mullen et al. (2010) and does not come as a surprise, as subjects in our experiment do react to the increased incentives under P4P.

It has been argued that the key to an effective P4P system is the design of its elements (Epstein, 2012; Maynard, 2012; Kristensen et al., 2016). To tackle this argument, we take a closer look at cost and benefits for the different severities of illness as the systematic variation of physicians' incentive payments for severities is an important design element of our experiment. We find that patients' health benefits and physicians' remunerations significantly increase for all severities ($p < 0.010$, Wilcoxon signed rank-test). Under CAP, the increase in health benefit is highest for the severely ill patients (43.7%), while under FFS the increase is highest for mildly ill patients (39.1%). This results in an increase in remuneration

Table 6: Patients' benefits, costs for physicians' remuneration, and changes in costs and benefits

	Aggregated		Mild severity		Interm. severity		High severity	
	$\overline{H}$	$\overline{R}$	$\overline{H}$	$\overline{R}$	$\overline{H}$	$\overline{R}$	$\overline{H}$	$\overline{R}$
FFS	7.92	13.38	6.72	11.38	7.92	13.38	9.10	15.38
FFS+P4P	9.51	15.02	9.35	12.93	9.52	14.75	9.65	17.38
Change	1.59	1.64	2.63	1.55	1.60	1.37	0.55	2.00
Ratio ($\Delta R/\Delta H$)	1.03		0.59		0.86		3.64	
CAP	8.01	10.00	9.15	10.00	8.03	10.00	6.86	10.00
CAP+P4P	9.47	13.77	9.53	12.31	9.51	13.53	9.36	15.46
Change	1.46	3.77	0.38	2.31	1.48	3.53	2.50	5.46
Ratio ($\Delta R/\Delta H$)	2.58		6.08		2.39		2.18	

Notes. This table shows the average patients' health benefits $\overline{H}$ and remuneration $\overline{R}$ for FFS, CAP, FFS+P4P, and CAP+P4P, both aggregated and differentiated for severities of illness (mild, intermediate, high). It further shows the marginal payment, marginal patient health benefit, and the ratio of marginal payment to marginal patient health benefit, also aggregated and separately for the three severities of illness.

[17] In absolute terms, the maximum health benefit achieved by patient-optimal service provision is 10.33 on average.

by 54.6% for the severely ill patients under CAP and by 13.6% for the mildly ill patients in FFS. The difference in relative changes between payment systems indicates that remuneration costs need to be taken into account when assessing the effectiveness of P4P.

To investigate this further, we focus on the ratio between remuneration and patient health benefits between non-blended and blended P4P payment systems. We find, that the financial resources needed to induce a one-unit increase in patient health benefit by physicians' treatment decisions varied substantially between payment systems. On average, 2.58 monetary units in CAP conditions and 1.03 units in FFS conditions are needed for a one-unit increase in patient benefit. Under CAP, the ratio is lowest for severely ill patients (2.18), due to the large increase in patient health benefit. The ratio is highest for mildly ill patients (6.08), driven by the rather small increase of 4.2% in patient health benefit. Under FFS, the ratio is 0.59 for patients with a mild severity of illness, while for patients with an intermediate severity the ratio is 0.86. This implies an increase in remuneration by less than one monetary unit for an increase in patients' health benefit by one unit. For severely ill patients, the ratio is 3.64.

We are aware that calculating ratios of marginal payment and marginal patient health benefit from our experimental data can only serve as a rough qualitative benchmark for comparing the effectiveness of performance pay. Our results suggest that incentivizing physicians' medical service provision with P4P is advisable in general for health care policy-makers, aiming primarily at enhancing the patients' health benefit, regardless of the additional costs generated. Taken at face value, introducing P4P for mildly ill patients under FFS and for severely ill patients under CAP would be most effective.

We next take into account that changing the baseline payment system from CAP to FFS and vice versa could provide an alternative to introducing P4P. The ratio of marginal physician payment to marginal patient health benefit is 0.58 when switching from FFS to CAP for mildly ill patients, and 2.40 when moving from CAP to FFS for highly ill patients. Hence, the effects of interchanging the baseline payment systems are similar to those when introducing P4P. The latter option may be favorable as it leads to an increase in patients' benefits for all severity types at the aggregate.

5 Concluding remarks

While performance pay for physicians has increasingly become popular among health policy makers and has been implemented in practice, the effects on physicians' provision behavior and patients' health benefits are still not well understood. To contribute in narrowing this gap, we conducted controlled laboratory and artefactual field experiments to analyze the causal effect of pay for performance on medical service provision. At a within-subject level, we introduced P4P which either complements FFS or CAP – with performance thresholds tied to the patient-optimal treatment and adjusted for the patients' severity of illness levels. Under P4P, subjects increase, on average, the quality of health care provision compared to non-blended payments. We unpack the positive P4P effect by finding that its intensity is significantly driven by the patients' severity of illness. At a between-subject level, we determine further how the severity-specific behavioral responses to introducing P4P differ dependent on the baseline payment systems. For intermediately ill patients, the increase in quality of medical services is nearly the same under both payment systems when introducing P4P. Mildly ill patients, however, marginally benefit the most when P4P is complementing FFS, while, for highly ill patients, this is true when P4P is complementing CAP.

In our parsimonious experimental design, we reduced the complexity of a physician's treatment decisions, abstracted from multitasking, considered a one-dimensional quality, and we refrained from measurement issues of a physician's quality of treatment. In contrast, we focused on *exogenously* introducing P4P while keeping all other variables constant. By doing so, we take advantage of the features and possibilities of controlled economics experiments to test for causal evidence (Falk and Heckman, 2009)—conditions that are rarely provided in the field. Taking a more general perspective, effective research needs to combine and balance insights from highly internally and externally valid methods. A controlled lab experiment has high internal validity and serves as a complement rather than a substitute for other research methods with high external validity. It could, for instance, work as a 'wind tunnel study', which allows us to rather inexpensively test for the behavioral effects of important design elements of P4P prior to implementing these elements in a large-scale randomized controlled trial (RCT), or before introducing policy measures in the field (Galizzi and Wiesen, 2018). Moreover, a combination of theory and experiments by economic engineering has improved the design and functioning of markets and institutions (Falk and Heckman, 2009). Examples in health care are the matching of doctors to entry level positions in the general medical labor market (Roth, 2002) or testing clinical decision support systems to improve hospital discharge decisions (Cox et al., 2016a).

In our experiment, P4P characterized by a 20%-bonus effectively induced a higher quality of medical service provision. Moreover, adjusting P4P for the patients' severity of illness reduced the strong overtreatment of low-severity patients under FFS and the strong undertreatment of high-severity patients under CAP. We designed P4P bonus payments such that performance thresholds are tied to the patient-optimal care and precisely varied bonus sizes to account for severity-specific patient benefits. It might not always be feasible to adequately design P4P bonus payments outside the laboratory, yet a general distinction between patient groups of rather high and low medical needs should be possible. In such cases, patients belonging to the former group should always be treated under FFS, while patients with rather low medical needs should be treated under CAP. This approach would guarantee that the harm is kept small, which deviations from the patient-optimal medical care are causing induced by opposing financial incentives between physician profit and patient benefit.

The cost-effectiveness analyses of our data shows that the additional expenditures for physicians' bonuses rise disproportionately although introducing P4P does induce increases in the patients' health benefit. Given the design of the experiment, our calculations are limited to incentive costs. Yet, other 'cost categories' might be affected by introducing P4P like set up/development costs, running costs, provider costs when participating in the scheme, as well as cost savings (Meacock et al., 2014). The latter category seems likely to apply as P4P induces care with superior health outcomes, which in turn will reduce future health care costs. Obviously, a health care policy-maker needs to take these considerations into account when evaluating the (cost-)effectiveness of different P4P schemes.

Finally, our behavioral results also evidence heterogeneities in responses to P4P. This calls for future work to better understand what drives this heterogeneity. What is the role, for example, of individuals' underlying social preferences, attitudes or personality traits? These individual characteristics might not only explain health care workers' responses to performance pay (e.g., Donato et al., 2017) but also self-selection into payment systems (e.g., Dohmen and Falk, 2011; Brosig-Koch et al., 2017b). Understanding the preferences that predict sorting are therefore of great importance for researchers and health care policy-makers alike. Policy-makers also need to account for unintended consequences of incentive schemes that enable physicians to manipulate the systems by only treating patients for whom they can generate the performance-based bonus payments (Alexander, 2020). Finally, further cost-benefit analyses and feasibility research are needed to systematically compare the effects of different payment menus, e.g., of mixing different non-blended payment systems, as opposed to introducing performance pay.

References

Alexander, D. (2020). How do doctors respond to incentives? Unintended consequences of paying doctors to reduce costs. *Journal of Political Economy*, 128:4046–4096. doi:10.1086/710334.

Allard, M., Jelovac, I., and Léger, P. (2011). Treatment and referral decisions under different physician payment mechanisms. *Journal of Health Economics*, 30:880–893. doi:10.1016/j.jhealeco.2011.05.016.

Andreoni, J. (1989). Giving with impure altruism: Applications to charity and ricardian equivalence. *Journal of Political Economy*, 97:1447–1458. doi:10.1086/261662.

Angerer, S., Glätzle-Rützle, D., and Waibel, C. (2021). Monitoring institutions in health care markets: Experimental evidence. *Health Economics*, 30:951–971. doi:10.1002/hec.4232.

Anselmi, L., Borghi, J., Brown, G. W., Fichera, E., Hanson, K., Kadungure, A., Kovacs, R., Kristensen, S. R., Singh, N. S., and Sutton, M. (2020). Pay for performance: A reflection on how a global perspective could enhance policy and research. *International Journal of Health Policy and Management*, 9:365–369. doi:10.34172/ijhpm.2020.23.

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review*, 53:941–969. doi:10.1515/9780822385028–004.

Baicker, K. and Goldman, D. (2011). Patient cost-sharing and healthcare spending growth. *Journal of Economic Perspectives*, 25:47–68. doi:10.1257/jep.25.2.47.

Brosig-Koch, J., Hennig-Schmidt, H., Kairies-Schwarz, N., Kokot, J., and Wiesen, D. (2020). Physician performance pay: Experimental evidence. HERO Online Working Paper Series 2020:3, University of Oslo, Health Economics Research Programme.

Brosig-Koch, J., Hennig-Schmidt, H., Kairies-Schwarz, N., and Wiesen, D. (2016). Using artefactual field and lab experiments to investigate how fee-for-service and capitation affect medical service provision. *Journal of Economic Behavior & Organization*, 131, Part B:17–23. doi:10.1016/j.jebo.2015.04.011.

Brosig-Koch, J., Hennig-Schmidt, H., Kairies-Schwarz, N., and Wiesen, D. (2017a). The effects of introducing mixed payment systems for physicians: Experimental evidence. *Health Economics*, 26:243 – 262. doi:10.1002/hec.3292.

Brosig-Koch, J., Kairies-Schwarz, N., and Kokot, J. (2017b). Sorting into payment schemes and medical treatment: A laboratory experiment. *Health Economics*, 26:52–65. doi:10.1002/hec.3616.

Byambadalai, U., Ma, A., and Wiesen, D. (2019). Changing preferences: An experiment and estimation of market-incentive effects on altruism. Working paper, Boston University.

Campbell, S. M., Reeves, D., Kontopantelis, E., Sibbald, B., and Roland, M. (2009). Effects of pay for performance on the quality of primary care in England. *New England Journal of Medicine*, 361:368–378. doi:10.1056/NEJMsa0807651.

Chalkley, M. and Malcomson, J. M. (1998). Contracting for health services when patient demand does not reflect quality. *Journal of Health Economics*, 17:1–19. doi:10.1016/S0167–6296(97)00019–2.

Choné, P. and Ma, C. (2011). Optimal health care contract under physician agency. *Annales d'Economie et de Statistique*, 101/102:229–256. doi:10.2307/41615481.

Clark, J. R. and Huckman, R. S. (2012). Broadening focus: Spillovers, complementarities, and specialization in the hospital industry. *Management Science*, 58:708–722. doi:10.1287/mnsc.1110.1448.

Clemens, J. and Gottlieb, J. D. (2014). Do physicians' financial incentives affect medical treatment and patient health? *American Economic Review*, 104:1320–1349. doi:10.1257/aer.104.4.1320.

Cox, J. C., Sadiraj, V., Schnier, K. E., and Sweeney, J. F. (2016a). Higher quality and lower cost from improving hospital discharge decision making. *Journal of Economic Behavior Organization*, 131, Part B:1 – 16. doi:10.1016/j.jebo.2015.03.017.

Cox, J. C., Sadiraj, V., Schnier, K. E., and Sweeney, J. F. (2016b). Incentivizing cost-effective reductions in hospital readmission rates. *Journal of Economic Behavior & Organization*, 131, Part B:24 – 35. doi:10.1016/j.jebo.2015.03.014.

Crea, G., Galizzi, M. M., Linnosmaa, I., and Miraldo, M. (2019). Physician altruism and moral hazard: (no) evidence from Finnish national prescriptions data. *Journal of Health Economics*, 65:153 – 169. doi:10.1016/j.jhealeco.2019.03.006.

DellaVigna, S., List, J. A., and Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *Quarterly Journal of Economics*, 127:1–56. doi:10.1093/qje/qjr050.

Di Guida, S., Gyrd-Hansen, D., and Oxholm, A. S. (2019). Testing the myth of fee-for-service and overprovision in health care. *Health Economics*, 28:717–722. doi:10.1002/hec.3875.

Dohmen, T. and Falk, A. (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American Economic Review*, 101:556–590. doi:10.1257/aer.101.2.556.

Donato, K., Miller, G., Mohanan, M., Truskinovsky, Y., and Vera-Hernández, M. (2017). Personality traits and performance contracts: Evidence from a field experiment among maternity care providers in India. *American Economic Review*, 107:506–10. doi:10.1257/aer.p20171105.

Doran, T., Fullwood, C., Gravelle, H., Reeves, D., Kontopantelis, E., Hiroeh, U., and Roland, M. (2006). Pay-for-performance programs in family practices in the United Kingdom. *New England Journal of Medicine*, 355:375–384. doi:10.1056/NEJMsa055505.

Dulleck, U. and Kerschbamer, R. (2006). On doctors, mechanics, and computer specialists: The economics of Credence Goods. *Journal of Economic Literature*, 44:5–42. doi:10.1257/002205106776162717.

Dulleck, U., Kerschbamer, R., and Sutter, M. (2011). The economics of Credence Goods: An experiment on the role of liability, verifiability, reputation, and competition. *American Economic Review*, 101:526–555. doi:10.1257/aer.101.2.526.

Eijkenaar, F., M. Emmert, M. Scheppach, and Oliver Schoeffski (2013). Effects of pay for performance in health care: A systematic review of systematic reviews. *Health Policy*, 110:115–130. doi:10.1016/j.healthpol.2013.01.008.

Ellis, R. P. and McGuire, T. G. (1986). Provider behavior under prospective reimbursement: Cost sharing and supply. *Journal of Health Economics*, 5:129–151. doi:10.1016/0167–6296(86)90002–0.

Ellis, R. P. and McGuire, T. G. (1990). Optimal payment systems for health services. *Journal of Health Economics*, 9:375–396. doi:10.1016/0167–6296(90)90001–J.

Emmert, M., Eijkenaar, F., Kemter, H., Esslinger, A., and Schöffski, O. (2012). Economic evaluation of pay-for-performance in health care: A systematic review. *European Journal of Health Economics*, 13:755–767. doi:10.1007/s10198–011–0329–8.

Epstein, A. M. (2012). Will pay for performance improve quality of care? The answer is in the details. *New England Journal of Medicine*, 367:1852–1853. doi:10.1056/NEJMe1212133.

Falk, A. and Heckman, J. J. (2009). Lab experiments are a major source of knowledge in social sciences. *Science*, 326:535–538. doi:10.1126/science.1168244.

Fischbacher, U. (2007). z-Tree: Zurich toolboox for readymade economic experiments – Experimenter's manual. *Experimental Economics*, 10:171–178. doi:10.1007/s10683–006–9159–4.

Galizzi, M. M. and Wiesen, D. (2017). Behavioural experiments in health: An introduction. *Health Economics*, 26:3–5. doi:10.1002/hec.3629.

Galizzi, M. M. and Wiesen, D. (2018). Behavioral experiments in health. In Hamilton, J., editor, *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press, Oxford.

Godager, G., Hennig-Schmidt, H., and Iversen, T. (2016). Does performance disclosure influence physicians' medical decisions? An experimental study. *Journal of Economic Behavior & Organization*, 131:36–46. doi:10.1016/j.jebo.2015.10.005.

Gravelle, H., Sutton, M., and Ma, A. (2010). Doctor behaviour under a pay for performance contract: Treating, cheating and case finding? *Economic Journal*, 120:129–156. doi:10.1111/j.1468–0297.2009.02340.x.

Green, E. P. (2014). Payment systems in the healthcare industry: An experimental study of physician incentives. *Journal of Economic Behavior & Organization*, 106:367–378. doi:10.1016/j.jebo.2014.05.009.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1:114–125. doi:10.1007/s40881–015–0004–4.

Greiner, B., Zhang, L., and Tang, C. (2017). Separation of prescription and treatment in health care markets: A laboratory experiment. *Health Economics*, 26:21–35. doi:10.1002/hec.3575.

Groß, M., Jürges, H., and Wiesen, D. (2021). The effects of audits and fines on upcoding in neonatology. *Health Economics Letter. doi:10.1002/hec.4272.*

Hellerstein, J. K. (1998). The importance of the physician in the generic versus trade-name prescription decision. *The Rand journal of economics*, 29:108–136.

Hennig-Schmidt, H., Jürges, H., and Wiesen, D. (2019). Dishonesty in health care practice: A behavioral experiment on upcoding in neonatology. *Health Economics*, 28:319–338. doi:10.1002/hec.3842.

Hennig-Schmidt, H., Selten, R., and Wiesen, D. (2011). How payment systems affect physicians' provision behavior – An experimental investigation. *Journal of Health Economics*, 30:637–646. doi:10.1016/j.jhealeco.2011.05.001.

Hennig-Schmidt, H. and Wiesen, D. (2014). Other-regarding behavior and motivation in health care provision: An experiment with medical and non-medical students. *Social Science & Medicine*, 108:156 – 165. doi:10.1016/j.socscimed.2014.03.001.

Huck, S., Lünser, G., Spitzer, F., and Tyran, J.-R. (2016). Medical insurance and free choice of physician shape patient overtreatment: A laboratory experiment. *Journal of Economic Behavior & Organization*, 131:78–105. doi:10.1016/j.jebo.2016.06.009.

Huesmann, K., Waibel, C., and Wiesen, D. (2020). Rankings in health care organizations. Working Paper, University of Cologne. doi:10.2139/ssrn.3690851.

Jack, W. (2005). Purchasing health care services from providers with unknown altruism. *Journal of Health Economics*, 24:73–93. doi:10.1016/j.jhealeco.2004.06.001.

Jia, L., Meng, Q., Scott, A., Yuan, B., and Zhang, L. (2021). Payment methods for healthcare providers working in outpatient healthcare settings. *Cochrane Database of Systematic Reviews*, 1. doi:10.1002/14651858.CD011865.pub2.

Keser, C., Peterle, E., and Schnitzler, C. (2014). Money talks-Paying physicians for performance. *Cege Discussion Paper, University of Göttingen, 173. doi:10.2139/ssrn.2357326.*

Kesternich, I., Schumacher, H., and Winter, J. (2015). Professional norms and physician behavior: Homo oeconomicus or homo hippocraticus? *Journal of Public Economics*, 131:1–11. doi:10.1016/j.jpubeco.2015.08.009.

Kolstad, J. T. (2013). Information and quality when motivation is intrinsic: Evidence from surgeon report cards. *American Economic Review*, 103:2875–2910. doi:10.1257/aer.103.7.2875.

Kristensen, S. R., Siciliani, L., and Sutton, M. (2016). Optimal price-setting in pay for performance schemes in health care. *Journal of Economic Behavior & Organization*, 123:57 – 77. doi:10.1016/j.jebo.2015.12.002.

Lagarde, M. and Blaauw, D. (2017). Physicians' responses to financial and social incentives: A medically framed real effort experiment. *Social Science & Medicine*, 179:147–159. doi:10.1016/j.socscimed.2017.03.002.

Li, J., Hurley, J., DeCicca, P., and Buckley, G. (2014). Physician response to pay-for-performance: Evidence from a natural experiment. *Health Economics*, 23:962–978. doi:10.1002/hec.2971.

Lindenauer, P. K., Remus, D., Roman, S., Rothberg, M. B., Benjamin, E. M., Ma, A., and Bratzler, D. W. (2007). Public reporting and pay for performance in hospital quality improvement. *New England Journal of Medicine*, 356:486–496. doi:10.1056/NEJMsa064964.

Liu, T. and Ma, C. (2013). Health insurance, treatment plan, and delegation to altruistic physician. *Journal of Economic Behavior & Organization*, 85:79 – 96. doi:10.1016/j.jebo.2012.11.002.

Ma, C. A. (1994). Health care payment systems: Cost and quality incentives. *Journal of Economics and Management Strategy*, 3:93–112. doi:10.1111/j.1430–9134.1994.00093.x.

Martinsson, P. and Persson, E. (2019). Physician behavior and conditional altruism: The effects of payment system and uncertain health benefit. *Theory and Decision*, 87:365–387. doi:10.1007/s11238–019–09714–7.

Mathes, T., Pieper, D., Morche, J., Polus, S., Jaschinski, T., and M., E. (2019). Pay for performance for hospitals. *Cochrane Database of Systematic Reviews*, 7. doi:10.1002/14651858.CD011156.pub2.

Maynard, A. (2012). The powers and pitfalls of payment for performance. *Health Economics*, 21:3–12. doi:10.1002/hec.1810.

McGuire, T. G. (2000). Physician Agency. In Cuyler and Newhouse, editors, *Handbook of Health Economics, Vol. 1 A*, pages 461–536. North-Holland, Amsterdam (The Netherlands).

Meacock, R., Kristensen, S. R., and Sutton, M. (2014). The cost-effectiveness of using financial incentives to improve provider quality: A framework and application. *Health Economics*, 23:1–13. doi:10.1002/hec.2978.

Milstein, R. and Schreyögg, J. (2016). Pay for performance in the inpatient sector: A review of 34 P4P programs in 14 OECD countries. *Health Policy*, 120:1125–1140. doi:10.1016/j.healthpol.2016.08.009.

Mimra, W., Rasch, A., and Waibel, C. (2016). Second opinions in markets for expert services: Experimental evidence. *Journal of Economic Behavior and Organization*, 131, Part B:106–125. doi:10.1016/j.jebo.2016.03.004.

Mullen, K., Frank, R., and Rosenthal, M. (2010). Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. *RAND Journal of Economics*, 41:64–91. doi:10.1111/j.1756–2171.2009.00090.x.

Olivella, P. and Siciliani, L. (2017). Reputational concerns with altruistic providers. *Journal of Health Economics*, 55:1 – 13. doi:10.1016/j.jhealeco.2017.05.003.

Oxholm, A.-S., Di Guida, S., and Gyrd-Hansen, D. (2021). Allocation of health care under pay for performance: Winners and losers. *Social Science & Medicine*, page 113939. doi:10.1016/j.socscimed.2021.113939.

Peckham, S. and Wallace, A. (2010). Pay for performance schemes in primary care: What have we learnt? *Quality in primary care*, 18:111–116.

Rammstedt, B. and John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German. *Journal of Research in Personality*, 41:203–212. doi:10.1016/j.jrp.2006.02.001.

Reif, S., Hafner, L., and Seebauer, M. (2020). Physician behavior under prospective payment schemes—Evidence from artefactual field and lab experiments. *International Journal of Environmental Research and Public Health*, 17:5540. doi:10.3390/ijerph17155540.

Roland, M. (2004). Linking physicians' pay to the quality of care - A major experiment in the United Kingdom. *New England Journal of Medicine*, 351:1448–1454. doi:10.1056/NEJMhpr041294.

Roland, M. and Campbell, S. (2014). Successes and failures of pay for performance in the United Kingdom. *New England Journal of Medicine*, 370:1944–1949. doi:10.1056/NEJMhpr1316051.

Rosenthal, M. B., Landon, B. E., Normand, S.-L. T., Frank, R. G., and Epstein, A. M. (2006). Pay for performance in commercial HMOs. *New England Journal of Medicine*, 355:1895–1902. doi:10.1056/NEJMsa063682.

Roth, A. E. (2002). The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica*, 70:1341–1378. doi:10.1111/1468–0262.00335.

Scott, A., Sivey, P., Ouakrim, D. A., Willenberg, L., Naccarella, L., Furler, J., and Young, D. (2011). The effect of financial incentives on the quality of health care provided by primary care physicians. *Cochrane Database of Systematic Reviews*, page 10.1002/14651858.CD008451.pub2.

Song, Z., Ji, Y., Safran, D. G., and Chernew, M. E. (2019). Health care spending, utilization, and quality 8 years into global payment. *New England Journal of Medicine*, 381:252–263. doi:10.1056/NEJMsa1813621.

Stokes, J., Struckmann, V., Kristensen, S. R., Fuchs, S., van Ginneken, E., Tsiachristas, A., Rutten van Mölken, M., and Sutton, M. (2018). Towards incentivising integration: A typology of payments for integrated care. *Health Policy*, 122:963 – 969. doi:10.1016/j.healthpol.2018.07.003.

Waibel, C. and Wiesen, D. (2020). An experiment on referrals in health care. *European Economic Review*, page 103612. doi:10.1016/j.euroecorev.2020.103612.

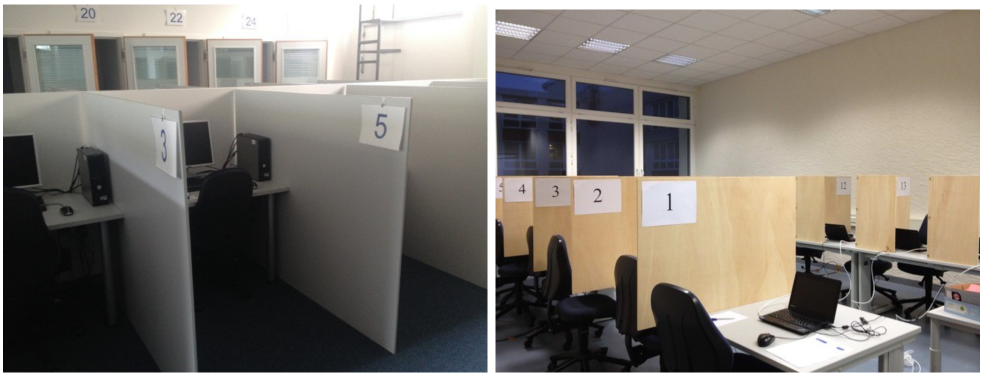
Wang, J., Iversen, T., Hennig-Schmidt, H., and Godager, G. (2020). Are patient-regarding preferences stable? Evidence from a laboratory experiment with physicians and medical students from different countries. *European Economic Review*, 125:103411. doi:10.1016/j.euroecorev.2020.103411.

Appendix

A Further details about the experiment

A.1 Laboratory setup

Figure A.1: Mobile and computer laboratory



Notes: This figure shows the laboratory setup for the computer laboratory experiments at elfe at the University of Duisburg-Essen (left panel) for our student sample and the mobile laboratory setup of elfe at the Academy for Training and Education in Bad Nauheim for our physician sample.

A.1.1 Sample

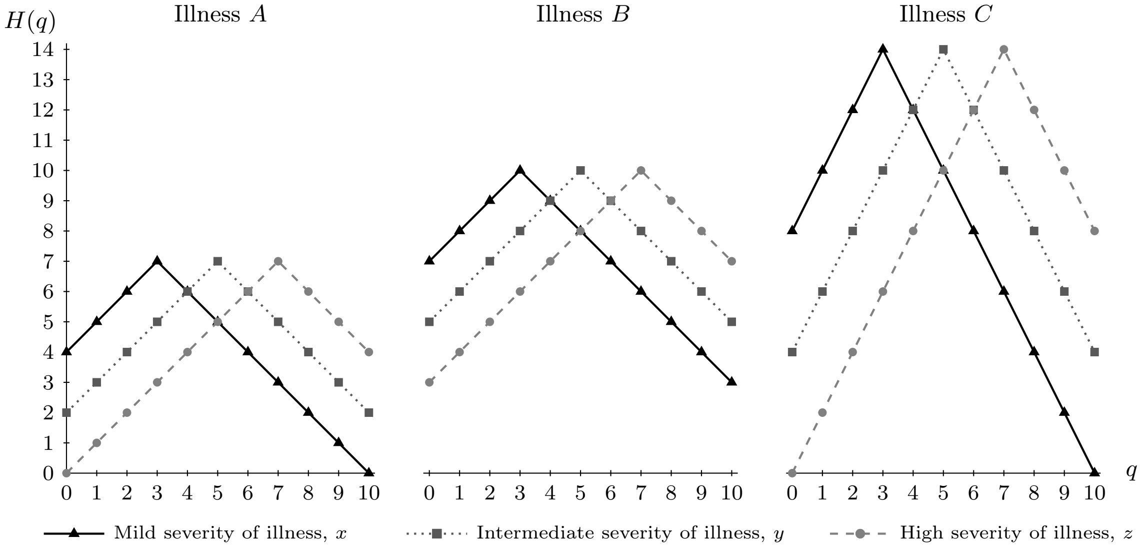
Table A.1: Sample characteristics

	Sample			
	Full	Med. students	Non-med. students	Physicians
A. All payment systems				
Main characteristics				
Male	40.2%	27.3%	53.5%	40.0%
Age (Mean, s.d.)	28.6 (9.8)	25.2 (7.0)	24.3 (3.4)	45.3 (6.7)
Personality traits (Mean, s.d.)				
Extraversion	3.6 (0.83)	3.7 (0.84)	3.5 (0.90)	3.5 (0.56)
Neuroticism	2.8 (0.97)	2.6 (0.91)	3.0 (0.92)	2.6 (1.10)
Openness	3.6 (0.92)	3.8 (0.88)	3.3 (0.97)	3.6 (0.82)
Conscientiousness	3.6 (0.81)	3.6 (0.69)	3.2 (0.81)	4.3 (0.53)
Agreeableness	3.1 (0.71)	3.3 (0.63)	2.8 (0.73)	3.1 (0.64)
N	107	44	43	20
B. FFS				
Main characteristics				
Male	44.2%	40.9%	50.0%	40.0%
Age (Mean, s.d.)	28.4 (9.9)	24.3 (4.5)	24.1 (4.0)	45.9 (7.2)
Personality traits (Mean, s.d.)				
Extraversion	3.7 (0.86)	3.8 (0.89)	3.7 (0.92)	3.4 (0.66)
Neuroticism	2.8 (0.92)	2.4 (0.82)	3.1 (0.85)	3.0 (0.08)
Openness	3.5 (0.93)	3.9 (0.85)	3.3 (1.00)	3.5 (0.80)
Conscientiousness	3.5 (0.79)	3.5 (0.66)	3.1 (0.82)	4.1 (0.61)
Agreeableness	3.1 (0.67)	3.3 (0.66)	2.9 (0.61)	3.2 (0.63)
N	52	22	20	10
C. CAP				
Main characteristics				
Male	36.4%	13.6%	56.5%	40.0%
Age (Mean, s.d.)	28.8 (9.9)	26.1 (8.8)	24.6 (2.9)	44.6 (8.2)
Personality traits (Mean, s.d.)				
Extraversion	3.4 (0.79)	3.5 (0.79)	3.4 (0.88)	3.4 (0.58)
Neuroticism	2.7 (1.02)	2.8 (0.99)	2.9 (0.99)	2.3 (1.11)
Openness	3.6 (0.91)	3.7 (0.92)	3.4 (0.95)	4.0 (0.69)
Conscientiousness	3.7 (0.82)	3.8 (0.70)	3.3 (0.81)	4.5 (0.44)
Agreeableness	3.0 (0.76)	3.2 (0.61)	2.9 (0.83)	3.0 (0.90)
N	55	22	23	10

Notes. This table presents summary statistics of subjects' characteristics for (i) the full sample of our experiment, (ii) for medical and (iii) non-medical students in the laboratory experiment and (iv) for physicians in the artifactual field experiment. We further differentiate between payment systems.

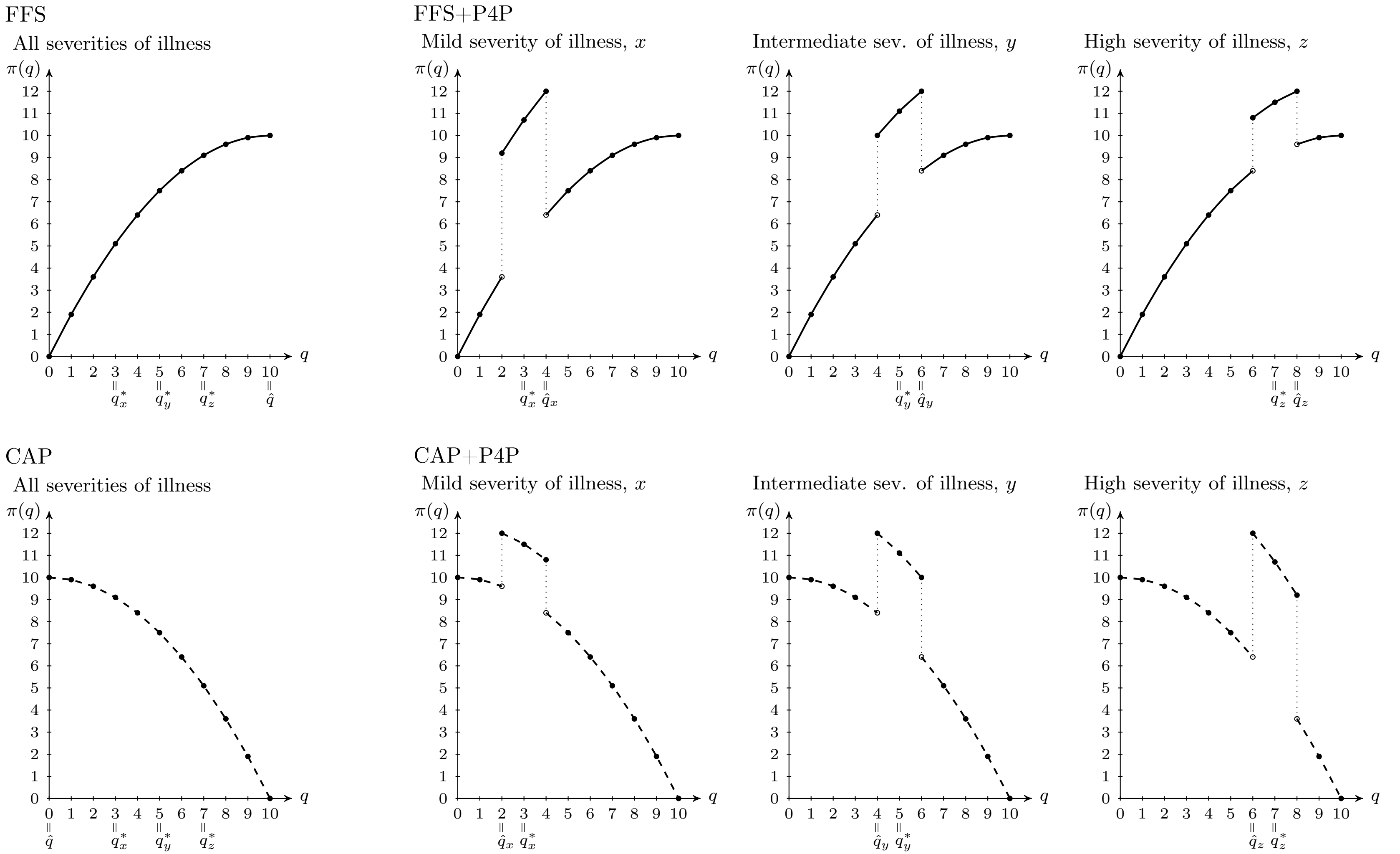
A.2 Parameters of the experiment

Figure A.2: Patient health benefits by illness and severity of illness



Notes: This figure illustrates patient health benefit parameters $H(q)$ for illnesses $k = A, B, C$ and severities of illness $l = x, y, z$ on the quantity interval from 0 to 10. The left panel shows patient benefits for illness A, the middle panel for illness B, and the right panel for illness C. The black solid line indicates severity of illness x, the grey dotted line severity of illness y, and the grey dashed line severity of illness z. For illness A and B, $\theta = 1$ and for illness C, $\theta = 2$. Notice that the patient health benefits are kept constant for all payment conditions.

Figure A.3: Profit parameters in FFS/FFS+P4P and CAP/CAP+P4P



FFS

All severities of illness

FFS+P4P

Mild severity of illness, x

Intermediate sev. of illness, y

High severity of illness, z

CAP

All severities of illness

CAP+P4P

Mild severity of illness, x

Intermediate sev. of illness, y

High severity of illness, z

Notes: The upper panel of the figure illustrates profits in FFS and FFS+P4P, the lower panel analogously illustrates profits in CAP and CAP+P4P. Under basic payments, profits increase (in FFS) and decrease (in CAP) continuously, regardless of the severity of illness on the quantity interval. In the pay for performance conditions, a bonus payment is granted if the performance threshold $|q - q^*| \leq 1$ is reached. As the patient-optimal quantity q^* depends on the severity of illness, the performance thresholds differ accordingly. In the basic payment condition, the profit-maximizing quantity is $\hat{q}$=10 in FFS and $\hat{q}$=0 in CAP, respectively. In the pay for performance condition, $\hat{q}$ changes depending on the severity of illness.

Table A.2: Parameters of main experimental conditions

	Quantity (q)										
	0	1	2	3	4	5	6	7	8	9	10
Patient benefit											
B_{Ax}	4	5	6	7	6	5	4	3	2	1	0
B_{Ay}	2	3	4	5	6	7	6	5	4	3	2
B_{Az}	0	1	2	3	4	5	6	7	6	5	4
B_{Bx}	7	8	9	10	9	8	7	6	5	4	3
B_{By}	5	6	7	8	9	10	9	8	7	6	5
B_{Bz}	3	4	5	6	7	8	9	10	9	8	7
B_{Cx}	8	10	12	14	12	10	8	6	4	2	0
B_{Cy}	4	6	8	10	12	14	12	10	8	6	4
B_{Cz}	0	2	4	6	8	10	12	14	12	10	8
Costs											
c	0.0	0.1	0.4	0.9	1.6	2.5	3.6	4.9	6.4	8.1	10.0
FFS											
p	0.0	2.0	4.0	6.0	8.0	10.0	12.0	14.0	16.0	18.0	20.0
π	0.0	1.9	3.6	5.1	6.4	7.5	8.4	9.1	9.6	9.9	10.0
CAP											
L	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
π	10.0	9.9	9.6	9.1	8.4	7.5	6.4	5.1	3.6	1.9	0.0
FFS+P4P											
p	0.0	2.0	4.0	6.0	8.0	10.0	12.0	14.0	16.0	18.0	20.0
b_x	0.0	0.0	5.6	5.6	5.6	0.0	0.0	0.0	0.0	0.0	0.0
b_y	0.0	0.0	0.0	0.0	3.6	3.6	3.6	0.0	0.0	0.0	0.0
b_z	0.0	0.0	0.0	0.0	0.0	0.0	2.4	2.4	2.4	0.0	0.0
π_x	0.0	1.9	9.2	10.7	12.0	7.5	8.4	9.1	9.6	9.9	10.0
π_y	0.0	1.9	3.6	5.1	10.0	11.1	12.0	9.1	9.6	9.9	10.0
π_z	0.0	1.9	3.6	5.1	6.4	7.5	10.8	11.5	12.0	9.9	10.0
CAP+P4P											
L	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
b_x	0.0	0.0	2.4	2.4	2.4	0.0	0.0	0.0	0.0	0.0	0.0
b_y	0.0	0.0	0.0	0.0	3.6	3.6	3.6	0.0	0.0	0.0	0.0
b_z	0.0	0.0	0.0	0.0	0.0	0.0	5.6	5.6	5.6	0.0	0.0
π_x	10.0	9.9	12.0	11.5	10.8	7.5	6.4	5.1	3.6	1.9	0.0
π_y	10.0	9.9	9.6	9.1	12.0	11.1	10.0	5.1	3.6	1.9	0.0
π_z	10.0	9.9	9.6	9.1	8.4	7.5	12.0	10.7	9.2	1.9	0.0

Notes: This table shows the parameters used in our experiment for all payment conditions. p is the fee per service rendered to a patient in FFS, L is the lump-sum payment in CAP, $b_l^\bullet$ is the bonus paid when the quality requirement is met in FFS+P4P (CAP+P4P), and π is the physician's profit.

A.3 Instructions of the experiment

Notice that the text in squared brackets denotes [Capitation, CAP] conditions.

Welcome to the Experiment!

You are participating in an economic experiment on decision behavior. You and the other participants will be asked to make decisions for which you can earn money. Your payoff depends on the decisions you make. At the end of the experiment, your payoff will be converted to Euro and paid to you in cash. During the experiment, all amounts are presented in the experimental currency Taler. 10 Taler equal 8 Euro. The experiment will take about 90 minutes and consists of two parts. You will receive detailed instructions before each part. Note that none of your decisions in either part have any influence on the other part of the experiment.

Part I of the experiment

Please read the instructions carefully. We will approach you in about five minutes to answer any questions you may have. If you have questions at any time during the experiment, please raise your hand and we will come to you. Part I of the experiment consists of 9 rounds of decision situations.

Decision situation

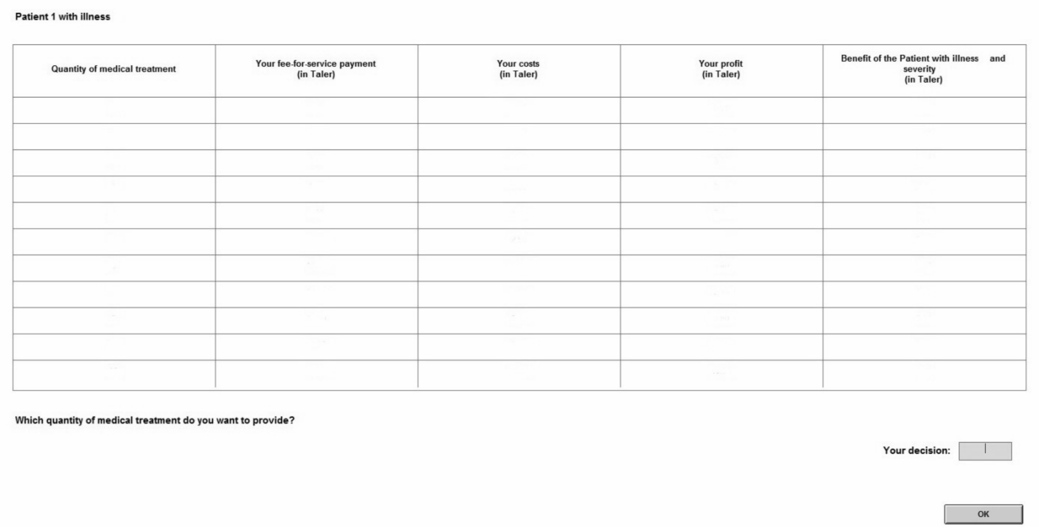
In each round, you are in the role of a physician and decide on medical treatment for a patient. That is, you determine the quantity of medical services you wish to provide to the patient for a given illness and a given severity of this illness. Each patient is characterized by one of three illnesses (A, B, C), each of which can occur in three different degrees of severity (x, y, z). In each consecutive decision round, you will face one patient who is characterized by one of the 9 possible combinations of illnesses and degrees of severity (in random order). Your decision is to provide each of these 9 patients with a quantity of $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$, or 10 medical services.

Payment

In each round, you receive a fee-for-service [capitation] remuneration for treating the patient. Your remuneration increases with the amount of medical treatment [irrespective of the amount of medical treatment] you provide. You also incur costs for treating the patient, which likewise depend on the quantity of services you provide. Your profit for each decision is calculated by subtracting these costs from the fee-for-service [capitation] remuneration. Each quantity of medical service yields a particular benefit for the patient—contingent on his illness and severity. Hence, in choosing the medical services you provide, you determine not only your own profit but also the patient's benefit.

In each round you will receive detailed information on your screen (see below) for the respective patient, the illness, your amount of fee-for-service [capitation] remuneration—for each possible amount of medical treatment—your costs, profit, as well as the benefit for the patient with the corresponding illness and severity.

Fee-for-service, FFS:



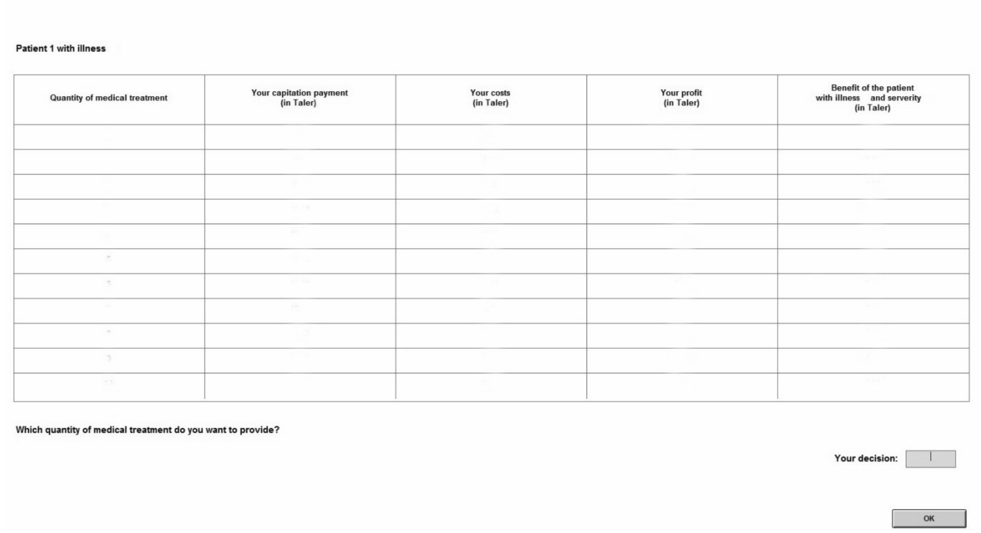
Patient 1 with illness

Quantity of medical treatment	Your fee-for-service payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the Patient with illness and severity (in Taler)

Which quantity of medical treatment do you want to provide?

Your decision:

[Capitation, CAP:]



Patient 1 with illness

Quantity of medical treatment	Your capitation payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness and serverity (in Taler)

Which quantity of medical treatment do you want to provide?

Your decision:

Payoff

At the end of the experiment, one of the 9 rounds in part *I* will be chosen at random. Your profit in that round will be paid to you in cash.

For this part of the experiment, no patients are physically present in the laboratory. Yet the patient benefit does accrue to a real patient: The amount resulting from your decision will be transferred to the Christoffel Blindenmission Deutschland e.V., 64625 Bensheim, which will use the money for enabling the treatment of patients with eye cataract.

The transfer of money to the Christoffel Blindenmission Deutschland e.V. will be carried out after the experiment by the experimenter and one participant. The participant completes a money transfer form, filling in the total patient benefit (in Euro) resulting from the decisions made by all participants in the randomly chosen situation. This form prompts the payment of the designated amount to the Christoffel Blindenmission Deutschland e.V. by the finance department of the University of Duisburg-Essen. The form is then sealed in a stamped envelope and deposited in the nearest mailbox by the participant and the experimenter.

After the entire experiment is completed, one participant is chosen at random to oversee the money transfer to the Christoffel Blindenmission Deutschland e.V. The participant receives an additional compensation of 5 Euro for this task. The participant certifies that the process has been completed as described here by signing a statement that can be inspected by all participants at the office of the Chair of Quantitative Economic Policy. A receipt of the bank transfer to the Christoffel Blindenmission Deutschland e.V. may also be viewed here.

Comprehension questions

Prior to the decision rounds, we kindly ask you to answer a few comprehension questions. They are intended to help you familiarize yourself with the decision situations. If you have any questions about this, please raise your hand. Part I of the experiment will begin once all participants have answered all comprehension questions correctly.

Part II of the experiment

Please read the instructions carefully. We will approach you in about five minutes to answer any questions you may have. If you have questions at any time during the experiment, please raise your hand and we will come to you. Part II of the experiment also consists of 9 rounds of decision situations.

Decision situation

As in part I of the experiment, you take on the role of a physician in each round and decide on medical treatment for a patient. That is, you determine the quantity of medical services you wish to provide to the patient for a given illness and a given severity of this illness.

Each patient is characterized by one of three illnesses (A, B, C), each of which can occur in three different degrees of severity (x, y, z). In each consecutive decision round, you will face one patient who is characterized by one of the 9 possible combinations of illnesses and degrees of severity (in random order). Your decision is to provide each of these 9 patients with a quantity of $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$, or 10 medical services.

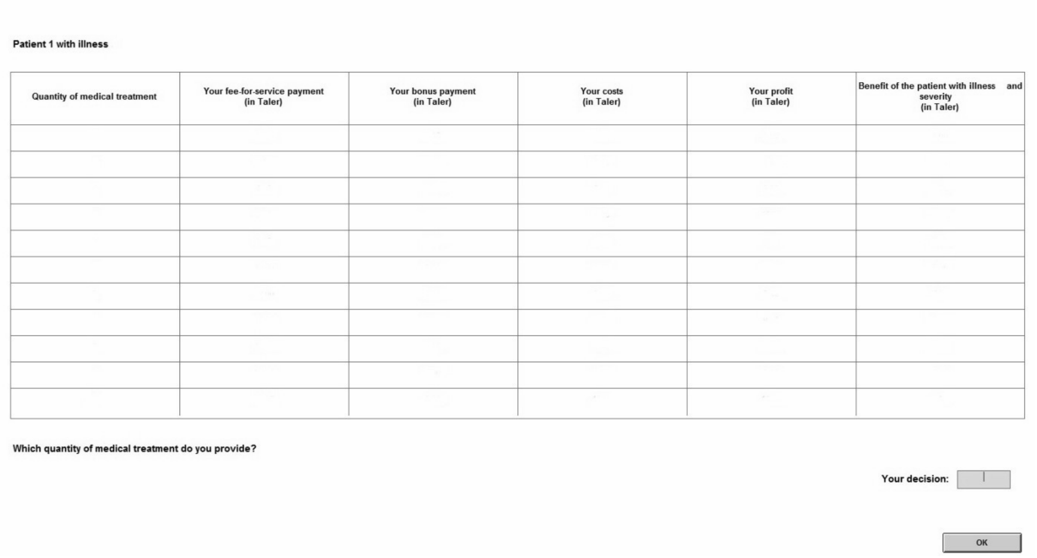
Payment

In each round, you are remunerated for treating the patient. In each round, you receive a fee-for-service [capitation] remuneration for treating the patient. Your remuneration increases with the amount of medical treatment [is irrespective of the amount of medical treatment] you provide. In addition to this, in each round you receive a bonus payment, in case the quantity of medical services you provide is equal to the one that results in the highest benefit for the patient, or deviates by one quantity from the latter. You also incur costs for treating the patient, which likewise depend on the quantity of services you provide. Your profit for each decision is calculated by subtracting these costs from the sum of your fee-for-service [capitation] remuneration and bonus payment.

As in part I, every quantity of medical service yields a particular benefit for the patient contingent on his illness and severity. Hence, in choosing the medical services you provide, you determine not only your own profit, but also the patient's benefit.

In each round, you will receive detailed information on your screen (see below) for the respective patient, the illness, your amount of fee-for-service [capitation] remuneration—for each possible amount of medical treatment, the amount of your bonus payment, your costs, profit, as well as the benefit for the patient with the corresponding illness and severity.

FFS+P4P:



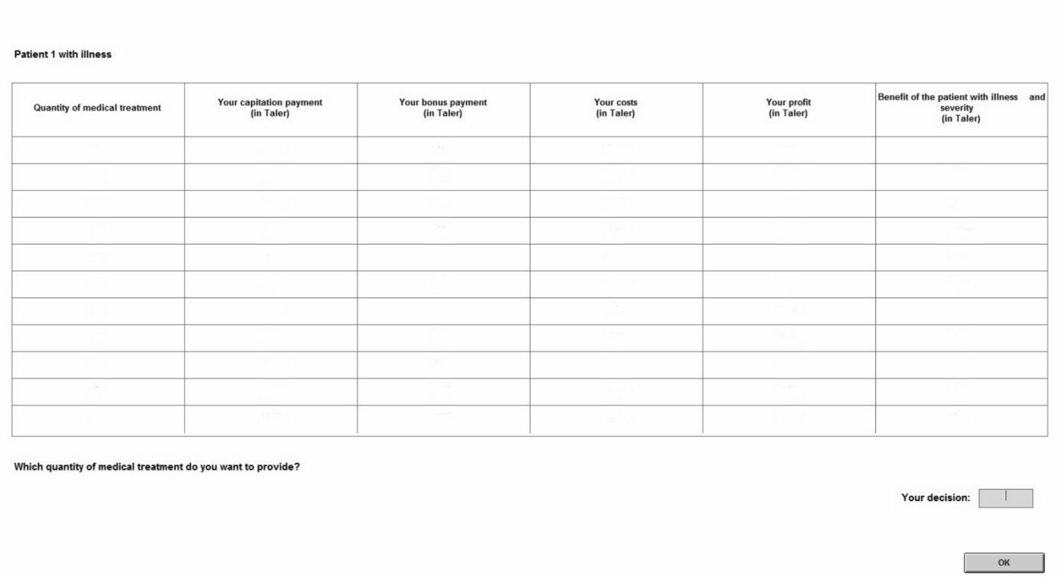
Patient 1 with illness

Quantity of medical treatment	Your fee-for-service payment (in Taler)	Your bonus payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness and severity (in Taler)

Which quantity of medical treatment do you provide?

Your decision:

[CAP+P4P:]



Patient 1 with illness

Quantity of medical treatment	Your capitation payment (in Taler)	Your bonus payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness and severity (in Taler)

Which quantity of medical treatment do you want to provide?

Your decision:

Payoff

At the end of the experiment, one of the 9 rounds of part II will be chosen at random. Your profit in this round will be paid to you in cash, in addition to your payment from the round chosen for part I of the experiment. After the experiment is over, please remain seated until the experimenter asks you to step forward. You will receive your payment at the front of the laboratory before exiting the room.

As in part *I*, no patients are physically present in the laboratory for part *II* of the experiment. Yet the patient benefit does accrue to a real patient: The amount resulting from your decision will be transferred to the Christoffel Blindenmission Deutschland e.V., 64625 Bensheim, which will use the money for enabling the treatment of patients with eye cataract.

The process for transferring the money to the Christoffel Blindenmission Deutschland e.V., as described for part *I* of the experiment, will be carried out by the experimenter and one participant.

Comprehension Questions

Prior to the decision rounds, we kindly ask you to answer a few comprehension questions. They are intended to help you familiarize yourself with the decision situations. If you have any questions about this, please raise your hand. Part *II* of the experiment will begin once all participants have answered all comprehension questions correctly.

Finally, we kindly ask you to not talk to anyone about the content of this session in order to prevent influencing other participants after you. Thank you for your cooperation!

B Behavioral Predictions

Let physician i choose the quantity of medical services q in order to maximize her utility

$$U_i(q) = \alpha_i H(q) + (1-\alpha_i)\pi(q), \tag{6}$$

with $\alpha_i \in [0,1)$. α_i is a measure for physician i's altruism. For a purely profit-maximizing physician, for example, $\alpha_i = 0$. A profit-maximizing physician therefore obtains the highest utility, in the absence of P4P in our experiment, when choosing 10 medical services in FFS and when choosing 0 medical services in CAP.

First, we consider physician i's behavior under the baseline payment systems, i.e., FFS and CAP. For profits and patient benefits given in our experiment, and the given altruism of physician i, we state the following lemma:[18]

Lemma 1. *Physician i overprovides medical services ($q > q^*$) if $p > q^*/5 + (\alpha_i/(1-\alpha_i))\,\theta$, and she underprovides medical services ($q < q^*$) if $p < q^*/5 - (\alpha_i/(1-\alpha_i))\,\theta$. Otherwise, physician i chooses the patient optimal quantity ($q = q^*$).*

Proof. Physician i's objective function $U_i(q) = \alpha_i H(q) + (1-\alpha_i)\pi(q)$ is concave. Payment $R(q) = L + pq$ is linear and $-c(q)$ is concave as $c(q)$ is convex, thus $\pi(q)$ is a concave function. As $H(q)$ is also a concave function (as defined in Equation 2) and $\alpha_i \geq 0$, it follows that $U_i(q)$ is concave.

Note that as $H(q)$ is not differentiable at $q = q^*$, with $q^* \in (0, 10)$. For $q < q^*$, the first-order condition $U_i'(q) = (1-\alpha_i)\left[p - \frac{q}{5}\right] + \alpha_i\theta$. For $q > q^*$, the first-order condition $U'(q) = (1-\alpha_i)_i\left[p - \frac{q}{5}\right] - \alpha_i\theta$. For $q > q^*$, consider $\lim_{q \to q^*} U_i'(q) = (1-\alpha_i)_i\left[p - \frac{q^*}{5}\right] - \alpha_i\theta$. If $p < q^*/5 - (\alpha_i/(1-\alpha_i))\,\theta$, $\lim_{q\to q^*} U_i'(q)$ is positive. Also, because $U_i(q)$ is concave, $U_i'(q) > 0 \; \forall \; q < q^*$. Therefore any q such that $q \leq q^*$ cannot be optimal, i.e., physician i chooses $q > q^*$.

Analogously for $q < q^*$, consider $\lim_{q\to q^*} U_i'(q) = (1-\alpha_i)_i\left[p - \frac{q^*}{5}\right] + \alpha_i\theta$. If $p < q^*/5 + (\alpha_i/(1-\alpha_i))\,\theta$, $\lim_{q\to q^*} U_i'(q)$ is negative. Also because $U_i(q)$ is concave, $U_i'(q) < 0 \; \forall \; q > q^*$. Therefore any q such that $q \geq q^*$ cannot be optimal, i.e., physician i chooses $q < q^*$. □

It directly follows from Lemma 1 that physician i's provision behavior depends on the severity of illness (i.e., the patient-optimal quantity q^* varying with severity of illness l), the fee for a medical service p, the marginal patient health benefit θ, and α_i, the physician i's degree of altruism. Intuitively, the higher physician i's altruism is towards her patient, the lower the degree of non-optimal service provision is. Based on Lemma 1, we expect that FFS induces overprovision of medical services, which decreases in the severity of a patient's illness and in patients' marginal health benefit. On the contrary, we expect that CAP induces underprovision of medical services, which increases in the severity of a patient's illness, it decreases in patients' marginal health benefit.

We now focus on the effect of introducing P4P on physicians' health care service provision. Comparing physician i's provision behavior between FFS (CAP) and FFS+P4P (CAP+P4P), we state the following proposition:

[18] Notice that Lemma 1 is a special case of Proposition 1 in Brosig-Koch et al. (2017a). They consider a physician's behavior under mixed payment systems with a weight on a FFS component and a lump-sum CAP.

Proposition 1. *Performance pay linked to the optimal patient's health benefit reduces physicians' overprovision of medical services in fee-for-service and underprovision in capitation.*

Proof. Let $q^{\text{Opt.}}$ be a physician's utility-maximizing choice for a patient j under FFS or CAP. Depending on a physician's quantity choice, we distinguish three cases. First, we consider $q^{\text{Opt.}} \in [q^* - 1, q^* + 1]$. As $b_l > 0$, it follows that a physician with $\alpha_i \in [0, 1)$ does not change her behavior since $b_l > 0$ is a constant. Second, we consider $q^{\text{Opt.}} > q^* + 1$. Here, the physician chooses q according to $\max\{U^{II}(q^{\text{Opt.}}), U(q^* + 1) + b_l\}$. That means a physician either does not change her behavior or chooses $q^* + 1$ when P4P has been introduced. Analogously for $q^{\text{Opt.}} < q^* - 1$, the same logic applies. □

Intuitively, whether a physician meets the quality threshold ($|q - q^*| \leq 1$) depends on physician i's degree of altruism towards the patient, aaccording to Lemma 1, counterbalancing the incentive effects in FFS and CAP. For a physician's given altruism with $\alpha_i \in [0, 1)$, introducing P4P, therefore, reduces non-optimal service provision under FFS and CAP. Since former experimental evidence shows that non-optimal service provision is highest for those patients where the difference between for whom the incentive effects in FFS and CAP and the patient's optimal quantity are the most misaligned, $\hat{q}$ i.e., for mild severe ill patients under FFS and high severe ill patients under CAP, the effect sizes of P4P are also likely to vary between severity types. Thus, for a physician's given altruism with $\alpha_i \in [0, 1)$, we expect a larger effect of P4P on non-optimal service provision with increasing severity of illness under CAP and decreasing severity under FFS.

C Additional analyses

Table C.1: Quantities and qualities of medical service provision by patients' health characteristics and payment schemes

	FFS		FFS+P4P			CAP		CAP+P4P		
	Mean	s.d.	Mean	s.d.	p-value	Mean	s.d.	Mean	s.d.	p-value
A. Quantity of medical services q										
Aggregate	6.69	2.07	5.59	1.66	<0.001	3.32	2.13	4.40	1.71	<0.001
Mild severity	5.69	2.44	3.70	0.64	<0.001	2.23	1.22	2.55	0.83	0.0095
Intermediate severity	6.69	1.74	5.58	0.53	<0.001	3.35	1.84	4.38	0.72	<0.001
High severity	7.69	1.36	7.50	0.56	0.0759	4.38	2.55	6.27	0.81	<0.001
Low marginal health benefit	6.69	2.12	5.61	1.67	<0.001	3.23	2.18	4.37	1.69	<0.001
High marginal health benefit	6.70	1.96	5.56	1.63	<0.001	3.49	2.03	4.47	1.74	<0.001
B. Absolute deviation from patient-optimal care ρ										
Aggregate	1.82	1.95	0.63	0.55	<0.001	1.77	2.01	0.65	0.75	<0.001
Mild severity	2.73	2.40	0.74	0.59	<0.001	0.90	1.12	0.59	0.73	0.0035
Intermediate severity	1.79	1.64	0.62	0.49	<0.001	1.75	1.75	0.62	0.72	<0.001
High severity	0.95	1.19	0.53	0.54	<0.001	2.66	2.51	0.75	0.78	<0.001
Low marginal health benefit	1.85	2.00	0.64	0.54	<0.001	1.84	2.08	0.67	0.74	<0.001
High marginal health benefit	1.76	1.85	0.60	0.56	<0.001	1.64	1.86	0.63	0.77	<0.001
C. Proportional health benefit $\hat{H}$										
Aggregate	0.71	0.31	0.90	0.09	<0.001	0.71	0.32	0.90	0.12	<0.001
Mild severity	0.61	0.34	0.89	0.08	<0.001	0.87	0.16	0.92	0.10	0.0039
Intermediate severity	0.64	0.33	0.88	0.1	<0.001	0.65	0.35	0.88	0.14	<0.001
High severity	0.86	0.17	0.92	0.08	<0.001	0.62	0.36	0.89	0.11	<0.001
High marginal health benefit	0.71	0.30	0.90	0.09	<0.001	0.73	0.30	0.90	0.13	<0.001
Observations	468		468			495		495		
Subjects	52		52			55		55		

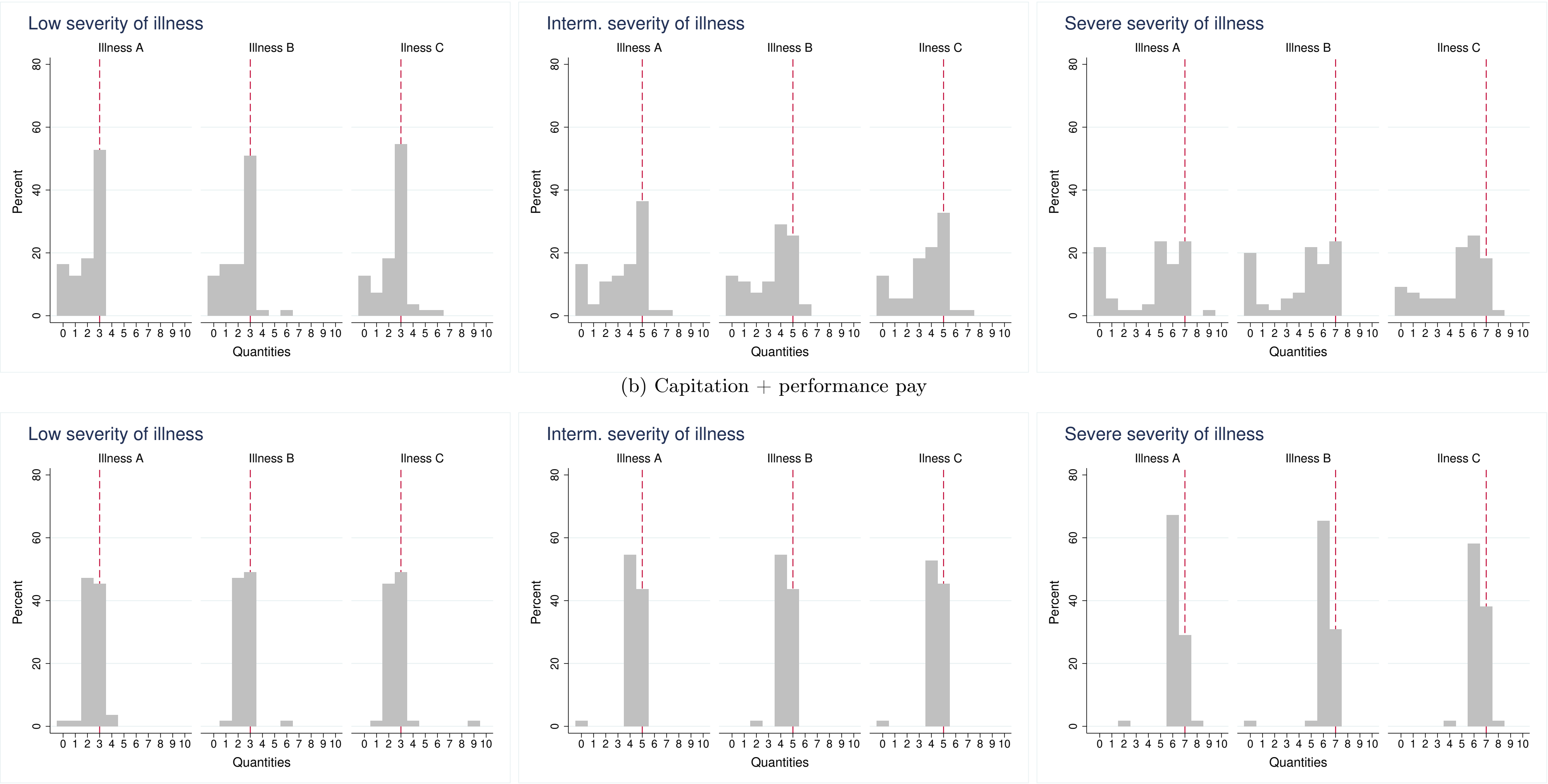
Notes: This table shows descriptive statistics on the quantity and quality of medical service provision for each payment condition, at the aggregate level and differentiated by patients' characteristics (levels of severity of illness and marginal health benefit). Two-sided p-values are shown for Wilcoxon sigend rank tests for differences in the quantity and quality measures across non-blended (FFS or CAP) and blended paymentsystem (FFS+P4P or CAP+P4P, respectively).

Table C.2: Quantity and quality of health care provision by payment system, illness, and severity of illness

	A. Quantity of medical services q				**B. Absolute deviation from optimal care ρ**				**C. Proportional health benefit $\hat{H}$**			
	unblended	+P4P	%-change	p-value	unblended	+P4P	%-change	p-value	unblended	+P4P	%-change	p-value
Fee-For-Service												
Mild severity of illness												
Illness A	5.77 (2.53)	3.65 (0.52)	-0.37	<0.001	2.81 (2.49)	0.69 (0.47)	-0.75	<0.001	0.60 (0.36)	0.90 (0.07)	0.50	<0.001
Illness B	5.67 (2.55)	3.73 (0.69)	-0.34	<0.001	2.75 (2.46)	0.77 (0.65)	-0.72	<0.001	0.61 (0.35)	0.89 (0.09)	0.47	<0.001
Illness C	5.63 (2.28)	3.71 (0.70)	-0.34	<0.001	2.63 (2.28)	0.74 (0.59)	-0.72	<0.001	0.62 (0.33)	0.89 (0.09)	0.43	<0.001
Interm. severity of illness												
Illness A	6.48 (1.82)	5.60 (0.53)	-0.14	<0.001	1.63 (1.68)	0.63 (0.49)	-0.61	<0.001	0.67 (0.34)	0.87 (0.10)	0.30	<0.001
Illness B	6.87 (1.69)	5.58 (0.54)	-0.19	<0.001	1.90 (1.65)	0.62 (0.49)	-0.67	<0.001	0.62 (0.33)	0.88 (0.10)	0.42	<0.001
Illness C	6.73 (1.73)	5.56 (0.54)	-0.17	<0.001	1.85 (1.60)	0.6 (0.5)	-0.68	<0.001	0.63 (0.32)	0.88 (0.10)	0.40	<0.001
Severe severity of illness												
Illness A	7.69 (1.57)	7.58 (0.64)	-0.01	0.5560	1.08 (1.33)	0.62 (0.6)	-0.43	0.0422	0.85 (0.19)	0.91 (0.09)	0.08	0.0422
Illness B	7.65 (1.37)	7.50 (0.50)	-0.02	0.4757	0.92 (1.20)	0.5 (0.50)	-0.46	0.0388	0.87 (0.17)	0.93 (0.07)	0.07	0.0655
Illness C	7.73 (1.12)	7.42 (0.54)	-0.04	0.0691	0.85 (1.04)	0.46 (0.50)	-0.46	0.0205	0.88 (0.15)	0.93 (0.07)	0.06	0.0205
Capitation												
Mild severity of illness												
Illness A	2.07 (1.15)	2.47 (0.69)	0.19	0.0088	0.93 (1.15)	0.6 (0.63)	-0.35	0.0592	0.87 (0.16)	0.91 (0.09)	0.05	0.0592
Illness B	2.20 (1.24)	2.55 (0.72)	0.16	0.1388	0.95 (1.13)	0.56 (0.63)	-0.32	0.0449	0.86 (0.16)	0.92 (0.09)	0.06	0.0717
Illness C	2.42 (1.26)	2.64 (1.04)	0.09	0.6947	0.84 (1.10)	0.62 (0.91)	-0.26	0.2786	0.88 (0.16)	0.91 (0.13)	0.04	0.2786
Interm. severity of illness												
Illness A	3.35 (1.94)	4.36 (0.78)	0.30	<0.001	1.76 (1.84)	0.64 (0.78)	-0.64	<0.001	0.65 (0.37)	0.87 (0.16)	0.35	<0.001
Illness B	3.24 (1.82)	4.40 (0.60)	0.36	<0.001	1.84 (1.74)	0.6 (0.6)	-0.67	<0.001	0.63 (0.35)	0.88 (0.12)	0.39	<0.001
Illness C	3.35 (1.80)	4.38 (0.78)	0.31	<0.001	1.65 (1.70)	0.62 (0.78)	-0.62	<0.001	0.67 (0.34)	0.88 (0.16)	0.31	<0.001
Severe severity of illness												
Illness A	4.27 (2.76)	6.25 (0.78)	0.46	<0.001	2.8 (2.68)	0.78 (0.74)	-0.72	<0.001	0.60 (0.38)	0.89 (0.11)	0.48	<0.001
Illness B	4.25 (2.59)	6.18 (0.98)	0.45	<0.001	2.75 (2.59)	0.82 (0.98)	-0.70	<0.001	0.61 (0.37)	0.88 (0.14)	0.45	<0.001
Illness C	4.60 (2.30)	6.38 (0.62)	0.39	<0.001	2.44 (2.26)	0.65 (0.58)	-0.73	<0.001	0.88 (0.16)	0.91 (0.08)	0.03	<0.001

Notes: This table shows descriptive statistics on the quantities and quality of medical service provision at the level of payment systems, illnesses, severities of illness (means and standard deviations in brackets). 23 non-medical, 22 medical students and 10 physicians decide in the CAP (amounting to a total 990 observations) and 20 non-medical, 22 medical students and 10 physicians in the FFS condition (936 observations). Two-sided p-values are shown for Wilcoxon sigend rank tests for matched samples.

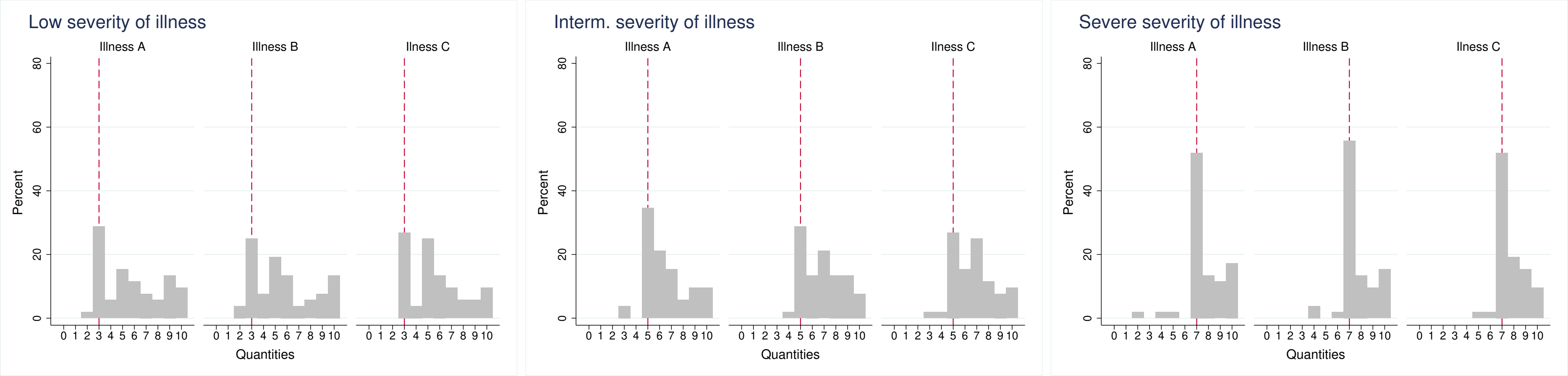
Figure C.1: Distributions of subjects' quantity choice by severity of illness under different payments scheme



(a) Capitation

(b) Capitation + performance pay

(c) Fee-For-Service



(d) Fee-For-Service + performance pay

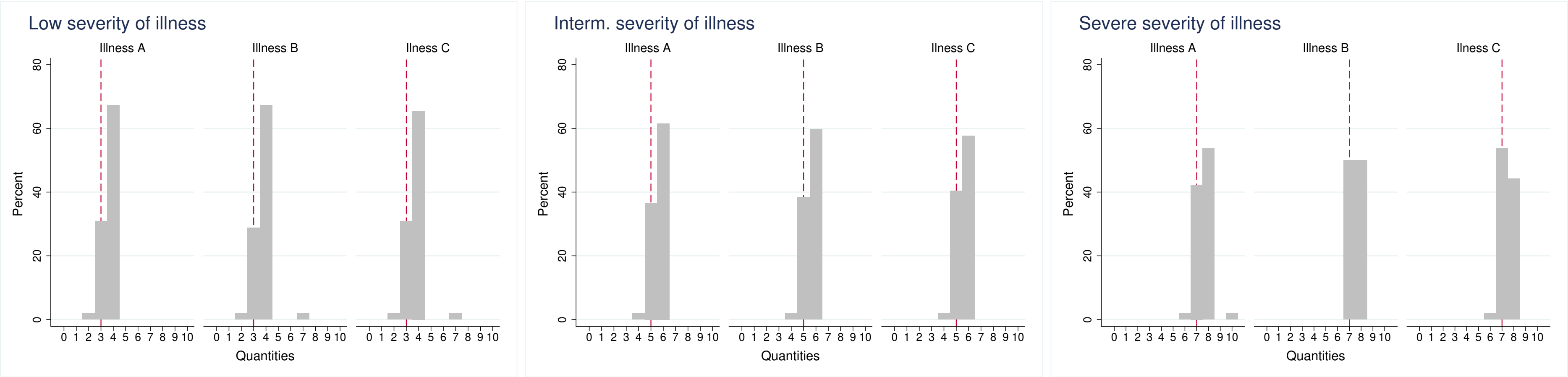


Table C.3: Regression models on the effect on quantity and quality between baseline CAP and FFS

	A. Quantity of medical services q		B. Absolute deviation from optimal care ρ		C. Proportional health benefit $\hat{H}$	
	(1)	(2)	(3)	(4)	(5)	(6)
CAP	-3.375***	-3.443***	-0.053	-0.063	0.009	0.009
	(0.320)	(0.309)	(0.300)	(0.267)	(0.048)	(0.042)
IntermSev	1.059***	1.059***	-0.019	-0.019	-0.099***	-0.100***
	(0.100)	(0.101)	(0.133)	(0.133)	(0.023)	(0.023)
HighSev	2.075***	2.075***	0.037	0.037	-0.006	-0.009
	(0.167)	(0.168)	(0.247)	(0.248)	(0.035)	(0.035)
HighMHB	0.139**	0.139**	-0.136**	-0.136**	0.020**	0.021**
	(0.057)	(0.057)	(0.056)	(0.056)	(0.009)	(0.009)
Medical students		-0.471		-0.365		0.050
		(0.358)		(0.346)		(0.057)
Physicians		-1.212**		-1.442***		0.236***
		(0.505)		(0.382)		(0.057)
Male		-0.468		0.340		-0.057
		(0.346)		(0.317)		(0.052)
Extraversion		0.444		-0.077		0.010
		(0.383)		(0.346)		(0.053)
Neuroticism		0.119		-0.126		0.018
		(0.321)		(0.315)		(0.054)
Openness		-0.097		-0.017		0.001
		(0.359)		(0.322)		(0.050)
Conscientiousness		0.930**		-0.347		0.053
		(0.456)		(0.409)		(0.061)
Agreeableness		0.721		-0.728*		0.121*
		(0.450)		(0.392)		(0.063)
Constant	5.601***	5.876***	1.864***	2.284***		
	(0.280)	(0.380)	(0.280)	(0.360)		
Observations	963	963	963	963	963	963
Observations	107	107	107	107	107	107
(Pseudo) R^2	0.493	0.534	0.001	0.127	0.009	0.066

Notes: This table shows parameter estimates from OLS regressions (Panel A and B) and average marginal effects from fractional probit response regressions (Panel C). Robust standard errors clustered for subjects are shown in parentheses. CAP = 1 if physicians are remunerated by CAP, and = 0 otherwise (by FFS). IntermSev and HighSev are dummy variables for intermediate and high severities of illness. HighMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). All models control for individual characteristics which comprise gender, medical background (non-medical student, medical student, physician), and personality traits.
* $p < 0.10$, ** $p < 0.05$, and, *** $p < 0.01$.

Table C.4: Regression models on the effect on quantity and quality under FFS, without individual control

	A. Quantity of medical services q			B. Absolute deviation from optimal care ρ			C. Proportional health benefit $\hat{H}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
P4P	-1.100***			-1.199***			0.191***		
	(0.184)			(0.171)			(0.027)		
IntermSev	1.439***	1.000***	1.439***	-0.529***	-0.936***	-0.529***	0.004	0.020*	0.004
	(0.086)	(0.147)	(0.086)	(0.085)	(0.148)	(0.085)	(0.010)	(0.012)	(0.010)
HighSev	2.901***	2.000***	2.901***	-0.997***	-1.782***	-0.997***	0.136***	0.188***	0.136***
	(0.128)	(0.229)	(0.128)	(0.141)	(0.255)	(0.141)	(0.019)	(0.026)	(0.019)
HighMHB	-0.016	-0.016	0.010	-0.054	-0.054	-0.074	0.008	0.008	0.007
	(0.053)	(0.053)	(0.088)	(0.051)	(0.051)	(0.086)	(0.008)	(0.008)	(0.010)
P4P × MildSev		-1.994***			-1.994***			0.188***	
		(0.276)			(0.276)			(0.024)	
P4P × IntermSev		-1.115***			-1.179***			0.163***	
		(0.191)			(0.182)			(0.023)	
P4P × HighSev		-0.192			-0.423***			0.077***	
		(0.132)			(0.111)			(0.017)	
P4P × LowMHB			-1.083***			-1.212***			0.172***
			(0.195)			(0.179)			(0.025)
P4P × HighMHB			-1.135***			-1.173***			0.154***
			(0.176)			(0.171)			(0.021)
Constant	5.251***	5.698***	5.243***	2.351***	2.749***	2.358***			
	(0.279)	(0.327)	(0.284)	(0.265)	(0.319)	(0.268)			
Wald test (p-value)									
H_0: P4P × MildSev =P4P × IntermSev		<0.001			<0.001			0.010	
H_0: P4P × MildSev= P4P × HighSev		<0.001			<0.001			<0.001	
H_0: P4P × IntermSev=P4P × HighSev		<0.001			<0.001			<0.001	
H_0: P4P × LowMHB=P4P × HighMHB			0.554			0.657			0.056
Observations	936	936	936	936	936	936	936	936	936
Subjects	52	52	52	52	52	52	52	52	52
(Pseudo) R^2	0.449	0.484	0.449	0.219	0.262	0.219	0.091	0.098	0.091

Notes: This table shows parameter estimates from OLS regressions (Panel A and B) and average marginal effects from fractional probit response regressions (Panel C). Robust standard errors clustered for subjects are shown in parentheses. P4P is a dummy variable indicating the introduction of P4P. IntermSev and HighSev are dummy variables for intermediate and high severities of illness. HighMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low).
* $p < 0.10$, ** $p < 0.05$, and, *** $p < 0.01$.

Table C.5: Regression models on the effect on quantity and quality under CAP, without individual controls

	A. Quantity of medical services			B. Abs. deviation from optimal care			C. Proportional health benefit		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
P4P	1.085***			-1.117***			0.178***		
	(0.188)			(0.179)			(0.029)		
IntermSev	1.473***	1.115***	1.473***	0.436***	0.848***	0.436***	-0.144***	-0.202***	-0.144***
	(0.100)	(0.138)	(0.100)	(0.073)	(0.137)	(0.074)	(0.018)	(0.025)	(0.018)
HighSev	2.933***	2.145***	2.933***	0.958***	1.758***	0.958***	-0.149***	-0.227***	-0.149***
	(0.151)	(0.244)	(0.151)	(0.133)	(0.249)	(0.133)	(0.020)	(0.030)	(0.020)
HighMHB	0.179***	0.179***	0.261***	-0.115**	-0.115**	-0.194**	0.017**	0.017**	0.023***
	(0.048)	(0.048)	(0.069)	(0.044)	(0.044)	(0.073)	(0.007)	(0.007)	(0.009)
P4P × MildSev		0.321**			-0.309***			0.057***	
		(0.140)			(0.108)			(0.018)	
P4P × IntermSev		1.036***			-1.133***			0.158***	
		(0.200)			(0.188)			(0.024)	
P4P × HighSev		1.897***			-1.909***			0.182***	
		(0.295)			(0.291)			(0.026)	
P4P × LowMHB			1.139***			-1.170***			0.166***
			(0.195)			(0.188)			(0.026)
P4P × HighMHB			0.976***			-1.012***			0.137***
			(0.192)			(0.172)			(0.021)
Constant	1.789***	2.171***	1.762***	1.345***	0.941***	1.372***			
	(0.180)	(0.147)	(0.183)	(0.173)	(0.133)	(0.178)			
Wald test (p-value)									
H_0: P4P × MildSev =P4P × IntermSev		<0.001			<0.001			<0.001	
H_0: P4P × MildSev= P4P × HighSev		<0.001			<0.001			<0.001	
H_0: P4P × IntermSev=P4P × HighSev		<0.001			<0.001			0.011	
H_0: P4P × LowMHB=P4P × HighMHB			0.096			0.048			0.004
Observations	990	990	990	990	990	990	990	990	990
Subjects	55	55	55	55	55	55	55	55	55
(Pseudo) R^2	0.432	0.457	0.432	0.180	0.221	0.180	0.082	0.090	0.082

Notes: This table shows parameter estimates from OLS regressions (Panel A and B) and average marginal effects from fractional probit response regressions (Panel C). Robust standard errors clustered for subjects are shown in parentheses. P4P is a dummy variable indicating the introduction of P4P. IntermSev and HighSev are dummy variables for intermediate and high severities of illness. HighMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low).
* $p < 0.10$, ** $p < 0.05$, and, *** $p < 0.01$.

Table C.6: Regression models on the effect on quantity and quality under FFS conditions with the full list of covariates

	A. Quantity of medical services q			**B. Absolute deviation from optimal care** ρ			**C. Proportional health benefit** $\hat{H}$		
Method:	OLS	OLS	OLS	OLS	OLS	OLS	Frac. Probit	Frac. Probit	Frac. Probit
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
P4P	-1.100***			-1.199***			0.189***		
	(0.185)			(0.172)			(0.025)		
IntermSev	1.439***	1.000***	1.439***	-0.529***	-0.936***	-0.529***	0.004	0.019	0.004
	(0.086)	(0.147)	(0.086)	(0.086)	(0.149)	(0.086)	(0.010)	(0.012)	(0.010)
HighSev	2.901***	2.000***	2.901***	-0.997***	-1.782***	0.997***	0.133***	0.184***	0.133***
	(0.129)	(0.230)	(0.129)	(0.141)	(0.256)	(0.141)	(0.016)	(0.023)	(0.016)
HighMHB	-0.016	-0.016	0.010	-0.054	-0.054	-0.074	0.009	0.009	0.008
	(0.053)	(0.053)	(0.089)	(0.051)	(0.051)	(0.087)	(0.008)	(0.008)	(0.011)
P4P × MildSev		-1.994***			-1.994***			0.187***	
		(0.277)			(0.277)			(0.021)	
P4P × IntermSev		-1.115***			-1.179***			0.162***	
		(0.192)			(0.183)			(0.021)	
P4P × HighSev		-0.192			-0.423***			0.079***	
		(0.132)			(0.111)			(0.017)	
P4P × LowMHB			-1.083***			-1.212***			0.171***
			(0.195)			(0.179)			(0.022)
P4P × HighMHB			-1.135***			-1.173***			0.153***
			(0.177)			(0.172)			(0.018)
Medical students	-0.402	-0.402	-0.402	-0.388	-0.388	-0.388	0.064	0.063	0.064
	(0.306)	(0.307)	(0.307)	(0.299)	(0.299)	(0.299)	(0.051)	(0.051)	(0.051)
Physicians	-1.909***	-1.909***	-1.909***	-1.520***	-1.520***	-1.520***	0.230***	0.228***	0.230***
	(0.270)	(0.271)	(0.270)	(0.266)	(0.267)	(0.267)	(0.041)	(0.040)	(0.041)
Male	0.043	0.043	0.043	0.050	0.050	0.050	-0.002	-0.002	-0.002
	(0.223)	(0.223)	(0.223)	(0.222)	(0.222)	(0.222)	(0.035)	(0.035)	(0.035)
Extraversion	0.227	0.227	0.227	0.275	0.275	0.275	-0.042	-0.042	-0.042
	(0.286)	(0.286)	(0.286)	(0.298)	(0.299)	(0.299)	(0.045)	(0.045)	(0.045)
Neuroticism	0.037	0.037	0.037	0.274	0.274	0.274	-0.046	-0.046	-0.046
	(0.255)	(0.255)	(0.255)	(0.258)	(0.258)	(0.258)	(0.042)	(0.042)	(0.042)
Openness	-0.151	-0.151	-0.151	-0.021	-0.021	-0.021	0.003	0.004	0.003
	(0.287)	(0.287)	(0.287)	(0.292)	(0.293)	(0.292)	(0.044)	(0.044)	(0.044)
Conscientiousness	0.477	0.477	0.477	0.428	0.428	0.428	-0.065	-0.065	-0.065
	(0.316)	(0.317)	(0.316)	(0.325)	(0.326)	(0.326)	(0.051)	(0.051)	(0.051)
Agreeableness	0.097	0.097	0.097	0.136	0.136	0.136	-0.030	-0.030	-0.030
	(0.306)	(0.307)	(0.306)	(0.315)	(0.315)	(0.315)	(0.053)	(0.053)	(0.053)
Constant	5.623***	6.070***	5.615***	2.621***	3.019***	2.627***			
	(0.315)	(0.350)	(0.318)	(0.315)	(0.354)	(0.317)			
Wald test (p-value)									
H_0: P4P × MildSev =P4P × IntermSev		<0.001			<0.001			0.010	
H_0: P4P × MildSev= P4P × HighSev		<0.001			<0.001			<0.001	
H_0: P4P × IntermSev=P4P × HighSev		<0.001			<0.001			<0.001	
H_0: P4P × LowMHB=P4P × HighMHB			0.556			0.658			0.062
Observations	936	936	936	936	936	936	936	936	936
Subjects	52	52	52	52	52	52	52	52	52
(Pseudo) R^2	0.563	0.599	0.563	0.336	0.379	0.336	0.150	0.157	0.150

Notes: This table shows parameter estimates from OLS regressions (Panel A and B) and average marginal effects from fractional probit response regressions (Panel C). Robust standard errors clustered for subjects are shown in parentheses. P4P is a dummy variable indicating the introduction of P4P. IntermSev and HighSev are dummy variables for intermediate and high severities of illness. HighMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). All models control for individual characteristics which comprise gender, medical background (non-medical student, medical student, physician), and personality traits.
* $p < 0.10$, ** $p < 0.05$, and, *** $p < 0.01$.

Table C.7: Regression models on the effect on quantity and quality under CAP conditions with the full list of covariates

	A. Quantity of medical services q			**B. Absolute deviation from optimal care ρ**			**C. Proportional health benefit $\hat{H}$**		
Method:	OLS	OLS	OLS	OLS	OLS	OLS	Frac. Probit	Frac. Probit	Frac. Probit
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
P4P	1.085***			-1.117***			0.175***		
	(0.189)			(0.180)			(0.026)		
IntermSev	1.473***	1.115***	1.473***	0.436***	0.848***	0.436***	-0.143***	-0.201***	-0.143***
	(0.100)	(0.139)	(0.100)	(0.074)	(0.138)	(0.074)	(0.016)	(0.024)	(0.016)
HighSev	2.933***	2.145***	2.933***	0.958***	1.758***	0.958***	-0.149***	-0.227***	-0.149***
	(0.151)	(0.245)	(0.151)	(0.134)	(0.250)	(0.134)	(0.019)	(0.028)	(0.019)
HighMHB	0.179***	0.179***	0.261***	-0.115**	0.115**	-0.194**	0.017**	0.017**	0.024***
	(0.048)	(0.048)	(0.069)	(0.044)	(0.044)	(0.073)	(0.007)	(0.007)	(0.009)
P4P × MildSev		0.321**			-0.309***			0.055***	
		(0.140)			(0.108)			(0.017)	
P4P × IntermSev		1.036***			-1.133***			0.157***	
		(0.201)			(0.189)			(0.021)	
P4P × HighSev		1.897***			-1.909***			0.180***	
		(0.296)			(0.292)			(0.023)	
P4P × LowMHB			1.139***			-1.170***			0.165***
			(0.195)			(0.188)			(0.024)
P4P × HighMHB			0.976***			-1.012***			0.135***
			(0.193)			(0.173)			(0.018)
Medical students	-0.165	-0.165	-0.165	0.192	0.192	0.192	-0.038	-0.039	-0.038
	(0.281)	(0.281)	(0.281)	(0.280)	(0.281)	(0.281)	(0.046)	(0.046)	(0.046)
Physicians	0.497	0.497	0.497	-0.438	-0.438	-0.438	0.073	0.072	0.073
	(0.324)	(0.324)	(0.324)	(0.322)	(0.323)	(0.323)	(0.048)	(0.048)	(0.048)
Male	-0.436*	-0.436*	-0.436*	0.385	0.385	0.385	-0.057	-0.057	-0.057
	(0.258)	(0.258)	(0.258)	(0.263)	(0.263)	(0.263)	(0.045)	(0.045)	(0.045)
Extraversion	0.497*	0.497*	0.497*	-0.430	-0.430	-0.430	0.059	0.059	0.059
	(0.283)	(0.283)	(0.283)	(0.296)	(0.296)	(0.296)	(0.045)	(0.045)	(0.045)
Neuroticism	0.354	0.354	0.354	-0.301	-0.301	-0.301	0.048	0.047	0.048
	(0.213)	(0.213)	(0.213)	(0.226)	(0.226)	(0.226)	(0.041)	(0.040)	(0.041)
Openness	-0.257	-0.257	-0.257	0.198	0.198	0.198	-0.036	-0.035	-0.036
	(0.237)	(0.237)	(0.237)	(0.252)	(0.252)	(0.252)	(0.039)	(0.039)	(0.039)
Conscientiousness	0.622**	0.622**	0.622**	-0.690**	-0.690**	-0.690**	0.098**	0.097**	0.098**
	(0.305)	(0.305)	(0.305)	(0.305)	(0.305)	(0.305)	(0.047)	(0.046)	(0.047)
Agreeableness	0.874***	0.874***	0.874***	-0.766**	-0.766**	-0.766**	0.121**	0.120**	0.121**
	(0.300)	(0.300)	(0.300)	(0.297)	(0.297)	(0.297)	(0.048)	(0.048)	(0.048)
Constant	1.725***	2.107***	1.698***	1.440***	1.036***	1.466***			
	(0.250)	(0.231)	(0.254)	(0.242)	(0.227)	(0.247)			
Wald test (p-value)									
H_0: P4P×MildSev=P4P×IntermSev		<0.001			<0.001			<0.001	
H_0: P4P×MildSev=P4P×HighSev		<0.001			<0.001			<0.001	
H_0: P4P×IntermSev=P4P×HighSev		<0.001			<0.001			0.015	
H_0: P4P×LowMHB=P4P×HighMHB			0.097			0.049			0.004
Observations	990	990	990	990	990	990	990	990	990
Subjects	55	55	55	55	55	55	55	55	55
(Pseudo) R^2	0.509	0.534	0.509	0.287	0.328	0.287	0.131	0.140	0.131

Notes: This table shows parameter estimates from OLS regressions (Panel A and B) and average marginal effects from fractional probit response regressions (Panel C). Robust standard errors clustered for subjects are shown in parentheses. P4P is a dummy variable indicating the introduction of P4P. IntermSev and HighSev are dummy variables for intermediate and high severities of illness. HighMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). All models control for individual characteristics which comprise gender, medical background (non-medical student, medical student, physician), and personality traits.
* $p < 0.10$, ** $p < 0.05$, and, *** $p < 0.01$.

Table C.8: Comparison of effects when introducing performance pay to fee-for-service and capitation with full list of covariates, effect splitted by marginal health benefit

	A. Absolute deviation from patient-optimal care ρ		**B. Proportional health benefit $\hat{H}$**	
Method:	OLS	OLS	Frac. Probit	Frac. Probit
Model:	(1)	(2)	(3)	(4)
CAP	-0.013	-0.018	0.001	0.001
	(0.309)	(0.283)	(0.037)	(0.034)
IntermSev	-0.033	-0.033	-0.063***	-0.063***
	(0.073)	(0.073)	(0.013)	(0.013)
HighSev	0.008	0.008	-0.001	-0.003
	(0.135)	(0.136)	(0.019)	(0.019)
HighMHB	-0.074	-0.074	0.007	0.008
	(0.086)	(0.086)	(0.010)	(0.010)
CAP×HighMHB	-0.120	-0.120	0.016	0.017
	(0.112)	(0.113)	(0.013)	(0.013)
CAP+P4P×LowMHB	-1.170***	-1.170***	0.151***	0.150***
	(0.187)	(0.187)	(0.019)	(0.018)
FFS+P4P×LowMHB	-1.212***	-1.212***	0.154***	0.154***
	(0.178)	(0.178)	(0.017)	(0.016)
CAP+P4P×HighMHB	-1.012***	-1.012***	0.133***	0.131***
	(0.172)	(0.172)	(0.016)	(0.015)
FFS+P4P×HighMHB	-1.173***	-1.173***	0.146***	0.144***
	(0.170)	(0.171)	(0.015)	(0.014)
Medical students		-0.136		0.016
		(0.215)		(0.035)
Physicians		-0.894***		0.145***
		(0.233)		(0.033)
Male		0.185		-0.030
		(0.189)		(0.031)
Extraversion		0.008		-0.004
		(0.213)		(0.032)
Neuroticism		-0.051		0.006
		(0.195)		(0.033)
Openness		0.078		-0.016
		(0.198)		(0.030)
Conscientiousness		-0.199		0.029
		(0.253)		(0.037)
Agreeableness		-0.470**		0.079**
		(0.236)		(0.037)
Constant	1.858***	2.048***		
	(0.245)	(0.274)		
Individual controls	No	Yes	No	Yes
Wald tests (p-value):				
H_0: CAP+P4P×LowMBH = FFS+P4P×LowMHB	0.507	0.508	0.425	0.404
H_0: CAP+P4P×HighMHB = FFS+P4P×HighMHB	0.871	0.872	0.863	0.856
Observations	1926	1926	1926	1926
Subjects	107	107	107	107
(Pseudo) R^2	0.135	0.207	0.064	0.099

Notes: This table shows parameter estimates from OLS regressions (Panel A and B) and average marginal effects from fractional probit response regressions (Panel C). Robust standard errors clustered for subjects are shown in parentheses. P4P is a dummy variable indicating the introduction of P4P. IntermSev and HighSev are dummy variables for intermediate and high severities of illness. HighMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). All models control for individual characteristics which comprise gender, medical background (non-medical student, medical student, physician), and personality trait. * $p < 0.10$, ** $p < 0.05$, and, *** $p < 0.01$.

Table C.9: Comparison of effects when introducing performance pay to fee-for-service and capitation, with full list of covariates

	A. Absolute deviation from patient-optimal care ρ		**B. Proportional health benefit $\hat{H}$**	
Method:	OLS	OLS	Frac. Probit	Frac. Probit
Model:	(1)	(2)	(3)	(4)
CAP	-1.828***	-1.832***	0.210***	0.209***
	(0.342)	(0.310)	(0.037)	(0.032)
IntermSev	-0.936***	-0.936***	0.020*	0.020
	(0.147)	(0.148)	(0.012)	(0.012)
HighSev	-1.782***	-1.782***	0.182***	0.178***
	(0.253)	(0.254)	(0.021)	(0.020)
HighMHB	-0.086**	-0.086**	0.013**	0.013**
	(0.033)	(0.034)	(0.005)	(0.005)
CAP×IntermSev	1.784***	1.784***	-0.243***	-0.241***
	(0.201)	(0.201)	(0.029)	(0.028)
CAP×HighSev	3.540***	3.540***	-0.492***	-0.484***
	(0.354)	(0.355)	(0.033)	(0.033)
CAP+P4P×MildSev	-0.309***	-0.309***	0.056***	0.054***
	(0.107)	(0.107)	(0.017)	(0.016)
FFS+P4P×MildSev	-1.994***	-1.994***	0.171***	0.170***
	(0.274)	(0.275)	(0.017)	(0.016)
CAP+P4P×IntermSev	-1.133***	-1.133***	0.148***	0.148***
	(0.187)	(0.188)	(0.018)	(0.017)
FFS+P4P×IntermSev	-1.179***	-1.179***	0.151***	0.150***
	(0.181)	(0.182)	(0.017)	(0.016)
CAP+P4P×HighSev	-1.909***	-1.909***	0.168***	0.167***
	(0.289)	(0.290)	(0.018)	(0.017)
FFS+P4P×HighSev	-0.423***	-0.423***	0.075***	0.077***
	(0.110)	(0.111)	(0.015)	(0.015)
Medical students		-0.136		0.015
		(0.215)		(0.035)
Physicians		-0.894***		0.142***
		(0.233)		(0.033)
Male		0.185		-0.031
		(0.189)		(0.030)
Extraversion		0.008		-0.004
		(0.214)		(0.032)
Neuroticism		-0.051		0.005
		(0.195)		(0.033)
Openness		0.078		-0.016
		(0.198)		(0.030)
Conscientiousness		-0.199		0.029
		(0.253)		(0.037)
Agreeableness		-0.470**		0.078**
		(0.236)		(0.037)
Constant	2.759***	2.950***		
	(0.316)	(0.324)		
Individual controls	No	Yes	No	Yes
Wald tests (p-value):				
H_0: CAP+P4P×MildSev = FFS+P4P×MildSev	<0.001	<0.001	<0.001	<0.001
H_0: CAP+P4P×IntermSev = FFS+P4P×IntermSev	0.860	0.860	0.872	0.884
H_0: CAP+P4P×HighSev = FFS+P4P×HighSev	<0.001	<0.001	<0.001	<0.001
Observations	1926	1926	1926	1926
Subjects	107	107	107	107
(Pseudo) R^2	0.240	0.312	0.094	0.129

Notes: For Panel A OLS estimates are reported with robust standard errors clustered for subjects (in brackets). For Panel B average marginal effects (AMEs) based on a fractional probit response are reported with robust standard errors clustered for subjects (in brackets). CAP = 1 if physicians are remunerated by CAP, and = 0 otherwise (by FFS). P4P is a dummy variable indicating the introduction of P4P. IntermSev and HighSev are dummy variables for intermediate and high severities of illness. HighMHB is a dummy for the marginal health benefit being 1 if $\theta = 2$ (high), and 0 otherwise ($\theta = 1$, low). Controls for subjects' individual characteristics comprise gender, medical background (non-medical student, medical student, physician), and personality traits. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.