# Physician performance pay: Experimental evidence

*Jeanette Brosig-Koch*
University of Duisburg-Essen and
Health Economics Research Centre
Essen (CINCH)

*Heike Hennig-Schmidt*
Department of Economics,
University of Bonn and
Department of Health Management
and Health Economics,
University of Oslo

*Nadja Kairies-Schwarz*
University of Duisburg-Essen and
Health Economics Research Centre
Essen (CINCH)

*Johanna Kokot*
University of Hamburg and
Hamburg Centre for Health
Economics

*Daniel Wiesen*
University of Cologne
Department of Health Care
Management

**UNIVERSITY
OF OSLO**
HEALTH ECONOMICS
RESEARCH NETWORK

Working paper 2020: 3

Health Economics Research Network at the University of Oslo
ISSN 1501-9071 (print version.), ISSN 1890-1735 (online), ISBN 978-82-7756-268-1

# Physician performance pay: Experimental evidence*

Jeannette Brosig-Koch, Heike Hennig-Schmidt, Nadja Kairies-Schwarz,

Johanna Kokot, Daniel Wiesen

May 7, 2020

*Brosig-Koch: University of Duisburg-Essen and Health Economics Research Center (CINCH) Essen, Berliner Platz 6-8, 45127 Essen, Germany, e-mail: jeannette.brosig@uni-due.de; Kairies-Schwarz: University of Duisburg-Essen and CINCH Essen, Berliner Platz 6-8, 45127 Essen, Germany, e-mail: nadja.kairies@ibes.uni-due.de; Hennig-Schmidt: University of Bonn, Department of Economics, Adenauerallee 24-42, 53113 Bonn, Germany, University of Oslo, Department of Health Management and Health Economics, Institute of Health and Society, University of Oslo, Norway, e-mail: hschmidt@uni-bonn.de (corresponding author); Kokot: University of Hamburg and Hamburg Center for Health Economics, Esplanade 36, 20354 Hamburg, Germany, e-mail: johanna.kokot@uni-hamburg.de; Wiesen: University of Cologne, Department of Health Care Management, Albertus-Magnus-Platz, 50923 Cologne, Germany, e-mail: wiesen@wiso.uni-koeln.de. We are highly indebted to the support of the Zi–*Zentralinstitut für die kassenärztliche Versorgung* of the Federal Republic of Germany. We are grateful for valuable comments and suggestions from Nick Bloom, Amitabh Chandra, John Cawley, Jim Cox, Robert Dur, Randy Ellis, Keith Ericsson, Glenn Harrison, Tor Iversen, Rudi Kerschbamer, Lawrence Katz, Albert Ma, Henry Mak, Robert Nuscheler, Anne Sophie Oxholm, Daniele Paserman, Jim Rebitzer, Meredith Rosenthal, Luigi Siciliani, Tony Scott, John F. Sweeney, John Van Reenen, and by participants of seminars and conferences at the AEA Philadelphia, BEH Workshop Atlanta, Boston University (Micro), City University London, Cornell University, Empirical Management Conference, Erasmus University Rotterdam, Helmholtz Institute Munich, Hertie School Berlin, Hochschule St. Gallen, iHEA Boston, Imperial College London, IUPUI Indianapolis, University of Cologne, University of Hannover, University of Innsbruck, University of Lucerne, University of Wuppertal, University of York, and *Verein für Socialpolitik Münster*. We also thank Kerstin Dannenberg, Anna Hanel, and Mona Gross for their excellent research assistance. Financial support by the German Research Foundation (grant: BR 2346/2-1/2), the German Federal Ministry of Education and Research (grant: 01EH1602A), and the University of Cologne is gratefully acknowledged. Hennig-Schmidt and Wiesen also gratefully acknowledge financial support from the Research Council of Norway (grant: 231776 IRECOHEX) during their employment at the Department of Health Management and Health Economics of the Institute of Health and Society, University of Oslo, Norway. Wiesen is also grateful for the hospitality and support from the Economics Department at Boston University where substantial parts of the paper were written.

## Abstract

We analyze the causal effect of performance pay on physicians' medical service provision and the quality of care. To address this effect, which is difficult to study in the field we conducted an online experiment with primary care physicians randomly drawn from a representative resident physician sample in Germany. Linking individual physicians' behavioral data with administrative data enables us to identify how practice characteristics account for the heterogeneity in individual physicians' responses to performance incentives, which field data do not allow in general. We find that performance pay reduces underprovision of medical care compared to lump-sum capitation. The effect increases with patients' severities of illness. Already small incentives are effective in enhancing the quality of care. Our results further indicate that physicians in high-profit practices and practicing in cities are most responsive to incentives.

**Keywords:** pay for performance, behavioral experiment, practice characteristics

**JEL-Classification:** I11, C93

A fundamental question in health policy around the world is that of how to incentivize health care providers to improve the quality of care. While the traditional approaches to pay physicians have focused on fee-for-service and capitation, there has been growing interest in directly measuring and incentivizing physicians' performance based on patients' health outcomes. In particular, to align better physician incentives with quality objectives, performance pay has become increasingly popular in health care.[1] This approach draws on the logic of performance pay in human resource management, which rewards workers for achieving pre-specified performance targets (e.g., Baker, 1992; Prendergast, 1999; Lazear,

[1] Performance pay is typically granted conditional on achieving a performance threshold. The idea of paying physicians (at least partially) on the basis of direct performance measures has attracted particular attention, as fee-for-service incentivizes physicians to overserve and capitation to underserve patients. Performance pay for physicians has been widely introduced, for example in the UK (see, e.g., Roland, 2004; Doran et al., 2006; Campbell et al., 2009; Roland and Campbell, 2014; Kristensen et al., 2014) and the US (e.g., Rosenthal et al., 2005; 2006).

2000).

While the idea of paying physicians for performance has made its way into health policy, the empirical evidence regarding its effect on the quality of care is quite limited—with identification of the causal impact of physician incentives being the main challenge. Establishing a causal link is particularly difficult due to the likely endogeneity of institutions (e.g., Baicker and Goldman, 2011), biases because of incomplete performance measures or measurement errors (e.g., Campbell et al., 2009), gaming of performance indicators (e.g., Gravelle et al., 2010; Maynard, 2012), and the frequent introduction of performance pay accompanied by other interventions (e.g., Lindenauer et al., 2007). It therefore comes as no surprise that the empirical evidence is quite mixed on whether performance pay helps to improve the quality of care.[2] If anything, rather moderate effects of performance pay are reported (e.g., Mullen et al., 2010; Li et al., 2014). As a consequence, one might argue that performance pay may be ill-suited for health care provision altogether (e.g., Frakt and Jha, 2018).

In this study, we employ a controlled behavioral experiment with primary care physicians drawn from a representative sample of German resident physicians to identify the effect of performance pay and to complement existing empirical research. Potential reasons that might lead to mixed evidence in the field are the following: First, field studies based on non-experimental data, typically consider aggregate effects of physician performance pay, while individuals' responses might be heterogeneous based on their individual and on practice characteristics. Health policies are usually introduced at the state or national level, not considering heterogeneity in these characteristics. Estimation results might thus be biased by, for example, physicians' personality traits and their practice characteristics such as location or profitability (e.g., Li et al., 2014; Donato et al., 2017). Second, it is not well understood how the design of a performance pay system (e.g., size of a bonus) the affects the provision and the quality of health

[2] For meta-studies on the effectiveness of pay for performance initiatives in OECD countries, see Scott et al. (2011), Eijkenaar et al. (2013), and Mendelson et al. (2017). Evidence from developing countries is also somewhat mixed Miller and Babiarz (2014).

care (Epstein, 2012; Roland, 2012; Kristensen et al., 2016). This lack of understanding is the more surprising, as behavioral evidence indicates that the size of incentives affects the behavior of individuals (e.g., Gneezy and Rustichini, 2000; Ariely et al., 2009). Finally, additional performance incentives may lead to a crowding out of patient-regarding behavior (e.g., Siciliani, 2009; Maynard, 2012). Compared to private-sector employees, public-sector physicians may be more intrinsically or prosocially motivated towards their patients (Arrow, 1963; Francois, 2000; Besley and Ghatak, 2005; Delfgaauw and Dur, 2008; Kolstad, 2013) and performance pay may dampen the effects of intrinsic and other-regarding motivation (e.g., Deci and Ryan, 2010; Kreps, 1997; Bénabou and Tirole, 2003; 2006; Gneezy et al., 2011).[3] While some experimental evidence for motivation crowding-out exists, for instance in real work settings (e.g., Gneezy and Rustichini, 2000; Ariely et al., 2009; Huffman and Bognanno, 2018) and in the case of blood donations (Mellström and Johannesson, 2008), evidence is lacking on whether performance pay affects physicians' altruistic (patient-regarding) behavior and therefore the quality of care.

Addressing these issues requires an exogenous variation of payment systems and the observability of individual responses. Running a controlled online-experiment with physicians[4] in a highly controlled decision-environment meets these requirements. A large-scale field experiment or RCT, which is also suitable, might be prohibitively costly and might adversely affect the health status of certain patient groups due to unintended effects of incentives. We include decision-makers in our controlled online experiment relevant to address our research question: namely primary care physicians. We recruited physicians via the pool of participants from the 'Physician Practice Panel' (Zi-Praxis-Panel,

[3] More generally, it is argued in the economics and psychology literature that economic incentives, being targeted at people who are intrinsically motivated, have been shown to be less effective than anticipated for purely profit-oriented individuals; see Bowles and Polania-Reyes (2012) for an excellent overview. For physicians, performance pay that yields financial incentives for good quality of care may thus crowd out their altruistic (patient-regarding) motivation of treating the patient optimally.

[4] Our behavioral experiment in health can be regarded as an artefactual field experiment according to the taxonomy of experiments by Harrison and List (2004). For a definition of behavioral experiments in health, see Galizzi and Wiesen (2018).

ZiPP) of the Zi – *Zentralinstitut für die kassenärztliche Versorgung* (The National Association of Statutory Health Insurance Physicians) in Germany. This is a representative sample of all resident physicians in Germany and is run annually with a sample size of about 5,000 physicians across all specializations. By combining individual-physician experimental data with real physicians' administrative data on their practice characteristics and individual self-reported characteristics we are able to account for the heterogeneity in these characteristics.

In order to avoid the complexity prevalent in the field, we implemented an abstract decision task, which ensures a high level of control while it still captures the main features and incentives inherent in physicians' health care provision. Physicians decide on the quantity of health care services for a set of different stylized patients varying in their severities of illness. Each decision simultaneously determines a physician's profit and the patient's health benefit. Reducing the complexity, which prevails in the field, in our experimental design ensures that individual physicians' responses are not confounded by different subjective interpretations of patients' health and heterogeneity in individuals' experience and ability. Prior to the experiment, we conducted several interviews with experts and physicians to ensure that the stylized decision situation in the experiment still captures the main features and incentives inherent in physicians' health care provision. In light of the feedback received from the participants in a post-experimental questionnaire, we are confident that physicians not only were fully aware of the trade-offs between patients' health benefits and their own financial concerns. They also pointed to the similarity between the experiment and their daily practice. In sum, we ensure internal validity and at the same time maintain a high degree of external validity by using the relevant subject pool of primary care physicians and relating their individual behavior in the experiment to detailed administrative data at a primary care physicians' practice level.

Our experimental design is well-grounded in economic theory, an approach which has been prominently advocated by economists (e.g., Heckman, 2010; List, 2011). In a parsimonious decision situation, physicians decide on the provision of medical services for a set of passive, abstract patients; quantity choices on a one-dimensional scale determine their profit and the patients' health benefits. The incentive to care for a patient is made salient, as real patients' health outside the experiment is affected by the subjects' decisions. Physicians were informed that the total health benefits (measured in monetary terms in the experiment) are transferred to the *Christoffel Blindenmission*, a charity coping with eye diseases. The money is earmarked to finance surgical treatment of cataract patients. Using a relatively cheap, but necessary medical treatment, we come close to a linear relationship between the patient health benefit provided in the experiment and the number of real patients who benefit from cataract surgery. For each patient in the experiment, trade-offs between the patient-optimal and the profit-maximizing quantities of care exist. With performance pay, the incentives of patients and physicians become more aligned, albeit not perfectly. The quality of care is non-perfectly contractible, as we assume asymmetric information between physician and payer about the optimal quality. Physicians are commonly assumed to be better informed than their patients (e.g., Dulleck and Kerschbamer, 2006; Dulleck et al., 2011), allowing for moral hazard (e.g., Gaynor and Gertler, 1995; Gaynor et al., 2004). These design features allow us to analyze whether a crowding-out of patient-regarding behavior results from the introduction of performance pay.

Performance pay is introduced at the within-subject level. Physicians are first incentivized by a lump-sum capitation, then performance pay is added in form of a discrete bonus complementing the baseline capitation. A bonus is granted whenever a physician meets a quality threshold linked to the patient's health benefit. Quality thresholds are set for different severities of illness and bonus rates are adjusted for the severities.[5] To test for the behavioral effect of the level

[5] The adjustment of the bonus rates based on illness severities can be interpreted as some form of risk adjustment (e.g., Glazer and McGuire, 2000). Patients with a high severity of illness, for example, face the highest 'risk' of being undertreated under capitation, a behavioral pattern that has been indicated by recent experimental findings (Hennig-Schmidt et al., 2011; Hennig-Schmidt and Wiesen, 2014; Kesternich et al., 2015; Brosig-Koch et al., 2016b, 2016a, 2017). Similarly, Clemens and Gottlieb (2014) report that the severity levels of the patients' illnesses relate to the physicians' responses to fee-for-service incentives.

of incentives, we implement two different bonus levels: a 5% bonus and 20% bonus. We randomly assign physicians to one of the two payment conditions.

We also link individual physicians' behavior to administrative data about their practice characteristics to address potential heterogeneity in individuals' health service provision and the quality of care in the experiment. We thus explore how experimental behavior relates to physicians' real-world characteristics and contribute to the generalizability of experimental results (Levitt and List, 2007, 2009; Czibor et al., 2019). The practice characteristics we consider comprise annual practice profit, location (city, outer conurbation, rural area), patient-related characteristics (share of patients in the statutory health insurance (SHI) scheme, revenue share, and time spent on SHI patients),[6] and whether more than one physician is employed in the practice.

These individual physician characteristics are important from a theoretical and empirical perspective concerning the quality of care. (i) Physicians' financial (profit) orientation is typically described as one key driver of physicians' health care provision in the economics and medical literature (Arrow, 1963; Pellegrino, 1987). However, empirical evidence on the relationship between physicians' practice profits and the quality of care is scarce. Estimates from experimental data indicate that medical students exhibit a considerable profit orientation within the confines of the experimental setup (Godager and Wiesen, 2013; Li, 2018). (ii) The location of physicians' practices might also relate to heterogeneity in the quality of care. Studies which compare the quality of health care between rural and urban areas usually report the former to be lower than the latter (e.g., Campbell et al., 2001; Burke et al., 2010; Kralewski et al., 2015)—often due to limited access to health care in rural areas. (iii) Patients' characteristics, such as their insurance status, have been shown to affect health care utilization (demand side) in the seminal RAND and the Oregon health insurance experiments (Manning et al., 1987; Newhouse and the Insurance Experiment Group, 1993;

[6] SHI patients are the ones under the statutory health insurance scheme. The alternative is to insure privately. Services rendered to these patients are typically reimbursed on a fee-for-service basis as oppossed to lump-sum incentives for SHI patients. For more details on the German physician remuneration, see Sections A.1 and A.2 in Appendix A.

Finkelstein et al., 2012; Baicker and Finkelstein, 2011; Baicker et al., 2013). We complement this seminal stream of the literature, in that we consider how physicians' behavior (supply side) relates to heir patients' characteristics such as their insurance status. (iv) Finally, evidence on how practice size (number of physicians employed in a practice), relates to health care quality is inconclusive (e.g., Campbell et al., 2001; Ng and Ng, 2013; Kralewski et al., 2015; Casalino et al., 2018). Linking behavioral data to physicians' practice characteristics enables us to shed light on potential drivers of heterogeneity in the physicians' behavioral responses to incentives in performance pay.

Our study yields three main results. First, physician performance pay affects health care service provision and enhances the quality of care. To quantify matters, the quality increases by about 7% on the aggregate compared to capitation. The performance-pay effect on quality increases with the patients' severity of illness. Second, we find that small incentives (a quarter of the size of the larger bonus) were effective in enhancing the quality of care. Implementing a performance-pay scheme that yields an incentive for physicians to earn 5% in addition to a baseline payment motivates a similar behavioral change compared to paying a 20% bonus. We also observe crowding-out of patient-regarding behavior, albeit to a rather small extent (for 7% of all patients). This finding suggests that crowding-out alone is not sufficient to explain the mixed effects of performance pay in the literature. Third, we find that physicians' practice characteristics significantly relate to physicians' health care choices and the quality of care in the experiment. Physicians in high-profit practices are also more profit-oriented in the experiment, resulting in lower qualities of care compared to low-profit practice physicians. Also, physicians practising in rural areas provide a significantly higher quality of care compared to physicians in cities. The quality of care is significantly higher among physicians from low profit practices, when practicing in rural areas, and it increases in the time spent on SHI patients. Other patient characteristics related to their insurance status do not significantly affect the behavior of physicians.

The rest of the paper is organized as follows. Section 1 provides a brief description of our physician sample and details the experimental design and procedure. In Section 2, we introduce a simple model to derive behavioral hypotheses for the experiment. Section 3 presents behavioral results on the effects of physician performance pay on health care service provision. Section 4 identifies relationships between the physicians' behavior in the experiment and their practice characteristics. Section 5 discusses implications and generalizability of behavioral results. Finally, Section 6 summarizes and concludes.

# 1 Experiment and sample

## 1.1 Our primary care physician sample

In our study, we use a representative sample of German primary care physicians contracting with Statutory Health Insurance (SHI). More details on the German primary care setting, the institutional background of the German SHI system, and the payment system for primary care physicians contracting with the SHI are relegated to Sections A.1 and A.2 in Appendix A. The 'Physician Practice Panel' (Zi-Praxis-Panel, ZiPP) of the Zi – *Zentralinstitut für die kassenärztliche Versorgung* (The National Association of Statutory Health Insurance Physicians) is a representative sample of all resident physicians in Germany and is run annually with a sample size of about 5,000 physicians across all specializations. It comprises about 5% of all practices in Germany. ZiPP is a unique data base, designed to analyze the cost structure, turnover, and surplus of SHI physician practices, to inform the annual negotiations on the budget for physicians' negotiations between sickness funds and the associations of SHI-physicians (*Kassenärztliche Vereinigung*, KV).

In 2016, primary care physicians comprised 32% (54,900) of all resident self-employed physicians contracting with the SHI. They were organized in 39,000 practices (77% in individual and 23% in group practices, see KBV 2016)[7]. This is the statistical population from which the subsample was randomly drawn (with a 9% selection probability); see ZiPP (2017). Compared to the KV's re-

[7] See *gesundheitsdaten.kbv.de/cms/html/17020.php* for the above data.

imbursement claims data, the ZiPP sample provides a good approximation of the general population of resident primary care physicians in Germany when measured by the number of medical treatments per physician, the remuneration per physician, the remuneration per medical treatment, and the ratio between remuneration and medical treatments required (ZiPP 2017, p. 19). The representative sample of resident primary care physicians is stratified according to three regional areas (city, outer conurbation, and rural).

Our study design was approved by the Scientific Board of the Zi Praxis Panel, which consists of independent scientists from medicine, health sciences, and economics. The research plan contained an experimental design which was analogous to the laboratory pre-study of Brosig-Koch et al. (2013).

Our experiment was run in April 2016 and was therefore based on those physicians who participated in the ZiPP survey wave from September to December 2015. For our experiment, the Zi randomly selected a subsample of 662 primary care physicians from the ZiPP who were invited to take part in our online experiment. In total, 104 resident primary care physicians participated in our experiment in our study. The number of participants was guided by our power and sample size calculations (see next subsection below). This is about 10% of all PCPs enrolled in the ZiPP. The ZiPP sample is also a rather good approximation of the general population of resident primary care physicians. Detailed sample characteristics are provided in Section 1.5.

## 1.2 General design

Our experimental study consists of two main experimental conditions and four control conditions. In the two main conditions, primary care physicians participating in our online experiment are randomly assigned either to the Low-bonus or the High-bonus condition ($N$=104). In the control conditions, we check for the robustness of our results and involve a medical students sample participating in online experiments ($N$=127). The general design and the decision situation are the same for all conditions.

We employ a medical frame in our experiment. While abstracting from the complexity of daily medical practice, the decision situation captures the main features and incentives primary care physicians face in their daily practice. This view has been supported in interviews with practicing physicians and leading experts involved in physician reimbursement at regional KVs.[8]

All subjects, be they primary care physicians or medical students in the role of physicians, decide on the provision of health care services for several different stylized 'patients'. Henceforth, we use these labels to indicate the roles in our experiment. In each experimental condition, physicians are exposed to two consecutive payment conditions. In the first part, each physician receives a lump-sum capitation (CAP) for providing health care services. In the second part, we introduce physician performance pay at a *within-subject level* (CAP+P4P).[9]

To determine the *a-priori* sample size needed to test for the effect of performance pay (within-subjects), our calculations showed that at least 39 physicians per treatment were necessary to detect a normalized effect of $dz = 0.6$. To determine this effect between CAP and CAP+P4P, we conducted a pilot experiment with medical students and non-medical students in the decision situation of the present experiment and used the parameters from the High–bonus (20%) condition; see Brosig-Koch et al. (2016b). For our analysis, we considered the means and standard deviations from treatment CAP and CAP+P4P with 45 participants; see Table 2 in Brosig-Koch et al. (2016b). Between the two conditions, we considered changes from CAP to CAP+P4P, using two-sided Wilcoxon signed rank tests, and assumed a power of 80% and a 5% significance level.

We use a threshold-based performance-pay system designed to mitigate the incentive to underserve patients in CAP. To this end, each physician is paid a discrete bonus in addition to the CAP payment if a quality threshold is reached

[8] This view is also supported by questionnaire data from our study. We asked our participating physicians about the motives for their decisions in a post-experimental open question. 98 of the 104 doctors were motivated by the patient benefit only or by both the benefit and their own profit. None of them commented that our design would be too artificial or simplistic. Only two participants referred to the experimental decision situation as somewhat theoretical and to be only vaguely reflective of their daily experiences in their practices, while admitting the realistic nature of the inherent incentives and tradeoffs in the decision situation.

[9] Note that the only one exception is condition C–High–bonus (20%)–First, which we deliberately introduced to test for order effects; see the notes of Table 1.

that is tied to the patients' optimal health outcome.[10] This feature of our experimental design is motivated by the main purpose of physician performance pay, namely to improve the quality of health care delivery (e.g., Rosenthal et al., 2006). To realize this goal, a payment is granted if a quality threshold is reached, which is often tied to health outcome measures (e.g., Gravelle et al., 2010). Moreover, we vary the size of the bonus payment at a *between-subject level* by assigning physicians either to a Low-bonus or a High-bonus condition, in which they receive a discrete bonus of either 5% or 20% in addition to the capitation payment; see Table 1.

Table 1: Experimental conditions

| | Part of the experiment | | Number of |
|---|---|---|---|
| | First part | Second part | subjects |
| **A. Main conditions: Primary care physicians** | | | |
| Low–bonus (5%) | CAP | CAP+P4P-5% | 51 |
| High–bonus (20%) | CAP | CAP+P4P-20% | 53 |
| **B. Control conditions: Medical students** | | | |
| C–Low–bonus (5%) | CAP | CAP+P4P-5% | 30 |
| C–High–bonus (20%) | CAP | CAP+P4P-20% | 33 |
| C–High–bonus (20%)–First | CAP+P4P-20% | CAP | 27 |
| C–CAP–High | CAP+20% | CAP+P4P-20% | 37 |
| *Total* | | | 231 |

*Notes.* This table provides an overview of our experimental conditions: the main conditions with primary care physicians and the control conditions with medical students who participated in online experiments. In all experimental conditions, we varied the payment system in the two parts of the experiment. This allows us to analyze the effect of performance pay on a within-subject level. We analyze the effect of bonus size on a between-subject level, comparing behavior between Low–bonus (5%) and High–bonus (20%) in the respective second parts of the experiment. In the control conditions, we conducted the online experiments with medical students.To have adequate financial incentives to reflect typical hourly wage levels for physicians and students, values for students were multiplied by 0.32. In C–Low–bonus (5%) and C–Low–bonus (20%), students decided in the same situation as primary care physicians in the main conditions. In additional control conditions, we test for the robustness of our results. In C–CAP-High, we check for income effects when introducing performance pay. To this end, we raise the capitation payment in the first part of the experiment by 20% (labeled CAP+20%). To test for order effects, in condition C–High–bonus (20%)–First, medical students were exposed to performance pay in the first part of the experiment followed by CAP in the second part.

Finally, we add control experiments to check for the robustness of our results towards (i) order effects, (ii) income effects, and (iii) subject pool effects; see Appendix A.4.

[10] Performance thresholds are quite common in practice; for example, in the Quality and Outcomes Framework in the UK (e.g., Roland, 2004; Roland and Campbell, 2014) and in many Health Maintenance Organizations' (HMO) P4P systems (e.g., Rosenthal et al., 2006).

### 1.3 Decision situation

A physician decides on the quantity of medical services $q \in [0, 10]$ for nine different patients ($j = 1, \ldots, 9$) in both payment systems. Patients differ in illnesses $k \in \{A, B, C\}$ and in severities of illness $l \in \{x, y, z\}$. A specific patient is a combination of one of the three illnesses and one of the three severities. We assume patients to be fully insured.[11] A patient's illness and severity of illness are the same in all payment schemes and conditions. This design feature implies that behavioral changes between payment schemes and conditions are not confounded by variations in the patient population.

With each decision, a physician determines his or her own profit and a patient's health benefit. While all physicians decide for *abstract* patients in the experiment, *real* patients' health is affected by their choices. Physicians are informed that the monetary equivalent of the patient health benefit resulting from their decisions is transferred to a charity that uses the money exclusively for surgical treatments of cataract patients; see Subsection 1.4 for procedural details. This mechanism ensures that the patients' health benefit is made salient.[12] For an illustration of the decision situation, see the screenshots in Figure A.3 and the instructions in Appendix A.5.

A physician's remuneration is $R(q) = \Lambda + b_l I_{b_l}$, with $\Lambda$ being the capitation payment in the experiment; $b_l$ is the bonus payment, which depends on the patient's severity of illness $l$ (for the bonus rates, see below); $I_{b_l}$ denotes an indicator variable, which equals 1 if the physician's chosen quantity meets the quality threshold. This is the case if a quantity choice does not differ by more than one unit from the patient optimal treatment $q^*$, that is, if $|q - q^*| \leq 1$;

[11] This seems to be a natural assumption in our setting, as patients insured under German SHI do not make co-payments in ambulatory health care for services provided by their doctors and covered by the standard SHI benefits package. Thus, the primary care physicians in our experiment are fairly familiar with fully insured patients. Moreover, full insurance is commonly assumed in theoretical models of physicians' behavior in the health economics literature; see McGuire (2000) for an comprehensive overview.

[12] This mechanism has been used in various recent behavioral experiments in health; see, for example, Hennig-Schmidt et al. (2011), Hennig-Schmidt and Wiesen (2014), Kesternich et al. (2015), and Brosig-Koch et al. (2016a, 2017).

$I_{b_l} = 0$ otherwise. In CAP, $b_l^{\mathrm{CAP}} = 0$. A physician's profit is given by

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

with $\Lambda$, $b_l > 0$, $c'(q) > 0$ and $c''(q) > 0$. We set $c(q) = q^2/4$ for both payment systems in the experiment.[13] For an illustration of the physicians' profits, see Figure A.1.

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

$$H(q) = H_k - \theta_k |q - q_l^*| \tag{2}$$

with $k \in \{A, B, C\}$ and $l \in \{x, y, z\}$. In particular, $H_A = 7$, $H_B = 10$, and $H_C = 14$, $\theta_A = \theta_B = 1$ and $\theta_C = 2$. The patient-optimal quantity $q^*$ varies with the severities of illness $l$. For mild ($x$), intermediate ($y$), and high ($z$) severity of illnesses, the patient-optimal quantities are $q_x^* = 3$, $q_y^* = 5$, and $q_z^* = 7$, respectively; for an illustration, see Figure A.2 in Appendix A.[14] We are able, first, to analyze the deviation from patient-optimal health care service provision (e.g., underprovision) and, second, to introduce a 'clean' outcome-based performance measure tied to a measurable health outcome $H(q^*)$. We thus avoid measurement errors of health care quality, often assumed to confound effects of performance pay in empirical studies. All parameters of the experiment, remuneration, cost, profit, and patient health benefit corresponding to $q$ are common knowledge to the physicians. All monetary values like remuneration, cost, profit, benefit, and patient health benefit are indicated in Euro.

[13] The assumption of convex costs it often made in the theoretical health economic literature; see McGuire (2000) for a summary.

[14] Varying the patients' characteristics is motivated by recent empirical findings indicating that the effect of financial incentives on physicians' behavior differs with patients' characteristics (e.g., Clemens and Gottlieb, 2014) and by experimental evidence (e.g., Hennig-Schmidt et al., 2011; Brosig-Koch et al., 2017).

Taking a theoretical perspective, our performance-pay system captures the asymmetric information between physician and payer (e.g., Ma and McGuire, 1997) with regard to the patient-optimal quantity of medical services. While the physician does observe $q^*$, our performance threshold implies that the payer only observes $q^* + \epsilon$, with a noise $\epsilon \in \{-1, 1\}$. Therefore, $q^*$ is not fully contractible in our performance-pay system.

We set bonus rates such that incentives are comparable across severities of illness. The bonus rates are adjusted for patients' severities of illness[15] and are as follows: In Low–bonus (5%), the bonus is $b_x = 2.25$, $b_y = 5.25$, and $b_z = 10.25$ for the patients with mild, intermediate, and high severity of illness, respectively. In High–bonus (20%), the bonus amounts to $b_x = 6$, $b_y = 9$, and $b_z = 14$ for the patients with mild ($x$), intermediate ($y$), and high ($z$) severity of illness, respectively. For the full set of parameter values, see Table A.1 in Appendix A.3.

We now qualify the trade-offs a physician faces. In CAP, the maximum profit $\pi(\hat{q})$ for a physician is €25. Choosing $q^*$ pays the physician €22.75, 18.75, and 12.75 for patients with a mild, intermediate, and high severity of illness, respectively. This means a reduction in profit by 9% (25%, 49%) compared to the maximum profit.

Under CAP+P4P, the trade-off between profit maximization and patient health benefit optimization is reduced. $\pi(\hat{q})$ in CAP+P4P-5% is €26.25, while choosing $q^*$ yields €25.00 (24.00, 23.00) for $x$ ($y, z$, respectively). $\pi(q^*)$ compared to $\pi(\hat{q})$ is reduced by about 5% (9%, 12%). In comparison to CAP, profit reductions are cut by 4.24 (16.43, 37.62) percentage points. In CAP+P4P-20%, the maximum profit is €30. Choosing $q^*$, however, yields profits of €28.75 (27.75, 26.75) for $x$ ($y, z$). The decrease in profits is about 4% (8%, 11%), which means that, compared to CAP, profit reductions are lowered by 4.88 (17.50, 37.17) percentage points.

[15] Notice that the adjustment of the bonus rates based on the severity of illness can be interpreted as a kind of risk adjustment (for a definition, see, for example, Glazer and McGuire, 2000; van de Ven and Ellis, 2000).

## 1.4 Protocol

In the main conditions, we employed a double-blind procedure according to the data protection guidelines of ZiPP that all ZiPP studies have to follow and about which participants were informed. Invitations to primary care physicians, including log-in data and IDs were sent out via a trustee at Zi. All decisions in the online experiment were made using these IDs; we can therefore only relate the primary physicians' choices to these IDs. The payment to participants was made via a notary authorized by Zi, who received a list containing the participants' names and IDs from the trustee and a list of IDs and payoffs from the IT department of Zi. The notary transferred the money to the banking accounts of the participants without being informed about their decisions.

The main conditions of the online experiment were programmed using the software SoPHIE (www.sophielabs.com), and were conducted in April 2016. The experimental procedure was as follows: Primary care physicians logged in with their IDs and were alternately assigned to one of the two conditions: Low–bonus (5%) or High–bonus (20%); i.e., the primary care physician who logged in first was assigned to Low–bonus (5%), the second one to High–bonus (20%), the third one again to Low–bonus (5%), and so forth. This procedure ensured that we had a random assignment of physicians to the two conditions.[16] Physicians then received onscreen instructions for the first part of the experiment. Moreover, a link to the instructions was provided on every subsequent screen during the experiment. Primary care physicians were informed that the experiment consisted of two parts, but received detailed instructions for the second part only after having finished the first part of the experiment. To check for each primary care physician's understanding of the decision task, he or she had to answer a set of control questions. The experiment did not start unless the primary care physician had answered all control questions correctly (instructions and control questions are in Sections A.5 and A.6 of Appendix A).

[16] We stopped this procedure at 110 physicians. However, we ended up with 104 participants (53 in the High–bonus and 51 in the Low–bonus condition) as six physicians did not complete the experiment and were dropped from the sample.

In each part of the experiment, primary care physicians subsequently decided on the quantity of medical services for each of the nine patients. The order of patients was randomly determined and kept constant for each participant in all conditions: $Bx$; $Cx$; $Az$; $By$; $Bz$; $Ay$; $Cz$; $Ax$; $Cy$. Before making their decision for a specific patient, primary care physician were informed about their remuneration, cost, bonus, and profit, as well as about the patient benefit for each quantity from 0 to 10. All monetary amounts were given in EUR. The procedure was exactly the same in the second part of the experiment.

After having finished the second part of the experiment, we asked primary care physicians to complete a questionnaire on social demographics (age and gender), on risk preferences (based on questions included in the German Socio Economic Panel; see Dohmen et al., 2011), on the social traits altruism and competitiveness (based on questions included in the European Values Study; European Values Study, 2016), and on their general attitude regarding pay for performance. For the full set of questionnaire items we employed in our study, see Appendix A.7.

We employed a random-choice payment technique. At the end of the experiment, one decision in each part of the experiment was randomly determined to be relevant for a primary care physician's actual payoff and for the patient benefit. Physicians were paid according to these two randomly determined choices. We paid only one decision per part to rule out income effects. The Zi notary transferred the money to the primary care physicians by the double-blind payment procedure. He also transferred the sum of patient benefits resulting from the two randomly determined decisions to *Christoffel Blindenmission*, which used the money exclusively to support surgical treatments of cataract patients in a hospital in Masvingo (Zimbabwe) staffed by ophthalmologists from the charity.[17].

[17] Similar or equivalent mechanisms have been employed in recent behavioral experiments in health analyzing physician behavior (Hennig-Schmidt et al., 2011; Hennig-Schmidt and Wiesen, 2014; Kesternich et al., 2015; Godager et al., 2016; Brosig-Koch et al., 2016b,a, 2017; Lagarde and Blaauw, 2017; Wang et al., forthcoming; Di Guida et al., 2019; Martinsson and Persson, 2019)

Physicians earned, on average, about €45.93 for the experiment, which lasted on average for 25 minutes.[18] In total, €5,002.50 were transferred to Christoffel Blindenmission, on average €47.64 per patient. The average cost for a cataract operation amounts, according to Christoffel Blindenmission, to about €30. Thus, the main experiment allowed 166 patients to be treated. The procedure in the control experiments was very similar to main experiments. For details, see Appendix C.1.

## 1.5 Sample characteristics

Besides physicians' main characteristics (age, gender, and experience in practice, we observe detailed administrative data at a primary care physicians' practice level. In our analysis, we also link the behavioral data from the experiment with physicians' practice characteristics. The first column of Table 2 presents an overview on physician and practice characteristics of our full sample. The second and the third columns show descriptives for primary care physicians in the Low-bonus and High-bonus conditions, respectively.

Our sample of primary care physicians is similar in terms of age and gender, compared to the entire population of primary care physicians in Germany. Comparing the sample to data from the federal registry of physicians in Germany (*Bundesarztregister*) in the year 2015 shows that our sample is very similar to all primary care physicians in Germany with regard to gender and age. The fraction of females is 34.6 percent in our sample, compared to 39.2 percent in Germany. Also, the age of the participants is quite similar. While the average age in the experiment is 55.9, it is 55.5 for all primary care physicians.

As our sample is a subsample of primary care physicians of the Zi-Praxis-Panel (ZiPP), we compare our sample with all primary care physicians in the ZiPP in 2015. Here, the similarity holds for age and gender. In the ZiPP, 38.9 percent of primary care physicians are female, and 72.1 percent are not older

[18] The payment is equivalent to an hourly payment of €110.23 and is about three times as high as the primary care physicians' average net hourly rate of €35 reported by Zi for 2015. However, it roughly corresponds to gross hourly rates of at least €65.

than 60 years. In our sample, this fraction is 71.2 percent. Also, our samples' annual profit, share of SHI patients, revenue share from SHI patients, and time spent with SHI patients are not significantly different from the primary care physicians in the ZiPP sample ($p$ >0.466, two-sided $t$-tests).

## Table 2: Sample characteristics

| | Full sample | High bonus | Low bonus |
|---|---|---|---|
| | ($N = 104$) | ($N = 53$) | ($N = 51$) |
| **A. Physician characteristics** | | | |
| *Main characteristics* | | | |
| Gender | | | |
| % female | 34.6 | 37.7 | 31.4 |
| Age (Mean, s.d.) | 55.80 (7.16) | 55.50 (7.61) | 56.20 (6.71) |
| Practice years (Mean, s.d.) | 27.80 (7.62) | 27.52 (8.14) | 28.09 (7.11) |
| *Self-reported attitudes* | | | |
| Risk (Mean, s.d.) | | | |
| General | 4.77 (2.35) | 5.11 (2.45) | 4.11 (2.19) |
| Own health | 4.71 (2.30) | 4.57 (2.36) | 4.86 (2.24) |
| Patients' health | 2.87 (1.45) | 2.64 (1.33) | 3.10 (1.54) |
| Altruism (Mean, s.d.) | 4.76 (2.30) | 4.55 (2.37) | 4.98 (2.23) |
| Competition (Mean, s.d.) | 3.64 (1.98) | 3.58 (1.95) | 3.71 (2.02) |
| **B. Practice characteristics** | | | |
| Annual profit | | | |
| $< Median$ (€147,000) | 45.5% | 47.1% | 42.9% |
| Location of practice | | | |
| City | 29.8% | 41.5% | 17.7% |
| Outer conurbation | 35.6% | 30.2% | 41.2% |
| Rural | 34.6% | 28.3% | 41.2% |
| Share of SHI patients | | | |
| < 87% | 16.3% | 5.9% | 27.7% |
| 87% – 90% | 22.5% | 21.6% | 23.4% |
| 90% – 93% | 25.5% | 29.4% | 21.3% |
| 93% – 96% | 19.4% | 23.5% | 14.9% |
| > 96% | 16.3% | 19.6% | 12.8% |
| Revenue share from SHI patients | | | |
| < 77% | 19.0% | 14.9% | 22.9% |
| 77% – 85% | 24.2% | 21.3% | 27.1% |
| 85% – 90% | 16.8% | 17.0% | 16.7% |
| 90% – 94% | 19.0% | 27.7% | 10.4% |
| > 94% | 21.1% | 19.2% | 22.9% |
| Share of time spend on SHI patients | | | |
| < 80% | 18.1% | 14.6% | 21.7% |
| 80% – 87% | 26.6% | 29.2% | 23.9% |
| 87% – 90% | 19.2% | 15.6% | 23.9% |
| 90% – 94% | 17.0% | 18.8% | 15.2% |
| > 94% | 19.2% | 22.9% | 15.2% |
| Physicians working in practice | | | |
| 1 | 55.0% | 56.9% | 53.1% |
| 2 | 32.0% | 27.5% | 36.7% |
| 3+ | 13.0% | 15.7% | 10.2% |

*Notes:* This table presents summary statistics of practices' and individual physicians' characteristics for (i) the full physician sample of our experiment, (ii) for physicians in the experimental condition High-bonus and (iii) for physicians in the Low-bonus condition. The practice characteristics and the physicians' gender are from an administrative data set of the Zi, and the remaining variables are from a self-reported questionnaire. Subjects could choose on a scale from 1 to 10 with 1 being the lowest and 10 the highest value for risk, altruism or competition, respectively. Table A.2 in Appendix A gives a full description of all variables. The administrative data were released in 2015.

Table 2 also shows subjects' self-reported attitudes towards risk, altruism, and competitiveness to be chosen on a scale from 1 to 10 with 1 being the lowest and 10 the highest value for each of the three attitudes. About 61.5 percent of physicians are risk averse, in the sense that they choose at most a number of five on the scale of general risk attitudes (average 4.77). The average willingness to take risks is to some extent higher for their own health (4.71) and much lower for the patients' health (2.87). 50 percent of physicians choose a value of at least four on the scale of altruism (average 4.76). That means there is a slight tendency towards the attitude that "most of the time people are mostly just looking out for themselves". Also, the majority views competition as harmful (average: 3.64, median 3).

Physicians' characteristics and practice characteristics are not significantly different between the two experimental conditions. With the exception of the practice location and the share of SHI patients, the experimental conditions High bonus and Low bonus are not significantly different ($p > 0.135$, two-sided Mann-Whitney-U tests). High-bonus differs significantly from low bonus regarding the location of the practice ($p = 0.031$) and the share of SHI patients ($p = 0.011$). As Table 2 shows, in condition High bonus more physicians practice in the city and considerably fewer physicians treat fewer than 87 percent of the SHI patients.

## 2 Behavioral hypotheses

To organize our thoughts and to derive behavioral hypotheses on the introduction of performance pay in the experiment, we introduce an illustrative model of physician behavior. In our model, we assume that the physician is altruistic on behalf of the patient, an assumption which has become common in modeling the behavior of physicians since Arrow (1963) coined the importance of physicians' patient-regarding motivation.

Similar to earlier models of physician behavior (e.g., Ellis and McGuire, 1986, 1990), we assume that a physician derives utility from increasing her own profit and the patients' health benefit. The weight the physician attaches to the patient's health benefit is interpreted as physician altruism. A physician chooses the quantity of medical services $q$ to maximize her utility:

$$U(q) = \alpha H(q) + \beta(\Lambda - c(q)) + \gamma b_l I_{b_l}, \tag{3}$$

with $\alpha, \beta$, and $\gamma \geq 0$. $\alpha$ is the weight the physicians attaches to the patient's health benefit $H(q)$, the patient-regarding altruism; $\beta$ is the physician's weight on profit from the lump-sum capitation payment $(\Lambda - c(q))$; and $\gamma$ is a measure for the physician's weight on the performance-based discrete bonus $b_l I_{b_l}$. We assume that a physician derives utility from receiving the performance-based discrete bonus. [19] We also assume the weights such that $\alpha + \beta + \gamma = 1$.

**Capitation (CAP).** Under CAP, $b_l = 0$. The quantity of health care services maximizing the physician's utility is $q^{\text{Max}} = 2\theta\frac{\alpha}{\beta}$. This means the utility-maximizing quantity increases in the marginal health benefit (as we only consider $\theta \geq 0$), and the concern for the patient's health ($\alpha$), while it decreases in the physician's weight on her profit margin ($\beta$). The extent to which a patient is underserved depends on the severity of illness which determines the patient-optimal quantity $q^*$. For given values of the constants ($\theta$, $\alpha$, and $\beta$), it follows that the intensity of underprovision is highest for patients with a high severity of illness ($q^* = 7$) and less so for patients with an intermediate (with $q^* = 5$), or mild severity of illness (with $q^* = 3$). For a profit-maximizing physician (with $\alpha = 0$), for example, the utility-maximizing quantity is $q^{\text{Max}} = 0$, illustrating that undertreatment is most pronounced for high-severity patients. We expect CAP to incentivize underprovision of medical services, which increases with the patients' severity of illness, while it decreases with the patient's marginal health benefit.

**Performance pay (CAP+P4P).** We now address the effect of physician performance pay with $b_l > 0$. The discrete bonus system we consider links a bonus

[19] As we explicitly model the effect of introducing performance pay has on a physician's utility, we make use of a multi-attribute utility function (e.g., Keeney and Raiffa 1976). This also allows us to consider potential adverse effects due to the introduction of performance pay such as crowding-out of altruistic behavior, for more see Appendix E.

payment to the patient's health benefit; performance pay thus aligns the interests of the physician and the patient. Since we assume $\gamma$ to be constant for the moment, the physician's utility increases in the size of bonus payment $b_l$ if the quantity of health care services is within $q^* + \epsilon$ with $\epsilon \in \{-1, 1\}$ (i.e., the performance pay interval). Since the physician's utility under performance pay shows discontinuities at $q^* + \epsilon$, we distinguish between the following cases.

First, physicians with a very high preference for their own profit margin (with a high $\beta$) provide a quantity below the performance pay interval ($q^{\text{Max}} < q^* - 1$). These physicians do not change their provision behavior. Second, highly altruistic physicians (with a high $\alpha$) do not change their provision behavior either, since they already provided a quantity within the performance pay interval even without performance pay ($q^{\text{Max}} > q^* - 1$). For those physicians, the performance pay is an additional payment that does not alter their behavior. Third, the intermediate type of physicians, who value both the patient's benefit as well as their own profit margin as important change the provision behavior under performance pay. Without performance pay, these physicians would chose $q^{\text{Max}}$, but with performance pay the medical service quantity is $q^* - 1 > q^{\text{Max}}$. The performance payment $b_l > 0$ compensates them for the higher quantity of health care services and underprovision is reduced.

Figure 1 illustrates the best responses for physicians with different patient-regarding motivations. Physicians with a high preference for their own profit are illustrated in area $A$, , the intermediate type in $B$, and highly altruistic physicians in $C$. We summarize in:

**Hypothesis 1.** *Performance pay reduces underprovision of medical services and enhances the quality of medical care.*

We also investigate whether the effect of performance pay is specific to the patient's severity of illness ($l$) and to the patient's marginal health benefit ($\theta$). First, higher severities increase $q^* - 1$. Physicians' utility trade-off varies between the largely profit-maximizing physicians (area $A$) and the intermediate types (area $B$). For a given constant performance pay $b_l$, this implies an in-

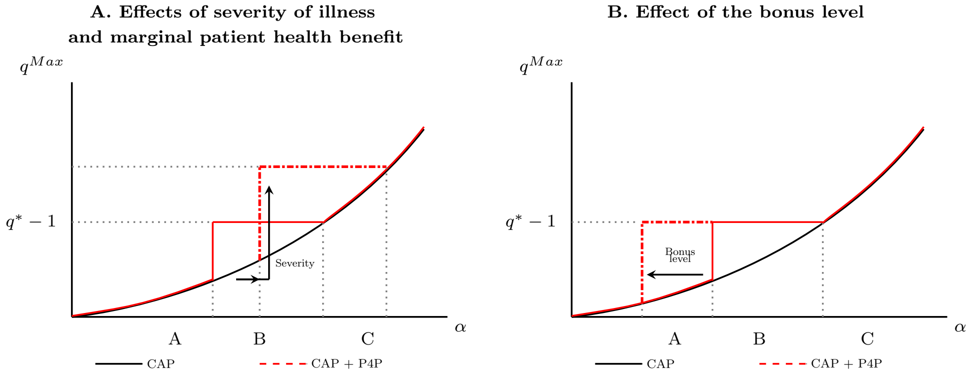


Figure 1: Illustration of the behavioral predictions

*Notes:* This figure shows optimal provision behavior for given patient-regarding altruism under CAP and CAP+P4P. The areas A, B, and C separate physicians with low, medium, and high altruism preferences.

crease in underprovision with a higher severity since area $B$ decreases. However, there is a counterveiling effect, since the utility trade-off between the medium and high altruism types is also influenced by severity. A higher severity of an illness means that the performance threshold is higher, which, *ceteris paribus*, decreases underprovision.

Whenever area A increases only weakly, due to a larger incentive, and area B increases more strongly (see Figure 1), underprovision decreases with a higher bonus payment. Note that in general the trade-off is ambiguous. Second, an increase in $\theta$ implies two potential effects. A higher health benefit increases ceteris paribus $q^{\text{Max}}$. The area $B$ increases to the left. On the contrary, the area $C$ of the high-altruism type of physicians increases to the detriment of the intermediate types, since $q^{\text{Max}}$ increases faster in $\alpha$ the higher $\theta$. In sum, we state:

**Hypothesis 2.** *The effect of performance pay implying a decrease in underprovision of health care increases in the patients' severity of illness and a higher marginal health benefit.*

Given the intuition above, it follows naturally that the level of bonus $b_l$ affects the intermediate altruistic types. A higher bonus biases the utility comparison between $U(q^{\text{Max}})$ and $U(q^* - 1)$ towards the latter. Therefore, area $B$ in Figure 1 increases to the 'burden' to rather profit-maximizing physicians. As a result, more physicians provide $q^* - 1$ instead of $q^{\text{Max}} < q^* - 1$; see Figure 1 for an illustration. In sum, we hypothesize:

**Hypothesis 3.** *An increase in the bonus level further reduces the underprovision of medical services.*

# 3 Behavioral results

## 3.1 Behavior under capitation

Before we analyze our main hypothesis on the effect of pay for performance (Hypothesis 1), we investigate the physicians' medical service provision under baseline capitation (CAP). Under CAP, the average quantity of medical services is 4.27 in both bonus conditions, which indicates a tendency to underserve the average patient.[20] Underprovision occurs for all patients under CAP; see Table 3. CAP incentivizes physicians significantly to underprovide health care services for all nine patients in the Low-bonus condition. In the High-bonus condition, eight of nine patients are significantly underserved ($p$ <0.014). Only patient $Ax$ with illness $A$ at a mild severity $x$ is not significantly underserved ($p = 0.207$); see Table B.1 in Appendix B. We also find that underprovision is not significantly affected by the marginal health benefit; see the estimation results from OLS regressions and Wald test results in Table B.2 in Appendix B.

The behavior of physicians under CAP implies that the quality of medical care is suboptimal. To quantify further the incentive effect on the quality of care, we consider a relative quality measure which is comparable across severities of illness: $\rho_{ikl} = (q_{ikl} - |q_{ikl} - q^*_l|)/q^*_l$. When physician $i$ does not deviate from the patient-optimal quantity $q^*_l$, the quality of care is optimal ($\rho_{ikl} = 1$) for a patient with illness $k$ and severity $l$. We find the relative quality to vary across patients

[20] Notice that we find no significant difference between the two conditions ($p \geq 0.700$, Mann-Whitney U-test). Throughout the paper, $p$-values are reported from two-sided tests. For between-subject analyses, we employ Mann-Whitney U-tests; for within-subject analyses, we use the Wilcoxon signed rank-test. In addition, we provide test statistics of Fisher-Pitman permutation tests for paired and unpaired samples.

between 0.79 and 0.90;[21] see Column "CAP" in Panel B of Table 3. The quality of care is significantly different from the optimal care ($p < 0.004$). This finding are in line with empirical studies (e.g., Cutler, 1995) and behavioral experiments (e.g. Hennig-Schmidt et al., 2011; Brosig-Koch et al., 2017).

### 3.2 The effect of performance pay on health care provision

We now analyze whether physicians behave according to Hypothesis 1 by comparing physicians' medical service provision under CAP and CAP+P4P. When complementing CAP with performance pay, we find that physicians choose a higher quantity of medical services. On aggregate, the quantity of medical services increases from 4.27 under CAP to 4.58 and to 4.63 under Low-bonus and High-bonus, respectively; see Table 3. This is an increase by about 7% under the Low-bonus and by about 8% under the High-bonus condition. The underserving of the average patient is reduced by about 43% under the Low-bonus and by 49% under the High-bonus condition. For the distribution of the physicians' quantity choices in both bonus conditions, see Figures B.1 and B.2 in Appendix B.

To quantify further the effect of performance pay, we use a linear multilevel mixed effects model fit by restricted maximum likelihood, and we include random effects for subjects and patients. We employ this model as it is well-suited for our hierarchical panel data structure. The model comprises subjects (physicians) specified at level 3 of clustering, patients at level 2, and the experimental stage $s$ with $s = 1$ for CAP and $s = 2$ for the P4P systems at level 1. The specification is as follows:

$$\begin{aligned} q_{sij} &= \beta_0 + \beta_1 \mathrm{T}_j + \beta_2 \mathrm{P4P}_s + \beta_3 \mathrm{S}_i + \beta_4 \mathrm{MHB}_i + \beta_5 \mathrm{S}_i \times \mathrm{P4P}_s \\ &+ \beta_6 \mathrm{PHY}_j + u_{0j} + u_{1j} \times \mathrm{P4P}_s + u_{0i} + \epsilon_{sij}, \end{aligned} \tag{4}$$

where $q_{sij}$ denotes physician $j$'s quantity choice (level 3) for patients $i$ (level 2) in

[21] It is plausible to limit relative quality to a lowest value of 0. The four choices out of the total of 1,872 decisions where our definition leads to negative relative quality were set to zero as well.

Table 3: Quantity and quality of health care provision by payment system, illness, and severity of illness

| | A. Quantity, $q_{kl}$ | | | | B. Relative quality, $\rho_{kl}$ | | | |
|---|---|---|---|---|---|---|---|---|
| | CAP | CAP+P4P | %-change | $p$-value | CAP | CAP + P4P | %-change | $p$-value |
| **Low-bonus (5%)** | | | | | | | | |
| Mild severity of illness | | | | | | | | |
| Illness *A* | 2.76 (0.62) | 2.84 (0.86) | 2.83 | 0.786 [0.711] | 0.88 (0.19) | 0.88 (0.23) | -0.74 | 0.671 [0.833] |
| Illness *B* | 2.55 (0.90) | 2.73 (0.67) | 6.92 | 0.188 [0.276] | 0.83 (0.29) | 0.88 (0.21) | 5.47 | 0.345 [0.321] |
| Illness *C* | 2.73 (0.67) | 2.82 (0.71) | 3.60 | 0.328 [0.542] | 0.90 (0.22) | 0.88 (0.21) | -2.19 | 0.369 [0.760] |
| Intermediate severity of illness | | | | | | | | |
| Illness *A* | 4.36 (1.13) | 4.57 (0.83) | 4.95 | 0.265 [0.246] | 0.85 (0.22) | 0.91 (0.16) | 5.97 | 0.139 [0.152] |
| Illness *B* | 4.14 (1.00) | 4.52 (1.05) | 9.48 | 0.001 [0.006] | 0.81 (0.19) | 0.89 (0.20) | 9.66 | 0.002 [0.006] |
| Illness *C* | 4.35 (0.84) | 4.69 (0.95) | 7.66 | 0.003 [0.023] | 0.85 (0.16) | 0.91 (0.18) | 5.96 | 0.005 [0.032] |
| High severity of illness | | | | | | | | |
| Illness *A* | 5.78 (1.80) | 6.25 (1.25) | 8.13 | 0.014 [0.009] | 0.83 (0.26) | 0.89 (0.18) | 8.14 | 0.014 [0.009] |
| Illness *B* | 5.86 (1.41) | 6.31 (1.39) | 7.69 | 0.001 [0.092] | 0.83 (0.19) | 0.90 (0.20) | 8.47 | 0.002 [0.024] |
| Illness *C* | 5.92 (1.44) | 6.45 (1.24) | 8.94 | 0.000 [0.001] | 0.84 (0.20) | 0.92 (0.17) | 9.00 | 0.000 [0.001] |
| Aggregated | 4.27 (1.73) | 4.58 (1.77) | 6.69 | | 0.84 (0.21) | 0.89 (0.19) | 5.52 | |
| | | | | | | | | |
| **High-bonus (20%)** | | | | | | | | |
| Mild severity of illness | | | | | | | | |
| Illness *A* | 3.02 (1.20) | 2.81 (0.48) | -6.88 | 0.540 [0.265] | 0.84 (0.26) | 0.91 (0.15) | 8.21 | 0.154 [0.059] |
| Illness *B* | 2.64 (0.79) | 2.75 (0.62) | 4.29 | 0.448 [0.413] | 0.86 (0.25) | 0.89 (0.19) | 4.41 | 0.306 [0.304] |
| Illness *C* | 2.64 (1.15) | 2.74 (0.62) | 3.57 | 0.359 [0.642] | 0.79 (0.31) | 0.89 (0.20) | 12.8 | 0.012 [0.006] |
| Intermediate severity of illness | | | | | | | | |
| Illness *A* | 4.32 (1.12) | 4.60 (0.57) | 6.55 | 0.057 [0.096] | 0.84 (0.21) | 0.91 (0.11) | 8.52 | 0.034 [0.017] |
| Illness *B* | 4.34 (1.37) | 4.58 (0.69) | 5.65 | 0.018 [0.306] | 0.82 (0.24) | 0.92 (0.14) | 12.50 | 0.000 [0.000] |
| Illness *C* | 4.28 (1.13) | 4.60 (0.72) | 7.49 | 0.003 [0.074] | 0.83 (0.21) | 0.91 (0.14) | 9.50 | 0.003 [0.003] |
| High severity of illness | | | | | | | | |
| Illness *A* | 6.06 (1.39) | 6.38(1.15) | 5.30 | 0.095 [0.130] | 0.87 (0.20) | 0.91 (0.16) | 5.2 | 0.097 [0.122] |
| Illness *B* | 5.51 (1.69) | 6.62 (0.69) | 20.21 | 0.000 [0.000] | 0.79 (0.24) | 0.95 (0.10) | 20.2 | 0.000 [0.000] |
| Illness *C* | 5.58 (1.78) | 6.57 (0.64) | 17.57 | 0.000 [0.000] | 0.82 (0.25) | 0.94 (0.09) | 17.57 | 0.000 [0.000] |
| Aggregated | 4.27 (1.79) | 4.63 (1.69) | 8.43 | | 0.82 (0.24) | 0.91 (0.15) | 11.00 | |

*Notes:* This table shows descriptive statistics on the quantities and relative quality of health care provision at the level of payment systems, illnesses, and severities of illness (means and standard deviations in brackets). Of the 104 German resident primary care physicians, 53 decide in the High-bonus (20%) and 51 in the Low-bonus (5%) condition. Two-sided $p$-values of pairwise comparisons between quantities and qualities in CAP versus CAP+P4P are shown for Wilcoxon signed rank tests for matched samples, and for Fisher-Pitman permutation tests for paired samples in squared brackets. Notice that percentage changes between payment systems are based on non-rounded values.

experimental stage $s$ (level 1). Indicator $i$ denotes the second level of clustering, which accounts for observations for each patient $i$ over stage $s$.

The fixed effects part of the model contains the constant $\beta_0$, fixed effects for the treatment group T, a dummy equal to 1 indicating the High-bonus and equal to 0 the Low-bonus condition, which is time invariant, a dummy for the second stage of the experiment where performance pay is introduced, a vector for patients' severities (S), a dummy for patients' marginal health benefit (MHB), and the interaction between severities and performance pay S × P4P. The vector PHY contains the physicians' individual time-invariant characteristics such as experience (years in practice) and self-reported attitudes towards risk, competition, and altruism. $\beta_1$ denotes changes due to the variation of the bonus level, while $\beta_2$ indicates average differences in the dependent variable over stages 1 and 2 of the experiment. Averaged over all subjects, $\beta_3$ measures the effect of the patients' severity of illness ($l$) being either low, intermediate, or high, while $\beta_4$ the effect of the marginal health benefit $\theta$ being 1 or 2, $\beta_5$ measures the effect of the interaction of performance pay and the severity of illness, and $\beta_6$ measures the effect of physicians' individual characteristics.

The random effects are assumed to be independent of each other between levels, and all random effects are independent of the residuals. The overall residuals $\epsilon_{sij}$ are assumed to be independent and normally distributed with a mean of 0 and a constant variance $\sigma^2$. Further, we assume $u_i \sim (0, \sigma^2)$ for the random errors at the patient level (level 2). The joint distribution of the two random effects associated with physician $j$ (i.e., the random intercept denoted by $u_{0j}$ and the random slope for stage 2 of the experiment denoted by $u_{1j}$) is $u_j = \binom{u_{0j}}{u_{1j}} \sim N(0, \mathbf{D})$ (level 3). We specify an unstructured $\mathbf{D}$ matrix by $\begin{pmatrix} Var(u_{0j}) & cov(u_{0j}, u_{1j}) \\ cov(u_{0j}, u_{1j}) & Var(u_{1j}) \end{pmatrix}$.

Estimation results indicate that the quantity of care under CAP+P4P is significantly larger compared to CAP; see Model (1) in Panel A of Table 4. This effect is also robust when controlling for the physicians' individual characteristics; see Model (2) in Panel A of Table 4. The observed significant increase in health care quantity under CAP+P4P relative to CAP is in line with Hypothesis 1.

We also assess the impact of performance pay on the quality of care. Under CAP, the average relative quality is 0.84 in the Low-bonus condition and 0.82 in the High-bonus condition. Quality increases by 5 and 9 percentage points under CAP+P4P to 0.89 and 0.91, respectively. To quantify the effect of performance pay on quality further, we employ a model specification, similar to equation (4):

$$\rho_{sj} = \beta_0 + \beta_1 \mathrm{T}_j + \beta_2 \mathrm{P4P}_s + \beta_3 \mathrm{PHY}_j + u_{0j} + u_{1j} \times \mathrm{P4P}_s + \epsilon_{sj}. \tag{5}$$

Estimation results indicate that the relative quality of care under P4P increases highly significantly by about 7% on average compared to CAP; see Models (5) and (6) in Table 4. We summarize our findings as follows:

**Result 1.** *Performance pay significantly reduces underprovision of health care services prevalent under capitation, increasing the quality of care.*

### 3.3 The effects of patient characteristics and the bonus level

We now analyze whether the effect of performance pay is specific to the patients' severity of illness according to Hypothesis 2. Descriptive analyses indicate heterogeneity in the physicians' responses to performance pay according to the patients' severities of illness; see Panel A of Table 3. In particular, for patients with intermediate and high severities of illness, physicians provide significantly more health care services in both the Low- and the High-bonus condition under performance pay ($p \leq 0.095$, Wilcoxon signed rank-test). The heterogeneity in physicians' quantity choices under performance pay is illustrated in panels B and C of Figures B.1 and B.2 in Appendix B.

Estimation results confirm that the effect of performance pay on physicians' behavior is specific to the patients' severities of illness. The reduction in underprovision of health care services increases in the severities of patients' illnesses; see Models (3) and (4) in Table 4. This result supporting Hypothesis 2 is robust when controlling for the physicians' characteristics. The marginal health benefit does not affect the physicians' decisions significantly; see Models (3) and (4) in Table 4 and Table B.3 in Appendix B. In sum, we state the following result:

**Result 2.** *The magnitude of the effect of performance pay on the physicians' underprovision significantly increases in the patients' severities of illness. The patients' marginal health benefit does not significantly affect physicians' behavior.*

Table 4: Physicians' health care service provision and quality of care under capitation and performance pay

| | A. Quantity $q$ | | | | B. Relative quality, $\rho_{kl}$ | |
|---|---|---|---|---|---|---|
| Model: | (1) | (2) | (3) | (4) | (5) | (6) |
| **Fixed effects** | | | | | | |
| Performance pay | 0.334*** | 0.334*** | 0.058 | 0.058 | 0.068*** | 0.068*** |
| | (0.071) | (0.071) | (0.087) | (0.087) | (0.013) | (0.013) |
| High bonus (= 1 if 20%-Bonus) | 0.036 | 0.106 | 0.036 | 0.106 | 0.010 | 0.030 |
| | (0.125) | (0.131) | (0.125) | (0.131) | (0.026) | (0.027) |
| Interm. severity of illness (= 1 if $l = y$) | 1.694*** | 1.694*** | .574*** | 1.574*** | | |
| | (0.049) | 0.049 | (0.065) | (0.065) | | |
| High severity of illness (= 1 if $l = z$) | 3.356*** | 3.356*** | 3.061*** | 3.061*** | | |
| | (0.049) | (0.049) | (0.065) | (0.065) | | |
| High marginal health benefit (= 1 if $\theta = 2$) | 0.016 | 0.016 | 0.016 | 0.016 | | |
| | (0.043) | (0.043) | (0.043) | (0.043) | | |
| Performance pay× Interm. severity | | | 0.240** | 0.240** | | |
| | | | (0.086) | (0.086) | | |
| Performance pay× High severity | | | 0.590*** | 0.590*** | | |
| | | | (0.086) | (0.086) | | |
| Physician characteristics | No | Yes | No | Yes | No | Yes |
| Constant | 2.562*** | 1.758*** | 2.700*** | 1.836*** | 0.831*** | 0.646*** |
| | (0.108) | (0.384) | (0.111) | (0.387) | (0.021) | (0.077) |
| **Random effects** | | | | | | |
| Subject level | | | | | | |
| Var(Patient/Performance pay) | 0.389** | 0.389** | 0.396** | 0.396** | 0.013*** | 0.013*** |
| | (0.073) | (0.073) | (0.073) | (0.073) | (0.002) | (0.002) |
| Var(Constant) | 0.522*** | 1.627*** | 1.659*** | 1.642*** | 0.061*** | 0.058 |
| | (0.278) | (0.277) | (0.278) | (0.277) | (0.010) | (0.009) |
| Cov(Patient/Performance pay, Constant) | -0.708*** | 0.705*** | -0.719*** | -0.716*** | -0.024*** | -0.024*** |
| | (0.026) | (0.134) | (0.135) | (0.135) | (0.005) | (.005) |
| Patient level | | | | | | |
| Var(Constant) | 0.074*** | 0.074*** | 0.090*** | 0.090*** | | |
| | (0.024) | (0.024) | (0.023) | (0.023) | | |
| Var(Residual) | 0.607*** | 0.607*** | 0.576*** | 0.576*** | 0.020 | 0.020 |
| | (0.030) | (0.030) | (0.028) | (0.028) | (0.001) | 0.001 |
| Observations | 1.872 | 1.872 | 1.872 | 1.872 | 1.872 | 1.872 |
| Physicians | 104 | 104 | 104 | 104 | 104 | 104 |

*Notes:* This table shows estimates from multilevel mixed-effects REML regressions. Standard errors are in parentheses. Models 1 to 4 include subject-specific and patient-specific random effects with stage at level 1, patients at level 2, and subjects at level 3. Models 5 and 6 include subject-specific random effects with patients at level 1and subjects at level 2. The reference category is the 'mild severity of illness', $l = x$. Physician characteristics comprise gender, practice years, a question each for the attitude towards altruism and competition from the European Values Study (European Values Study, 2016), and risk attitudes according to the German Socio-Economic Panel on the willingness to take risk in general, related to health (Dohmen et al., 2011) and one question related to a patient's health risk. OLS regressions and fractional probit models yield very similar estimation results; see Tables D.1 and D.2 in Appendix D. *** $p < 0.001$, ** $p < 0.01$, and *$p < 0.05$.

To analyze how the level of the bonus payment affects the physicians' medical service provision (Hypothesis 3), we compare the quantities and quality of health care services under Low-bonus (5%) and High-bonus (20%) conditions. For all severities, we find no significant differences in average medical services per subject between the two conditions ($p \geq 0.4964$, Mann-Whitney U-tests). Regression analyses show a tendency that the High-bonus incentivizes higher health care service provision and higher quality than Low-bonus. This effect is not significant, however; see Models (1) to (6) in Table 4. In sum, we state:

**Result 3.** *The level of the bonus payment being either low (5%) or high (20%) complementing a lump-sum capitation affects neither the physicians' health care service provision nor the quality of care.*

**Robustness of the main experimental results.** Our analyses of behavioral data from control experiments with medical students indicates that the effect of performance pay on health care service provision is robust (i) between physician and medical student samples, (ii) towards keeping the level of incentives constant between capitation and blended capitation plus performance pay, and (iii) concerning the order of payment systems. We detail these robustness checks in Appendix C.2.

### 3.4 Unintended consequence of performance pay: Crowding-out of patient-regarding behavior

Incentives such as pay-for-performance may have unintended consequences for the intrinsic motivation of service providers in the public domain, a very important motivation of individuals providing services (e.g., Bénabou and Tirole, 2003; 2006). Crowding out has been reported in contexts and areas other than health care (e.g., Gneezy and Rustichini, 2000; Heyman and Ariely, 2004; Mellström and Johannesson, 2008; Ariely et al., 2009; Huffman and Bognanno, 2018).

In health care provision, where physicians' other-regarding motivation is essential for high-quality patient care, its potential reduction has been pointed out as one of the major pitfalls of pay-for-performance systems. While many studies have focused on the cost effectiveness of the payment systems (see, e.g., Maynard 2012; Miller and Babiarz, 2014), Glasziou et al. (2012) stress the importance of systematically analyzing potentially harmful effects on patient outcomes since performance pay may incentivize physicians to change their behavior in an unintended way that is detrimental to the patients' health (e.g., Woolhandler et al., 2012). Our within-subject design allows us to analyze exactly this effect, namely whether the introduction of performance pay incentivizes an individual physician to crowd out patient-regarding behavior such that patients suffer under P4P compared to CAP. This has not been experimentally studied in a health context before.

For our descriptive analysis, we consider the individual physician × patient level data (1872 data points). We distinguish between three main treatment patterns: Quantity choices that maximize physician profit (PM), quantity choices maximizing the patient benefit (BM), and trade-off choices (TO), which capture Pareto-efficient quantity choices, but are neither PM nor BM. The category 'Other' comprises Pareto-inferior medical service provision. As we do not observe significant differences between Low- and High-bonus conditions for both parts of the experiment, we pool the data for our classification of choices.[22] We observe the following patterns under CAP: PM 1%, BM 54%, TO 42%, and Other 3% of all the choices. Under CAP+P4P, we find PM: 30%, BM: 64%, and Other: PM: 6%.

When performance pay is introduced, we observe that PM increases by 29 percentage points, BM by 10, and Other by 3 percentage points. Despite the rise in BM, we do find some evidence for a crowding-out of altruistic behavior, which is characterized by a physician's choice transitioning from BM to PM for the same patient. Thus, the physician moves away from the patient-optimal service provision under CAP to his profit-maximizing treatment choice under CAP+P4P, which is $q^* - 1$, i.e., one unit below the patient's optimum. In total,

[22] As the bonus is paid in case the chosen quantity does not differ by more than one unit from the patient-optimal treatment, the classification TO does not exist under P4P+CAP. All choices classified as 'Other' under P4P+CAP are Pareto-inferior.

crowding out amounts to 7% of all physicians' choices, which is 14% of all BM choices under CAP. 22% of choices transition from TO to PM. We also observe crowding-in (PM to BM) which amounts to 1%. Finally, 17% of choices transition from TO to BM.

We next analyze how the crowding-out of patient-regarding behavior relates to patients' characteristics. Table 5 shows estimation results (marginal effects) from probit regressions. We find that crowding-out is significantly affected by the marginal health benefit in that the likelihood of crowding-out is significantly lower when the patients' marginal health benefit is high; see Model (1) in Table 5. This finding is robust when including controls for physicians' characteristics (Model (2) in Table 5) and when we only consider the subsample of benefit-maximizing choices in CAP; see Models (3) and (4) in Table 5.

Table 5: Probit regressions on crowding-out of patient-regarding behavior, avg. marginal effects

| | Full sample | | Sample restricted to benefit maximizers in CAP | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| High bonus (= 1 if 20%-Bonus) | 0.012<br>(0.045) | 0.012<br>(0.043) | 0.012<br>(0.046) | 0.012<br>(0.043) |
| Intermediate severity (= 1 if $l = y$) | 0.012<br>(0.037) | 0.016<br>(0.036) | 0.012<br>(0.038) | 0.016<br>(0.036) |
| High severity (= 1 if $l = z$) | -0.018<br>(0.043) | -0.017<br>(0.042) | -0.018<br>(0.043) | -0.017<br>(0.041) |
| High marginal health benefit (= 1 if $\theta = 2$, Illness $C$) | -0.060*<br>(0.028) | -0.057*<br>(0.028) | -0.060*<br>(0.027) | -0.058*<br>(0.028) |
| Physician characteristics | No | Yes | No | Yes |
| Observations | 936 | 936 | 503 | 503 |

*Notes:* The table shows marginal effects from probit regressions with robust standard errors clustered for subjects (in parentheses). The reference category is 'mild severity', $l = z$. The variable 'High marginal health benefit' is a dummy equal to 1 if $\theta = 2$ for illness $C$, and = 0 if $\theta = 1$ for illnesses $A$ and $B$. Logit regressions yield very similar estimation results; see Table D.3 in the Appendix. *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$.

**Observation 1.** *Our behavioral data show some evidence for crowding-out of patient-regarding behavior when performance pay is introduced. Crowding-out is less pronounced for patients with a high marginal health benefit.*

In order to rationalize further the observed crowding-out of patient-regarding behavior, we introduce a behavioral model in Appendix E.

# 4 Quality of care and physicians' practice characteristics

In this section, we link the behavioral data to physicians' practice characteristics from our administrative data set. The characteristics comprise annual practice profit, practice location (city, outer conurbation, rural area), SHI patient-related characteristics (share, revenue share, time spent), and whether more than one physician is employed in the practice.

To estimate the impact of the above practice characteristics, we extend the model specification for predicting health care provision in our experiment (see equation (4)) by a vector ($\mathrm{PRAC}_j$), which yields the following new specification:

$$\begin{aligned} q_{sij} &= \beta_0 + \beta_1 \mathrm{T}_j + \beta_2 \mathrm{P4P}_s + \beta_3 \mathrm{S}_i + \beta_4 \mathrm{MH}_i + \beta_5 \mathrm{S}_i \times \mathrm{P4P}_s \\ &+ \beta_6 \mathrm{PHY}_j + \beta_7 \mathrm{PRAC}_j u_{0j} + u_{1j} \times \mathrm{P4P}_s + u_{0i} + \epsilon_{sij}. \end{aligned} \tag{6}$$

The estimation results based on the new specification in Equation (6) confirm our previous findings on the impact of performance pay and additionally indicate that the provision of health care services is significantly lower for physicians in high-profit practices; see Model (7) of Table 6. Apparently, high-profit practice physicians underprovide patients to a significantly larger extent than physicians from a low-profit practice. Moreover, service provision is significantly lower when a physician's practice is located in a city rather than in a rural region; see Models (2) and (7) of Table 6. For all other practice characteristics, we do not find a significant correlation with medical service provision in our experiment; see Models (3), (4), (5), (6), and (7) of Table 6.

Table 6: Link between quantity of health care services and physicians' practice characteristics

| Model | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| Performance pay | -0.027 | -0.027 | -0.027 | -0.027 | -0.027 | -0.027 | -0.027 |
| | (0.092) | (0.092) | (0.092) | (0.092) | (0.092) | (0.092) | (0.092) |
| High bonus | 0.138 | 0.207 | 0.121 | 0.120 | 0.125 | 0.100 | 0.246 |
| | (0.148) | (0.151) | (0.157) | (0.151) | (0.150) | (0.150) | (0.156) |
| Interm. severity of illness | 1.582*** | 1.582*** | 1.582*** | 1.582*** | 1.582*** | 1.582*** | 1.582*** |
| | (0.066) | (0.066) | (0.066) | (0.066) | (0.066) | (0.066) | (0.066) |
| High severity of illness | 3.019*** | 3.019*** | 3.019*** | 3.019*** | 3.019*** | 3.019*** | 3.019*** |
| | (0.066) | (0.066) | (0.066) | (0.066) | (0.066) | (0.066) | (0.066) |
| High marginal health benefit | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 |
| | (0.044) | (0.044) | (0.044) | (0.044) | (0.044) | (0.044) | (0.044) |
| Performance pay × Interm. severity | 0.241** | 0.241** | 0.241** | 0.241** | 0.241** | 0.241** | 0.241** |
| | (0.084) | (0.084) | (0.084) | (0.084) | (0.084) | (0.084) | (0.084) |
| Performance pay × High severity | 0.667*** | 0.667*** | 0.667*** | 0.667*** | 0.667*** | 0.667*** | 0.667*** |
| | (0.084) | (0.084) | (0.084) | (0.084) | (0.084) | (0.084) | (0.084) |
| High annual profit | -0.255 | | | | | | -0.330* |
| | (0.147) | | | | | | (0.148) |
| City | | -0.427* | | | | | -0.552** |
| | | (0.193) | | | | | (0.195) |
| Outer conurbation | | -0.100 | | | | | -0.196 |
| | | (0.172) | | | | | (0.177) |
| Share of SHI patients | | | 0.004 | | | | -0.040 |
| | | | (0.058) | | | | (0.070) |
| Revenue share from SHI patients | | | | 0.028 | | | 0.009 |
| | | | | (0.053) | | | (0.063) |
| Group practice (=1 if no. of physicians > 1) | | | | | 0.111 | | 0.109 |
| | | | | | (0.145) | | (0.144) |
| Time spent on SHI patients | | | | | | 0.073 | 0.088 |
| | | | | | | (0.050) | (0.063) |
| Constant | 2.114*** | 1.900*** | 1.799*** | 1.684*** | 1.748*** | 1.606*** | 2.276*** |
| | (0.479) | (0.444) | (0.477) | (0.513) | (0.460) | (0.472) | (0.528) |
| Observations | 1,566 | 1,566 | 1,566 | 1,566 | 1,566 | 1,566 | 1,566 |
| Physicians | 87 | 87 | 87 | 87 | 87 | 87 | 87 |

*Notes:* This table shows parameter estimates from multilevel mixed-effects REML regressions. Standard errors are shown in parentheses. All models include subject-specific random effects and, all models contain controls for physicians' gender, practice years, as well as attitudes towards altruism and competition from the European Values Study (European Values Study, 2016), and risk attitudes according to the German Socio-Economic Panel on the willingness to take risk in general, related to health (Dohmen et al., 2011) and one question eliciting attitudes related to a patient's health. *** $p < 0.001$, ** $p < 0.01$, and *$p < 0.05$.

Next, we analyze how the relative quality of care is influenced by practice characteristics. To this end, we also extend the model specification in equation (5) by a vector of our physician practice characteristics ($\text{PRAC}_j$):

$$\rho_{sj} = \beta_0 + \beta_1 \text{T}_j + \beta_2 \text{P4P}_s + \beta_3 \text{PHY}_j + \beta_4 \text{PRAC}_j + u_{0j} + u_{1j} \times \text{P4P}_s + \epsilon_{sj}. \quad (7)$$

Again, our estimation results support the previous findings on the impact of performance pay. They further reveal that the quality of care in the experiment is about 7% lower among physicians from high-profit practices and even about 10% lower among physicians from city practices; see Model (7) of Table 7. We also find a positive relationship between physicians' quality of care and the time spent with SHI patients in Model (6) of Table 7.

Further, we test whether associations of the practice characteristics differ for low-profit and high-profit practices. We add interaction terms of high profit practices with other practice characteristics to Model (7) of Table 7; see Table D.4 in Appendix D.[23] For practices in cities, one might argue that high-profit practices provide better quality than low-profit practices due to, for example, better medical equipment and facilities. However, we do not find a significant association between profit and location, which indicates that the effect of annual profit is independent of location of practice. Moreover, the effect of annual profit does not significantly correlate with the share of SHI patients, the revenue share from SHI patients, and time spent on SHI patients. In sum, we state:

**Observation 2.** *Physicians from high-profit practices underserve patients with a significantly larger intensity than physicians from low-profit practices as do physicians practicing in the city. The quality of care is significantly higher among physicians from low-profit practices, and for physicians practicing in rural areas, and it increases in the time spent on SHI patients.*

[23] For analogous analyses, we considered split sample regressions for low and high profit practices and for locations, see Tables D.6 and D.7 in Appendix D.

Table 7: Quality of care ($\rho_{kl}$) and physician practice characteristics

| Model | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| Performance_pay | 0.055*** | 0.055*** | 0.055*** | 0.055*** | 0.055*** | 0.055*** | 0.055*** |
| | (0.014) | (0.014) | (0.014) | (0.014) | (0.014) | (0.014) | (0.014) |
| High bonus | 0.030 | 0.042 | 0.025 | 0.026 | 0.028 | 0.022 | 0.048 |
| | (0.029) | (0.030) | (0.031) | (0.029) | (0.029) | (0.029) | (0.030) |
| High annual profit | -0.054 | | | | | | -0.069* |
| | (0.029) | | | | | | (0.028) |
| City | | -0.078* | | | | | -0.106** |
| | | (0.034) | | | | | (0.037) |
| Outer conurbation | | -0.030 | | | | | -0.049 |
| | | (0.034) | | | | | (0.034) |
| Share of SHI patients | | | 0.004 | | | | -0.008 |
| | | | (0.011) | | | | (0.013) |
| Revenue share from SHI patients | | | | 0.011 | | | 0.003 |
| | | | | (0.010) | | | (0.012) |
| Group practice (=1 if no. of physicians > 1) | | | | | 0.034 | | 0.035 |
| | | | | | (0.028) | | (0.027) |
| Time spent on SHI patients | | | | | | 0.020* | 0.022 |
| | | | | | | (0.010) | (0.012) |
| Constant | 0.680*** | 0.633*** | 0.608*** | 0.570*** | 0.598*** | 0.562*** | 0.644*** |
| | (0.093) | (0.087) | (0.093) | (0.099) | (0.089) | (0.091) | (0.101) |
| Observations | 1,566 | 1,566 | 1,566 | 1,566 | 1,566 | 1,566 | 1,566 |
| Physicians | 87 | 87 | 87 | 87 | 87 | 87 | 87 |

*Notes:* This table shows parameter estimates (fixed effects) from multilevel mixed-effects REML regressions. Standard errors are shown in parentheses. All models include subject-specific random effects, and all models control for the physicians' characteristics which comprise gender, practice years, a question each for the attitudes towards altruism and competition from the European Values Study (European Values Study, 2016), and risk attitudes according to the German Socio-Economic Panel on the willingness to take risk in general, related to health (Dohmen et al., 2011), as well as one questions eliciting attitudes related to a patient's health. *** $p < 0.001$, ** $p < 0.01$, and *$p < 0.05$.

# 5 Discussion

In this section, we discuss implications for health care policy-makers and address potential limitations of our experimental results. Taking a payer's perspective, with the introduction of performance pay the remuneration increases by about 37.4% in the High-bonus and by about 22.6% in the Low-bonus condition. The health benefit of the average patient improves by about 7.5% in the High-bonus and by about 8% in the Low-bonus condition. Table 8 presents this percentage change in benefit as an arc-elasticity, similar to Brot-Goldberg et al. (2017). Model (1) is based on descriptive patient benefit values. Models (2) and (3) present elasticities obtained from the OLS estimation of benefit. Here, we control for experimental and physician characteristics (2), and additionally for physician practice characteristics (3). We use the same explanatory variables as for Model (2) in Table 4 and Model (7) in Table 6. The arc-elasticity ranges from 0.08 to 0.27 and slightly varies with the bonus size.

For example, descriptive values reveal that a 10% increase in remuneration would yield a 2.6% increase in patient benefit in the High-bonus condition and a 2.7% increase in the Low-bonus condition. The physicians' medical treatment behavior yielding patients health benefit is relatively inelastic. That is, the introduction of a substantial performance pay led to a relatively small change in the amount of benefit a patient receives.

Yet, one has to be careful when generalizing these insights to the field despite the measures we took measures to safeguard our experiment against the threats to internal and external validity. A behavioral experiment based on a highly controlled decision environment has a high internal validity, which makes it particularly useful to test causal relationships implied by economic models and allows us to find out behavioral regularities that are prohibitively difficult to detect in the field (e.g., Falk and Heckman, 2009). These are exactly the characteristics of experimental economics research that we exploit in our study. We introduce and test behavioral hypotheses and, by using a within-subject design, we are able to identify heterogeneous individual behavioral changes that *ceteris*

Table 8: Descriptives and arc-elasticities of patient health benefits and remuneration changes

| Condition | (1) | (2) | (3) |
|---|---|---|---|
| High bonus | 0.26 | 0.20 | 0.18 |
| Low bonus | 0.27 | 0.14 | 0.08 |

*Notes:* Arc-elasticity represents the %-patient benefit change by a %-remuneration change. Column (1) is based on descriptives for patient health benefits. Columns (2) and (3) present elasticities obtained from the OLS estimation of benefit. Here, we control for experimental and physician characteristics (2), and additionally for physician practice characteristics (3). We present the arc-elasticity of a patient's benefit with respect to remuneration: $\frac{(b_2-b_1)/(b_2+b_1)}{(R_2-R_1)/(R_2+R_1)}$, with $b_i$ being the mean patient health benefit per physician, and $R_i$ being the mean remuneration for the physician in part $i$.

*paribus* result from introducing performance pay. At the same time, a high control of the experimental decision environment requires one to substantially reduce the complexities compared to the field environment, which potentially affects the external validity of results.

In our study, we carefully considered these issues to minimize the potential effects as much as possible. First of all, we used a representative sample of German primary care physicians. We thus observe the behavior of a representative share of those who are central to reforms introducing performance pay.

Second, we used a medically framed decision environment, a specific context in which physicians are used to make decisions and where introducing performance pay is a highly relevant issue. We also argue that abstracting from a specific medical environment is an advantage rather than a deficiency as the participating primary care physicians need not deliberate about the effectiveness of specific medical services or how to combine them for treating a patient optimally.

Third, a potential difficulty when translating our findings to the real world might be the artificial notion of quantity in our experiment compared to the complexity in reality. We obviously abstract from tangible services and treatments including anamneses, tests, examinations, and time spent with the patient.

Moreover in reality, there tend to be genuine uncertainties over optimal care provision, and physicians may have different views on what constitutes best practice treatment. In our stylized environment, optimal care provision is known with certainty. While being motivated by theory, this design choice can be rationalized by medical guidelines indicating the optimal number of services for a patient (e.g., Eilermann et al., 2019). This view is also supported by recent experimental evidence. Martinsson and Persson (2019) show that risky, ambiguous, and deterministic decision situations yield very similar behavioral results. Taking a more general perspective, designing counterfactual situations in experiments—in our study, the patient-optimal treatment is known to physicians with certainty—is important to get insight into physicians' potentially suboptimal medical service provision although they know exactly what is best for the patient. This knowledge is typically not available in the field.

Fourth, when choosing parameters of our experimental design like thresholds and bonus levels, we aligned these values with real-world ones within the restrictions of our theory-guided decision environment. For example, performance thresholds and discrete bonus payments are commonly used in the Quality and Outcomes Framework in the UK and in many US HMOs.

Fifth, we exploited the fact that experimental designs are replicable and checked the robustness of our results with regard to the subject pool, to the order in which subjects are exposed to incentives, and to income effects which might result from the introduction of an additional payment.

Finally, the fact that subjects are aware of being part of a study might imply some scrutiny which potentially affects decisions (Levitt and List, 2007, 2009; Czibor et al., 2019). To minimize such potential effects, we implemented a double-blind procedure in our experiment. Even more, our sample was already used to this procedure, as it is a regular part of the data management of the Zi Praxis Panel; for the relevance of this issue see, for example, Barmettler et al. (2012). Nevertheless, the estimation results from our experiment need to be interpreted in light of these potential limitations.

# 6 Conclusion

We studied the effect of performance pay on physicians' health care service provision and the quality of care. To this end, we ran a comprehensive behavioral experiment with a representative German primary care physician sample. At a within-subject level, we implemented a performance pay system with performance thresholds tied to the patient-optimal treatment which complements capitation. Our behavioral results are in line with our theoretical predictions. Under performance pay, physicians increase the quantity of health care services and also increase the quality of health care provision compared to non-blended capitation. However, the intensity of a response to performance pay is significantly increasing in the severity of the patients' illnesses.

In our parsimonious experimental design, we reduced the complexity of a physician's treatment decisions, abstracted from multitasking, considered one-dimensional quality, and refrained from measurement issues of a physician's quality of treatment. We focused on exogenously introducing performance pay while keeping all other variables constant. We incentivized physicians for certain health outcomes—in particular, if a physician's treatment choice either renders the patient's highest health benefit or deviates only by one unit from the patient-optimal treatment—which did not generate uncertainty in physicians' payoffs, as the outcomes of all patients are known.

Our results imply further that high-powered and low-powered incentives attained similar behavioral changes among physicians. An increase in a physician's maximum attainable payoffs by 5% and 20% are similarly effective in inducing a higher quality of care. Also, our behavioral data showed that adjusting performance pay for the patient's severity of illness is reasonable to cope with undertreatment of high-severity patients under capitation. Nonetheless, we observed some evidence for a crowding-out of patient-regarding behavior. This unintended consequence of performance pay incentives referred to in the literature (e.g., Gneezy and Rustichini, 2000) can be observed in our experiment with a representative sample of primary care physicians. While occurring only to a small extent in our experimental frame, the effects should nevertheless be taken seriously: In real-world practice, it could be that rather large shares of physicians provide patient-optimal care in the absence of performance pay. Those patient-regarding physicians might be disposed to crowding out under performance pay if they were given the opportunity. Moreover, as physicians do respond to performance pay, they may capitalize on the information asymmetry between the patient and the health policy-maker on the patient-optimal treatment.

Our behavioral results complement findings from more cumbersome and costly large-scale field interventions and are of particular relevance in light of the scarcity of causal evidence from randomized controlled trials on physician performance pay (e.g., Finkelstein and Taubman, 2015; Newhouse and Normand, 2017). More broadly, our results draw attention to the important challenge of better understanding the impact of the design of incentive schemes and of how individual addressees' characteristics relate to responses to incentives. To this end, an appealing feature of our parsimonious design is that it easily lends itself to further study. Thus, one of our contributions is that we have provided a valuable and easily modifiable design to extend our experimental paradigm for studying further factors that affect physicians' responses to incentives.

## References

Ariely, D., Uri Gneezy, George Loewenstein, and Nina Mazar (2009): "Large Stakes and Big Mistakes," *Review of Economic Studies*, 76, 451–469.

Arrow, K. J. (1963): "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, 53, 941–969.

Baicker, K. and A. Finkelstein (2011): "The Effects of Medicaid Coverage–Learning from the Oregon Experiment," *New England Journal of Medicine*, 365, 683–685.

Baicker, K. and D. Goldman (2011): "Patient Cost-Sharing and Healthcare Spending Growth," *Journal of Economic Perspectives*, 25, 47–68.

Baicker, K., S. L. Taubman, H. L. Allen, M. Bernstein, J. H. Gruber, J. P. Newhouse, E. C. Schneider, B. J. Wright, A. M. Zaslavsky, and A. N. Finkelstein (2013): "The Oregon Experiment — Effects of Medicaid on Clinical Outcomes," *New England Journal of Medicine*, 368, 1713–1722.

Baker, G. P. (1992): "Incentive Contracts and Performance Measurement," *Journal of Political Economy*, 100, 598–614.

Barmettler, F., E. E. Fehr, and C. Zehnder (2012): "Big Experimenter Is Watching You! Anonymity and Prosocial Behavior in the Laboratory," *Games and Economic Behavior*, 75, 17–34.

Bénabou, R. and J. Tirole (2003): "Intrinsic and Extrinsic Motivation," *Review of Economic Studies*, 70, 489–520.

——— (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96, 1652–1678.

Besley, T. and M. Ghatak (2005): "Competition and Incentives with Motivated Agents," *American Economic Review*, 95, 616–636.

Bowles, S. and S. Polania-Reyes (2012): "Economic Incentives and Social Preferences: Substitutes or Complements?" *Journal of Economic Literature*, 50, 368–425.

Brosig-Koch, J., Heike Hennig-Schmidt, Nadja Kairies-Schwarz, and Daniel Wiesen (2016a): "Using Artefactual Field and Lab Experiments to Investigate how Fee-for-Service and Capitation Affect Medical Service Provision," *Journal of Economic Behavior and Organization*, 131, Part B, 17–23.

Brosig-Koch, J., H. Hennig-Schmidt, N. Kairies, and D. Wiesen (2013): "How Effective are Pay-for-Performance Incentives for Physicians? A Laboratory Experiment," Ruhr Economic Papers, No. 413.

Brosig-Koch, J., H. Hennig-Schmidt, N. Kairies-Schwarz, and

D. Wiesen (2016b): "Physician Performance Pay: Evidence from a Laboratory Experiment," Ruhr Economic Papers, No. 658.

——— (2017): "The Effects of Introducing Mixed Payment Systems for Physicians: Experimental Evidence," *Health Economics*, 26, 243–262.

Brot-Goldberg, Z. C., A. Chandra, B. R. Handel, and J. T. Kolstad (2017): "What does a Deductible Do? The Impact of Cost-Sharing on Health Care Prices, Quantities, and Spending Dynamics," *Quarterly Journal of Economics*, 132, 1261–1318.

Buckley, N., K. Cuff, J. Hurley, S. Mestelman, S. Thomas, and D. Cameron (2015): "Support for Public Provision of a Private Good with Top-Up and Opt-Out: A Controlled Laboratory Experiment," *Journal of Economic Behavior and Organization*, 111, 177–196.

——— (2016): "Should I Stay or Should I Go? Exit Options within Mixed ystems of Public and Private Health Care Finance," *Journal of Economic Behavior and Organization*, 131, Part B, 62–77.

Burke, M. A., G. M. Fournier, and K. Prasad (2010): "Geographic Variations in a Model of Physician Treatment Choice with Social Interactions," *Journal of Economic Behavior and Organization*, 73, 418–432.

Campbell, S. M., David Reeves, Evangelos Kontopantelis, Bonnie Sibbald, and Martin Roland (2009): "Effects of Pay for Performance on the Quality of Primary Care in England," *New England Journal of Medicine*, 361, 368–378.

Campbell, S. M., M. Hann, J. Hacker, C. Burns, D. Oliver, A. Thapar, N. Mead, D. G. Safran, and M. O. Roland (2001): "Identifying Predictors of High Quality Care in English General Practice: Observational Study," *British Medical Journal*, 323, 784.

Casalino, L. P., P. Ramsay, L. C. Baker, M. F. Pesko, and S. M.

Shortell (2018): "Medical Group Characteristics and the Cost and Quality of Care for Medicare Beneficiaries," *Health Services Research*, 53, 4970–4996.

Clemens, J. and J. D. Gottlieb (2014): "Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health?" *American Economic Review*, 104, 1320–1349.

Cutler, D. M. (1995): "The Incidence of Adverse Medical Outcomes under Prospective Payment," *Econometrica*, 63, 29–50.

Czibor, E., D. Jimenez-Gomez, and J. A. List (2019): "The Dozen Things Experimental Economists Should Do (More of)," National Bureau of Economic Research No. 25451.

Deci, E. L. and R. M. Ryan (2010): "Intrinsic Motivation," *The Corsini Encyclopedia of Psychology*, 1–2.

Delfgaauw, J. and R. Dur (2008): "Incentives and Workers' Motivations in the Public Sector," *Economic Journal*, 118, 171–191.

Di Guida, S., D. Gyrd-Hansen, and A. S. Oxholm (2019): "Testing the Myth of Fee-for-Service and Overprovision in Health Care," *Health Economics*, 28, 717–722.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011): "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences," *Journal of the European Economic Association*, 9, 522–550.

Donato, K., G. Miller, M. Mohanan, Y. Truskinovsky, and M. Vera-Hernández (2017): "Personality Traits and Performance Contracts: Evidence from a Field Experiment among Maternity Care Providers in India," *American Economic Review*, 107, 506–510.

Doran, T., C. Fullwood, H. Gravelle, D. Reeves, E. Kontopantelis, U. Hiroeh, and M. Roland (2006): "Pay-for-Performance Programs in

Family Practices in the United Kingdom," *New England Journal of Medicine*, 355, 375–384.

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

Dulleck, U., R. Kerschbamer, and M. Sutter (2011): "The Economics of Credence Goods: An Experiment on the Role of Liability, Verifiability, Reputation, and Competition," *American Economic Review*, 101, 526–555.

Eckel, C. C. and P. J. Grossman (1996): "Altruism in Anonymous Dictator Games," *Games and Economic Behavior*, 16, 181–191.

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.

Eilermann, K., Katrin Halstenberg, Ludwig Kuntz, Kyriakos Martakis, Bernhard Roth, and Daniel Wiesen (2019): "The Effect of Expert Feedback on Antibiotic Prescribing in Pediatrics: Experimental Evidence," *Medical Decision Making*, 39, 781–795.

Ellis, R. P. and T. G. McGuire (1986): "Provider Behavior under Prospective Reimbursement: Cost Sharing and Supply," *Journal of Health Economics*, 5, 129–151.

——— (1990): "Optimal Payment Systems for Health Services," *Journal of Health Economics*, 9, 375–396.

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.

European Values Study (2016): "European Values Study: GESIS Datenarchiv, Cologne, Germany. http://dx.doi.org/10.4232/1.12458," .

Falk, A., E. Fehr, and C. Zehnder (2006): "Fairness Perceptions and Reservation Wages—The Behavioral Effects of Minimum Wage Laws," *Quarterly Journal of Economics*, 121, 1347–1381.

Falk, A. and J. J. Heckman (2009): "Lab Experiments Are a Major Source of Knowledge in Social Sciences," *Science*, 326, 535–538.

Fehr, E. and S. Gächter (2000): "Cooperation and Punishment in Public Goods Experiments," *American Economic Review*, 90, 980–994.

——— (2002): "Altruistic Punishment in Humans," *Nature*, 415, 137.

Finkelstein, A. and S. Taubman (2015): "Randomize Evaluations to Improve Health Care Delivery," *Science*, 347, 720–722.

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker, and Oregon Health Study Group (2012): "The Oregon Health Insurance Experiment: Evidence from the First Year," *Quarterly Journal of Economics*, 127, 1057–1106.

Frakt, A. B. and A. K. Jha (2018): "Face the Facts: We Need to Change the Way We Do Pay for Performance," *Annals of Internal Medicine*, 168, 291–292.

Francois, P. (2000): "'Public Service Motivation' as an Argument for Government Provision," *Journal of Public Economics*, 78, 275–299.

Galizzi, M. M. and D. Wiesen (2018): "Behavioral Experiments in Health," in *Oxford Research Encyclopedia of Economics and Finance*, ed. by J. Hamilton, Oxford University Press, Oxford, UK.

Gaynor, M. and P. Gertler (1995): "Moral Hazard and Risk Spreading in Partnerships," *RAND Journal of Economics*, 26, 591–613.

Gaynor, M., J. B. Rebitzer, and L. J. Taylor (2004): "Physician Incentives in Health Maintenance Organizations," *Journal of Political Economy*, 112, 915–931.

Glasziou, P. P., H. Buchan, C. Del Mar, J. Doust, M. Harris,

R. Knight, A. Scott, I. A. Scott, and A. Stockwell (2012): “When Financial Incentives Do More Good than Harm: A Checklist,” *British Medical Journal*, 345.

Glazer, J. and T. G. McGuire (2000): “Optimal Risk Adjustment of Health Insurance Premiums: An Application to Managed Care,” *American Economic Review*, 90, 1055–1071.

Gneezy, U., S. Meier, and P. Rey-Biel (2011): “When and Why Incentives (Don’t) Work to Modify Behavior,” *Journal of Economic Perspectives*, 25, 191–210.

Gneezy, U. and A. Rustichini (2000): “Pay Enough or Don’t Pay at All,” *Quarterly Journal of Economics*, 115, 791–810.

Godager, G., H. Hennig-Schmidt, and T. Iversen (2016): “Does Performance Disclosure Influence Physicians’ Medical Decisions? An Experimental Study,” *Journal of Economic Behavior and Organization*, 131, Part B, 36–46.

Godager, G. and D. Wiesen (2013): “Profit or Patients’ Health Benefit? Exploring the Heterogeneity in Physician Altruism,” *Journal of Health Economics*, 32, 1105–1116.

Gravelle, H., M. Sutton, and A. Ma (2010): “Doctor Behaviour under a Pay for Performance Contract: Treating, Cheating and Case Finding?” *Economic Journal*, 120, 129–156.

Harrison, G. W. and J. A. List (2004): “Field Experiments,” *Journal of Economic Literature*, 42, 1009–1055.

Heckman, J. J. (2010): “Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy,” *Journal of Economic Literature*, 48, 356–398.

Hennig-Schmidt, H., R. Selten, and D. Wiesen (2011): “How Payment Systems Affect Physicians’ Provision Behavior – An Experimental Investigation,” *Journal of Health Economics*, 30, 637–646.

Hennig-Schmidt, H. and D. Wiesen (2014): “Other-regarding Behavior and Motivation in Health Care Provision: An Experiment with Medical and Non-medical Students,” *Social Science and Medicine*, 108, 156–165.

Herrmann, B., C. Thöni, and S. Gächter (2008): “Antisocial Punishment Across Societies,” *Science*, 319, 1362–1367.

Heyman, J. and D. Ariely (2004): “Effort for Payment: A Tale of Two Markets,” *Psychological Science*, 15, 787–793.

Huffman, D. and M. Bognanno (2018): “High-Powered Performance Pay and Crowding Out of Nonmonetary Motives,” *Management Science*, 64, 4669–4680.

Keeney, R. L. and H. Raiffa (1976): *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, J. Wiley, New York.

Kesternich, I., H. Schumacher, and J. Winter (2015): “Professional Norms and Physician Behavior: Homo Oeconomicus or Homo Hippocraticus?” *Journal of Public Economics*, 131, 1–11.

Kolstad, J. T. (2013): “Information and Quality When Motivation Is Intrinsic: Evidence from Surgeon Report Cards,” *American Economic Review*, 103, 2875–2910.

Kralewski, J., Dowd Bryan, Knutson David, Tong Junliang, and Savage Megan (2015): “The Relationships of Physician Practice Characteristics to Quality of Care and Costs,” *Health Services Research*, 50, 710–729.

Kreps, D. M. (1997): “Intrinsic Motivation and Extrinsic Incentives,” *The American Economic Review*, 87, 359–364.

Kristensen, S. R., R. Meacock, A. J. Turner, R. Boaden, R. McDonald, M. Roland, and M. Sutton (2014): “Long-term Effect of Hospital Pay for Performance on Mortality in England,” *New England Journal of Medicine*, 371, 540–548.

Kristensen, S. R., L. Siciliani, and M. Sutton (2016): "Optimal Price-setting in Pay for Performance Schemes in Health Care," *Journal of Economic Behavior and Organization*, 123, 57–77.

Lagarde, M. and D. Blaauw (2017): "Physicians' Responses to Financial and Social Incentives: A Medically Framed Real Effort Experiment," *Social Science and Medicine*, 179, 147–159.

Lazear, E. P. (2000): "Performance Pay and Productivity," *American Economic Review*, 90, 1346–1361.

Levitt, S. D. and J. A. List (2007): "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives*, 21, 153–174.

——— (2009): "Field Experiments in Economics: The Past, the Present, and the Future," *European Economic Review*, 53, 1–18.

Li, J. (2018): "Plastic Surgery or Primary Care? Altruistic Preferences and Expected Specialty Choice of U.S. Medical Students," *Journal of Health Economics*, 62, 45–59.

Li, J., J. Hurley, P. DeCicca, and G. Buckley (2014): "Physician Response to Pay-for-Performance–Evidence from a Natural Experiment," *Health Economics*, 23, 962–978.

Lindenauer, P. K., D. Remus, S. Roman, M. B. Rothberg, E. M. Benjamin, A. Ma, and D. W. Bratzler (2007): "Public Reporting and Pay for Performance in Hospital Quality Improvement," *New England Journal of Medicine*, 356, 486–496.

List, J. A. (2011): "Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off," *Journal of Economic Perspectives*, 25, 3–16.

Ma, C.-t. A. and T. G. McGuire (1997): "Optimal Health Insurance and Provider Payment," *American Economic Review*, 87, 685–704.

Manning, W. G., J. P. Newhouse, N. Duan, E. B. Keeler, and A. Leibowitz (1987): "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment," *American Economic Review*, 251–277.

Martinsson, P. and E. Persson (2019): "Physician Behavior and Conditional Altruism: The Effects of Payment System and Uncertain Health Benefit," *Theory and Decision*, 1–23.

Maynard, A. (2012): "The Powers and Pitfalls of Payment for Performance," *Health Economics*, 21, 3–12.

McGuire, T. G. (2000): "Physician Agency," in *Handbook of Health Economics, Vol. 1 A*, ed. by Cuyler and Newhouse, North-Holland, Amsterdam (The Netherlands), 461–536.

Mellström, C. and M. Johannesson (2008): "Crowding out in Blood Donation: Was Titmuss right?" *Journal of the European Economic Association*, 64, 845–863.

Mendelson, A., K. Kondo, C. Damberg, A. Low, M. Motuapuaka, M. Freeman, M. O'Neil, R. Relevo, and D. Kansagara (2017): "The Effects of Pay-for-Performance Programs on Health, Health Care Use, and Processes of Care: A Systematic Review," *Annals of Internal Medicine*, 166, 341–353.

Miller, G. and K. S. Babiarz (2014): "Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs," in *Encyclopedia of Health Economics*, ed. by Anthony J. Culyer, Elsevier, 457 – 466.

Mullen, K. J., R. G. Frank, and M. B. Rosenthal (2010): "Can You Get What You Pay For? Pay-For-Performance and the Quality of Healthcare Providers," *RAND Journal of Economics*, 41, 64–91.

Newhouse, J. P. and S.-L. T. Normand (2017): "Health Policy Trials," *New England Journal of Medicine*, 376, 2160–2167.

Newhouse, J. P. and the Insurance Experiment Group (1993): *Free for*

*All: Lessons from the RAND Health Insurance Experiment*, Harvard University Press.

Ng, C. W. L. and K. P. Ng (2013): "Does Practice Size Matter? Review of Effects on Quality of Care in Primary Care," *British Journal of General Practice*, 63, e604–e610.

Pellegrino, E. D. (1987): "Altruism, Self-interest, and Medical Ethics," *Journal of the American Medical Association*, 258, 1939–1940.

Prendergast, C. (1999): "The Provision of Incentives in Firms," *Journal of Economic Literature*, 37, 7–63.

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.

——— (2012): "Pay-for-Performance: Not a Magic Bullet," *Annals of Internal Medicine*, 157, 912–913.

Roland, M. and S. Campbell (2014): "Successes and Failures of Pay for Performance in the United Kingdom," *New England Journal of Medicine*, 370, 1944–1949.

Rosenthal, M. B., R. G. Frank, Z. Li, and A. M. Epstein (2005): "Early Experience with Pay-for-Performance: From Concept to Practice," *Journal of the American Medical Association*, 294, 1788–1793.

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

Scott, A., P. Sivey, D. A. Ouakrim, L. Willenberg, L. Naccarella, J. Furler, and D. Young (2011): "The Effect of Financial Incentives on the Quality of Health Care Provided by Primary Care Physicians," *Cochrane Database of Systematic Reviews*, https://doi.org/10.1002/14651858.CD008451.pub2.

Siciliani, L. (2009): "Paying for Performance and Motivation Crowding Out," *Economics Letters*, 103, 68–71.

van de Ven, W. P. and R. P. Ellis (2000): "Risk Adjustment in Competitive Health Plan Markets," in *Handbook of Health Economics*, ed. by Culyer, Anthony J. and Newhouse, Joseph P. , Elsevier, vol. 1, Part A, 755 – 845.

Wang, J., T. Iversen, H. Hennig-Schmidt, and G. Godager (2017): "How Changes in Payment Schemes Influence Provision Behavior," University of Oslo, Health Economics Research Programme, HERO Working paper No. 2017:2.

——— (forthcoming): "Are Patient-Regarding Preferences Stable? Evidence from a Laboratory Experiment with Physicians and Medical Students from Different Countries," *European Economic Review.*

Woolhandler, S., D. Ariely, and D. U. Himmelstein (2012): "Why Pay for Performance May Be Incompatible with Quality Improvement," *British Medical Journal*, 345, e5015.

# Appendices

## A Further background, information on the experiment, and variables

### A.1 Statutory health insurance in Germany

In Germany, health insurance is mandatory for all citizens and permanent residents. Health insurance is offered by two systems: (i) non-governmental health insurance funds (sickness funds) in the SHI system and (ii) private health insurance (PHI). Sickness funds are financed by compulsory payroll-deducted contributions of employees as a percentage of their gross income (14.6% in 2016). The vast majority of health care is provided under the SHI scheme: about 87% of the German population (i.e., 73 million people) were enrolled in SHI and about 11% in PHI. Sickness funds are represented by the National Association of Statutory Health Insurance Funds (*GKV-Spitzenverband*), the central representation of the health insurance funds at federal level. Its key responsibility is the determination of payments for health care services.

Overall, about 169,900 primary care physicians and physicians from other specialists in ambulatory care contracted with SHI in 2016.[24] By law, they are mandatory members in one of the 17 regional associations of SHI physicians (*Kassenärztliche Vereinigung*, KV). The regional KVs act as financial intermediaries between the sickness funds and the physicians in ambulatory care, who typically are self-employed owners of private practices. KVs and statutory sickness funds enter into collective agreements on reimbursement for health care services. Sickness funds pool the designated funds into a joint budget, which is then distributed by the regional KVs.

[24] For more details, see *KBV 2017-Stat, Kassenärztliche Bundesvereinigung (2017). Statistische Informationen aus dem Bundesarztregister—Bundesgebiet insgesamt, as of December 31st, 2017, http://www.kbv.de/media/sp/2017_12_31_BAR_Statistik.pdf.*

## A.2 Reimbursement of primary care physicians

The reimbursement of primary care physicians contracting with the SHI contain main elements of lump-sum capitation (CAP). In particular, medical service provision is reimbursed according to a standardized fee schedule (*Einheitlicher Bewertungsmaßstab*). Sickness funds pay an overall morbidity-adjusted capped budget to the regional KVs. Each primary care physician bills his or her regional KV on a case-to-case basis according to the volume of provided health care services (rather than directly charging the sickness funds).

The payment a primary care physician receives for medical services provided to a patient in the SHI (*Regelleistungsvolumen*, RLV) is given by: *number of cases/patients* × *case value for primary care physicians (in EUR)* × *age-based risk-adjustment weight.*[25] That means German primary care physicians face a capped budget (RLV) based on the number of patients treated in a baseline period. One could therefore regard the RLV as a capitation system, as the overall payment is capped, leaving the physician with a payment per patient enrolled with his or her practice. Primary care physicians are familiar with capitation, as RLV limits their service provision for the total number of patients enrolled in their practices. We thus implement capitation as the baseline payment in our experiment; for more details, see Section 1.

[25] To curtail excessive provision of standard services according to the SHI reimbursement scheme, the following reduction of case values are applied: If a primary care physician's total claims are higher than his/her RLV and his/her number of cases exceeds the average number of cases for the group of primary care physicians by more than 50%, case values are reduced by 25% to 80% of full reimbursement values.

## A.3 Parameters of the experiment

Table A.1: Parameters

| | Quantity ($q$) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Patient benefit | | | | | | | | | | | |
| $B_{Ax}$ | 10 | 12.5 | 15 | 17.5 | 15 | 12.5 | 10 | 7.5 | 5 | 2.5 | 0 |
| $B_{Ay}$ | 5 | 7.5 | 10 | 12.5 | 15 | 17.5 | 15 | 12.5 | 10 | 7.5 | 5 |
| $B_{Az}$ | 0 | 2.5 | 5 | 7.5 | 10 | 12.5 | 15 | 17.5 | 15 | 12.5 | 10 |
| $B_{Bx}$ | 17.5 | 20 | 22.5 | 25 | 22.5 | 20 | 17.5 | 15 | 12.5 | 10 | 7.5 |
| $B_{By}$ | 12.5 | 15 | 17.5 | 20 | 22.5 | 25 | 22.5 | 20 | 17.5 | 15 | 12.5 |
| $B_{Bz}$ | 7.5 | 10 | 12.5 | 15 | 17.5 | 20 | 22.5 | 25 | 22.5 | 20 | 17.5 |
| $B_{Cx}$ | 20 | 25 | 30 | 35 | 30 | 25 | 20 | 15 | 10 | 5 | 0 |
| $B_{Cy}$ | 10 | 15 | 20 | 25 | 30 | 35 | 30 | 25 | 20 | 15 | 10 |
| $B_{Cz}$ | 0 | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 30 | 25 | 20 |
| Costs | | | | | | | | | | | |
| $c$ | 0.0 | 0.25 | 1 | 2.25 | 4 | 6.25 | 9 | 12.25 | 16 | 20.25 | 25 |
| **Capitation (CAP)** | | | | | | | | | | | |
| CAP | | | | | | | | | | | |
| $\Lambda$ | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 |
| $\pi$ | 25 | 24.75 | 24 | 22.75 | 21 | 18.75 | 16 | 12.75 | 9 | 4.75 | 0 |
| **Performance pay (CAP+P4P)** | | | | | | | | | | | |
| Low-bonus condition (5%) | | | | | | | | | | | |
| $\Lambda$ | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 |
| $b_x$ | 0 | 0 | 2.25 | 2.25 | 2.25 | 0 | 0 | 0 | 0 | 0 | 0 |
| $b_y$ | 0 | 0 | 0 | 0 | 5.25 | 5.25 | 5.25 | 0 | 0 | 0 | 0 |
| $b_z$ | 0 | 0 | 0 | 0 | 0 | 0 | 10.25 | 10.25 | 10.25 | 0 | 0 |
| $\pi_x$ | 25 | 24.75 | 26.25 | 25 | 23.25 | 18.75 | 16 | 12.75 | 9 | 4.75 | 0 |
| $\pi_y$ | 25 | 24.75 | 24 | 22.75 | 26.25 | 24 | 21.25 | 12.75 | 9 | 4.75 | 0 |
| $\pi_z$ | 25 | 24.75 | 24 | 22.75 | 21 | 18.75 | 26.25 | 23 | 19.25 | 4.75 | 0 |
| High-bonus condition (20%) | | | | | | | | | | | |
| $\Lambda$ | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 |
| $b_x$ | 0 | 0 | 6 | 6 | 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| $b_y$ | 0 | 0 | 0 | 0 | 9 | 9 | 9 | 0 | 0 | 0 | 0 |
| $b_z$ | 0 | 0 | 0 | 0 | 0 | 0 | 14 | 14 | 14 | 0 | 0 |
| $\pi_x$ | 25 | 24.75 | 30 | 28.75 | 27 | 18.75 | 26.25 | 23 | 19.25 | 4.75 | 0 |
| $\pi_y$ | 25 | 24.75 | 24 | 22.75 | 30 | 27.75 | 25 | 12.75 | 9 | 4.75 | 0 |
| $\pi_z$ | 25 | 24.75 | 24 | 22.75 | 21 | 18.75 | 30 | 26.75 | 23 | 4.75 | 0 |

*Notes:* This table shows the parameters used in our experiment for all payment conditions. $B$ denotes the patient's health benefit for all combinations of illnesses $A, B$, and $C$, and the severities of illnesses ($x, y$, and $z$). $\Lambda$ is the lump-sum payment in CAP, $b_l^{\bullet}$ is the bonus paid when the quality requirement is met in CAP+P4P, and $\pi$ is the physician's profit. For the control conditions with medical students, the parameter values are multiplied by 0.32.

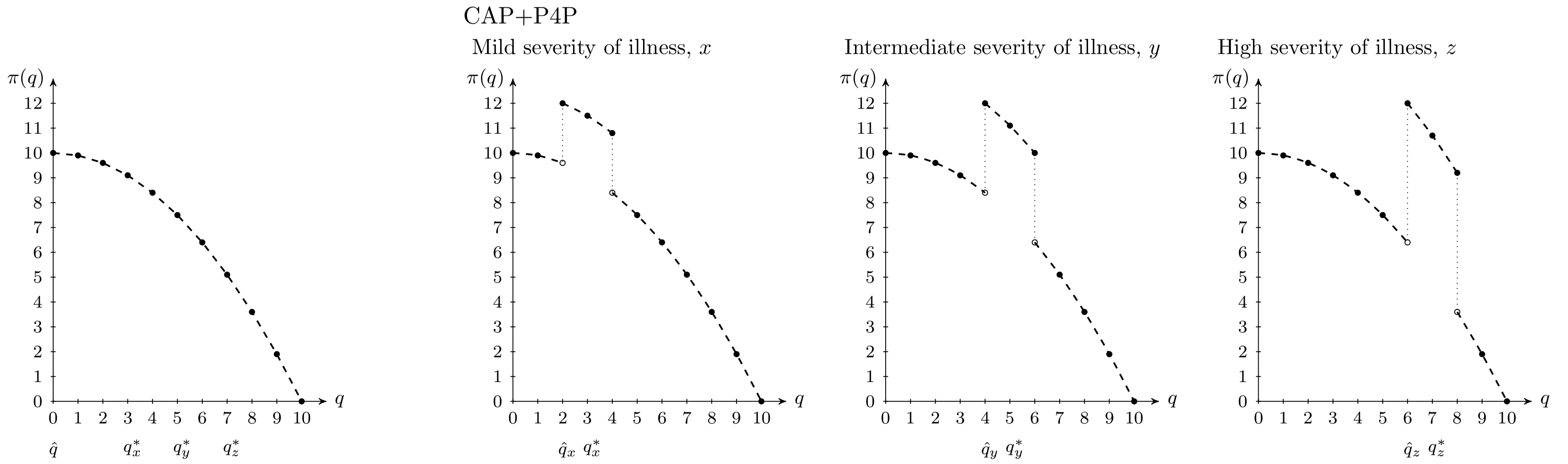


Figure A.1: Illustration of profit parameters under capitation (CAP) and performance pay (CAP+P4P)

*Notes:* This figure illustrates profits in CAP and CAP+P4P. Notice that, for the main experiments, profit parameters in the figure need to be multiplied by 2.5, and, for the control conditions by 0.8 to reflect actual Euro values. Under basic payments, profits decrease in CAP monotonically, 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 $\hat{q}$=0 in CAP, respectively. In the pay for performance condition, $\hat{q}$ changes depending on the severity of illness.

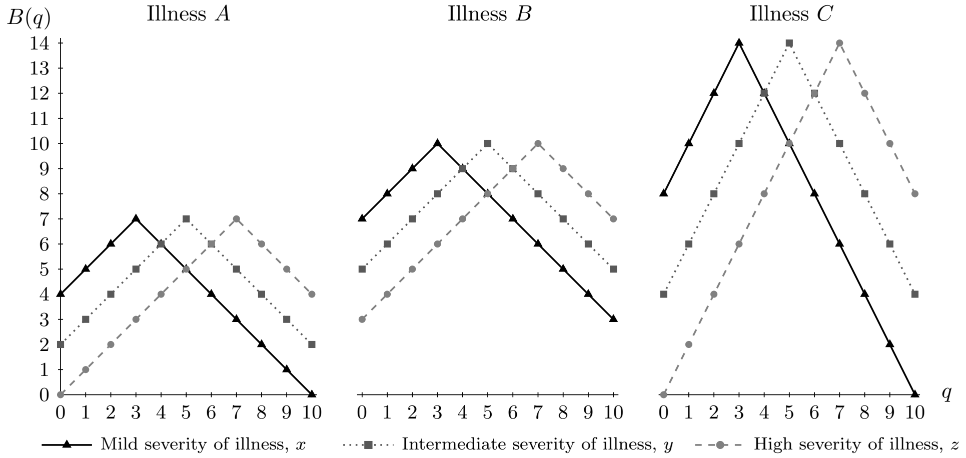
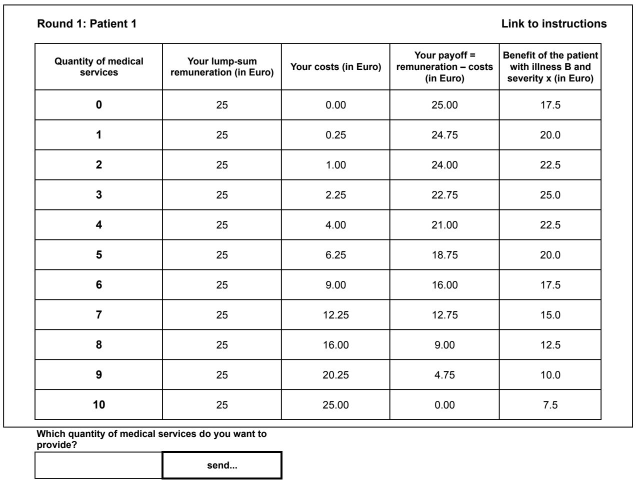


Figure A.2: Illustration of patient health benefits by illness and severity of illness

*Notes:* This figure illustrates patient health benefit parameters $B(q)$ for illnesses $k = A, B, C$ and severities of illness $l = x, y, z$ on the quantity interval from 0 to 10. Notice that, for the main experiments, benefit parameters in the figure need to be multiplied by 2.5, and, for the control conditions by 0.8 to reflect actual Euro values. 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: Sample decision situation in the Low-bonus condition

CAP:

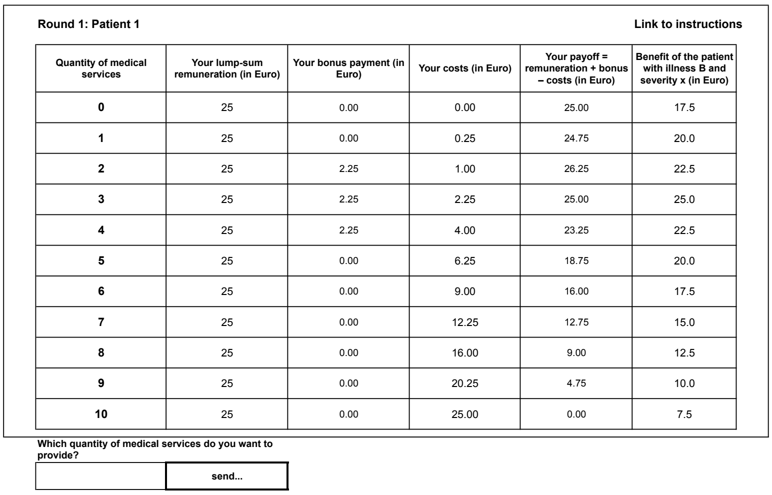


Round 1: Patient 1 Link to instructions

| Quantity of medical services | Your lump-sum remuneration (in Euro) | Your costs (in Euro) | Your payoff = remuneration – costs (in Euro) | Benefit of the patient with illness B and severity x (in Euro) |
|---|---|---|---|---|
| **0** | 25 | 0.00 | 25.00 | 17.5 |
| **1** | 25 | 0.25 | 24.75 | 20.0 |
| **2** | 25 | 1.00 | 24.00 | 22.5 |
| **3** | 25 | 2.25 | 22.75 | 25.0 |
| **4** | 25 | 4.00 | 21.00 | 22.5 |
| **5** | 25 | 6.25 | 18.75 | 20.0 |
| **6** | 25 | 9.00 | 16.00 | 17.5 |
| **7** | 25 | 12.25 | 12.75 | 15.0 |
| **8** | 25 | 16.00 | 9.00 | 12.5 |
| **9** | 25 | 20.25 | 4.75 | 10.0 |
| **10** | 25 | 25.00 | 0.00 | 7.5 |

Which quantity of medical services do you want to provide?

CAP+P4P:



Round 1: Patient 1 Link to instructions

| Quantity of medical services | Your lump-sum remuneration (in Euro) | Your bonus payment (in Euro) | Your costs (in Euro) | Your payoff = remuneration + bonus – costs (in Euro) | Benefit of the patient with illness B and severity x (in Euro) |
|---|---|---|---|---|---|
| **0** | 25 | 0.00 | 0.00 | 25.00 | 17.5 |
| **1** | 25 | 0.00 | 0.25 | 24.75 | 20.0 |
| **2** | 25 | 2.25 | 1.00 | 26.25 | 22.5 |
| **3** | 25 | 2.25 | 2.25 | 25.00 | 25.0 |
| **4** | 25 | 2.25 | 4.00 | 23.25 | 22.5 |
| **5** | 25 | 0.00 | 6.25 | 18.75 | 20.0 |
| **6** | 25 | 0.00 | 9.00 | 16.00 | 17.5 |
| **7** | 25 | 0.00 | 12.25 | 12.75 | 15.0 |
| **8** | 25 | 0.00 | 16.00 | 9.00 | 12.5 |
| **9** | 25 | 0.00 | 20.25 | 4.75 | 10.0 |
| **10** | 25 | 0.00 | 25.00 | 0.00 | 7.5 |

Which quantity of medical services do you want to provide?

### A.4 Control experiments

In our control experiments, we check for the robustness of our results towards (i) order effects, (ii) income effects, and (iii) subject pool effects. First, it could matter if participants are exposed to incentives in CAP first and in CAP+P4P second or vice versa. While the main purpose of our experiment is testing the introduction of performance pay, a test of order effects allows us to explore whether taking it away affects behavior. It has been argued that the order of payment systems might influence decisions. Yet, we do not know of any study that analyzes this question in the health domain, and the experimental evidence is not clear-cut even in health settings.[26] To analyze order effects in our experiment, in condition 'C–High bonus (20%)–First', we conducted CAP+P4P-20% in part one of the experiment, followed by CAP in part two of the experiment; see Panel B of Table 1 for all control conditions.

The second control condition is related to the fact that in CAP+P4P the maximum attainable profit is higher than in CAP. This potential income effect may lead to more pronounced behavioral responses under performance pay. To check whether this effect does exist, we conducted a control condition C–CAP–High. We raised the lump-sum reimbursement in CAP by 20%, while keeping the CAP+P4P-20% payment constant in the second part of the experiment.

Third, we analyze whether primary care physicians and medical students differ in their behavior. This is important as we conducted all control conditions with medical students enrolled at the University of Cologne, Germany. In order to be sure that the above control conditions can be used as a valid check for the robustness of the results of our main conditions, we had to assess that the behavior of the medical students did not differ significantly from that of primary care physicians. To this end, we conducted the experiments C–Low–bonus (5%) and C–High–bonus (20%), using the same experimental parameters as for the

[26] Supporting evidence for the absence of order effects is reported in health contexts (e.g., Buckley et al., 2015, 2016) and in public good games (e.g., Fehr and Gächter, 2000, 2002; Herrmann et al., 2008). On the contrary, evidence for an order effect is reported in experiments in a medical frame (e.g., Wang et al., 2017) and, for example, when introducing or removing a minimum wage (e.g., Falk et al., 2006).

primary care physicians; see Table A.1 in Appendix A.3. The only difference is that we adjusted the payment for the medical students to one third of that for the physicians to have adequate financial incentives reflecting typical hourly wage levels.

## A.5 Instructions of the experiment

Notice that the text in squared brackets [] denotes text on the computer screen not contained in the instructions.

[*Text on Computer Screen*: **Welcome to the Experiment!** ]

Thank you very much for your participation. During the study, you will be asked to make decisions for which you will receive an allowance. This allowance we call payoff in the following. Your payoff depends on the decisions you make. At the end of the study, your total payoff will be transferred to you by the notary of the Zentralinstitut für die kassenärztliche Versorgung in Deutschland. Thereby, anonymity of your decisions is guaranteed.

The experiment will take about 30 minutes and consists of two parts. Before each part, you will receive detailed instructions that you can download during the respective part of the study using the 'Link to Instructions'. If possible, please print the instructions for your assistance before the respective part of the study starts.

Pls. note that neither your decisions in part *I* nor in part *II* will have any influence on the respective other part of the study. The study ends by a small questionnaire. Pls. klick OK to proceed to the instructions of part *I* of the study. ]

### Instructions to part *I*

In part *I* of the study you will participate in nine decision rounds.

*Description of decision rounds*

In each round, you decide as a physician on the medical treatment for a patient. That means, in each round you have to determine the quantity of medical services you wish to provide to this 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 of the nine decision rounds, you will consecutively and in random order face one patient who is characterized by one of the nine possible combinations of illnesses and degrees of severity. Each of these nine patients you can provide with a quantity of $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$, or 10 medical services. Providing the medical treatment for each patient is independent from that for the other patients.

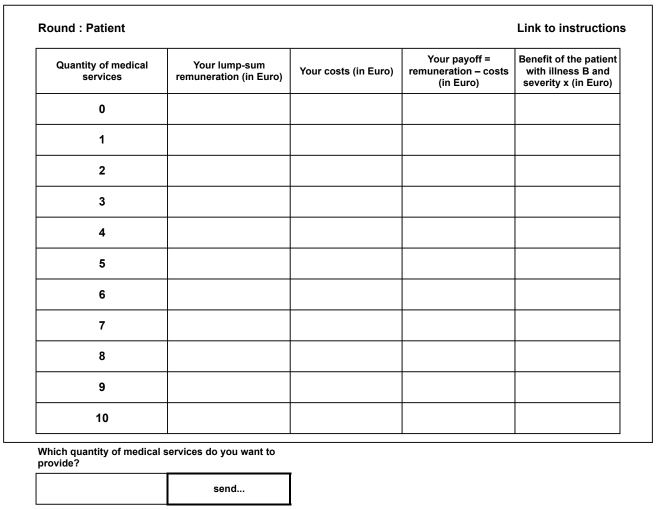
*Payoff*

In each round, you receive a lump-sum remuneration for treating the patient irrespective of the amount of medical treatment you provide. You also incur costs for treating the patient, which depend on the quantity of services you provide. Your payoff in each decision round is calculated by subtracting these costs from the lump-sum remuneration for treating the patient. Your remuneration, your costs and your payoff will be stated in Euro.

Each quantity of medical services yields a particular health status—contingent on illness and severity—, i.e., a particular benefit for the patient. Hence, in choosing the medical services you provide, you determine not only your own payoff but also the patient's benefit. The benefit is stated in monetary units (Euro).

Before taking your decision, in each round you will be shown on your screen the illness ($A$, $B$, or $C$), the severity of the illness (x, y or z), and—for each possible amount of medical treatment—your lump-sum remuneration, your costs, your payoff, as well as the benefit for the patient. You, therefore, need not calculate these values yourself.

*Payment*



Round : Patient

Link to instructions

| Quantity of medical services | Your lump-sum remuneration (in Euro) | Your costs (in Euro) | Your payoff = remuneration – costs (in Euro) | Benefit of the patient with illness B and severity x (in Euro) |
|---|---|---|---|---|
| 0 | | | | |
| 1 | | | | |
| 2 | | | | |
| 3 | | | | |
| 4 | | | | |
| 5 | | | | |
| 6 | | | | |
| 7 | | | | |
| 8 | | | | |
| 9 | | | | |
| 10 | | | | |

Which quantity of medical services do you want to provide?

At the end of the study, one of the nine rounds of this part of the study will be chosen at random. Your payoff in that round together with your payoff from part *II* of the study will be transferred to you by the notary of the Zentralinstitut für die kassenärztliche Versorgung in Deutschland.

The benefit (in Euro) that a patient gets from your medical treatment in the chosen round, will be beneficial for a real patient. The amount will be transferred to the Christoffel Blindenmission Deutschland e.V., 64625 Bensheim, which will use the money exclusively for enabling the treatment of patients with eye cataract. Transferring the money to the Christoffel Blindenmission Deutschland e.V. will also be carried by the notary of the Zentralinstitut für die kassenärztliche Versorgung in Deutschland.

[*Text on Computer Screen*: In the following, you are kindly asked to answer some comprehension questions. Pls. note, that the comprehension questions are not meant to recommend taking a specific decision in the study to follow. The questions are only intended to improve and sharpen your understanding of the decision situation you will be facing in the study.]

*Comprehension questions*

Prior to the decision rounds, we kindly ask you to answer a few comprehension questions. They are intended to help familiarize yourself with the decision situation. Having answered all questions correctly, part $I$ of the study will begin immediately. Otherwise you are asked to answer the respective question again.

## Instructions to part $II$

In part $II$ of the study you will again participate in nine decision rounds.

*Description of decision rounds*

As in part $I$ of the study, in each round, you decide as a physician on the medical treatment for a patient. That means, you have to determine in each round the quantity of medical services you wish to provide to this patient for a given illness and a given severity of this illness.

As in part $I$, you will in the nine decision rounds consecutively and in random order face one patient who is characterized by one of the nine patients who is characterized by one of the three illnesses $(A, B, C)$, and by one of the three different degrees of severity $(x, y, z)$. Each of these nine patients you can provide with a quantity of $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$, or 10 medical services. Providing the medical treatment for each patient is independent from that for the other patients.

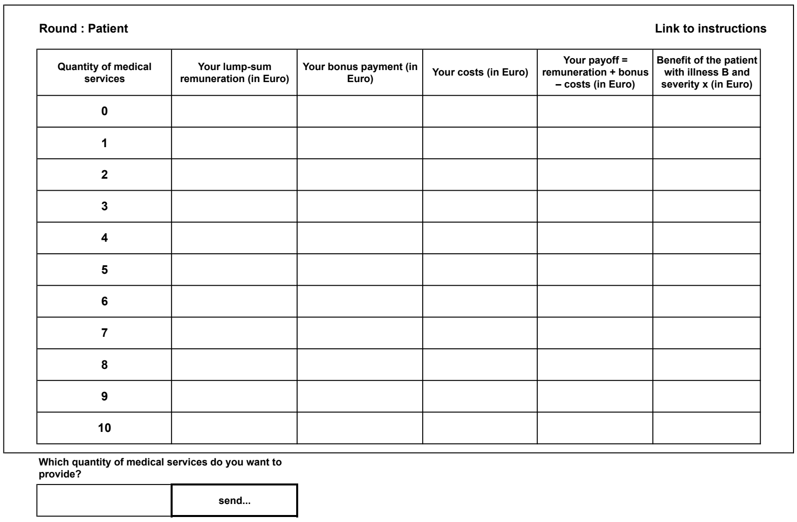
*Payoff*

In each round, you receive a lump-sum remuneration for treating the patient 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 depend on the quantity of services you provide. **Your payoff in each decision round is**

**calculated by the sum of the lump-sum remuneration and the bonus payment minus the costs from treating the patient.** Your lump-sum remuneration, your costs, your bonus payment and your payoff will be stated in Euro.

As in part $I$, each quantity of medical service yields a particular health status—contingent on illness and severity—, i.e., a particular benefit for the patient. Hence, in choosing the medical services you provide, you determine not only your own payoff but also the patient's benefit. The benefit is stated in monetary units (Euro).

**Before taking your decision, in each round you will be shown on your screen the illness ($A, B$, or $C$), the severity of the illness (x, y or z), and—for each possible amount of medical treatment—the amounts of your lump-sum remuneration and the bonus payment, your costs, your payoff, as well as the benefit for the patient.** You, therefore, need not calculate these values yourself.



Round : Patient

Link to instructions

| Quantity of medical services | Your lump-sum remuneration (in Euro) | Your bonus payment (in Euro) | Your costs (in Euro) | Your payoff = remuneration + bonus – costs (in Euro) | Benefit of the patient with illness B and severity x (in Euro) |
|---|---|---|---|---|---|
| 0 | | | | | |
| 1 | | | | | |
| 2 | | | | | |
| 3 | | | | | |
| 4 | | | | | |
| 5 | | | | | |
| 6 | | | | | |
| 7 | | | | | |
| 8 | | | | | |
| 9 | | | | | |
| 10 | | | | | |

Which quantity of medical services do you want to provide?

*Payment*

At the end of the study, one of the nine rounds of this part of the study will be chosen at random. Your payoff in that round together with your payoff from part *I* of the study will be transferred to you by the notary of the Zentralinstitut für die kassenärztliche Versorgung in Deutschland.

As in part *I* of the study, the benefit (in Euro) that a patient gets from your medical treatment in the chosen round, will be beneficial for a real patient. The amount together with the amount from part *I* will be transferred by the notary to the Christoffel Blindenmission Deutschland e.V., 64625 Bensheim, which will use the money exclusively for enabling the treatment of patients with eye cataract.

[*Text on Computer Screen*: In the following, you are kindly asked to answer some comprehension questions. Pls. note, that the comprehension questions are not meant to recommend taking a specific decision in the study to follow. The questions are only intended to improve and sharpen your understanding of the decision situation you will be facing in the study.]

*Comprehension questions*

Prior to the decision rounds, we again kindly ask you to answer a few comprehension questions. They are intended to help familiarize yourself with the decision situation. Having answered all questions correctly, part *II* of the study will begin immediately. Otherwise you are asked to answer the respective question again.

## A.6 Comprehension questions

*The questions were asked for different benefit functions and for both CAP and CAP+P4P.*

- Assume you want to provide for the patient shown in the table the quantity of services that yields the **lowest benefit** for this patient. Which quantity of medical services you have to choose?
- Assume you want to provide for the patient shown in the table the quantity of services that yields the **highest payoff** for you. Which quantity of medical services you have to choose?
- Assume you want to provide for the patient shown in the table the quantity of services that yields the **highest benefit** for this patient. Which quantity of medical services you have to choose?
- Assume you want to provide for the patient shown in the table the quantity of services that yields the **lowest payoff** for you. Which quantity of medical services you have to choose?

## A.7 Description of variables

Table A.2: Description of all physician variables

| Variable | Description |
|---|---|
| *Physician practice characteristics* | |
| Annual profit | Annual profit per practice owner before tax in 2014. Data are given as median spilt based on all general practitioners in the ZiPP. Source: Zi-Praxis-Panel, wave 2015. |
| Time for treatment SHI patients | Time for treatment of patients with statutory health insurance in 2014, given as proportionate share relative to patients with private insurance. Data are given in 5 percentiles with class limits based on all general practitioners in the ZiPP. Source: Zi-Praxis-Panel, wave 2015. |
| Revenue from treating SHI patients | Revenue for treatment of patients with statutory health insurance in 2014, given as proportionate share relative to patients with private insurance. Data are given in 5 percentiles with class limits based on all general practitioners in the ZiPP. Source: Zi-Praxis-Panel, wave 2015. |
| Number of SHI patients | Number of patients with statutory health insurance in 2014, given as proportionate share relative to patients private insurance. Data are given in 5 percentiles with class limits based on all general practitioners in the ZiPP. Source: Zi-Praxis-Panel, wave 2015. |
| Number of physicians in practice | Number of physicians in the practice $[1, 2, 3$ or more$]$ in 2014. Source: Zi-Praxis-Panel, wave 2015. |
| Location of practice | Location of practice classified by density of inhabitants: city ($>$ 300 inhabitants/km$^2$), outer conurbation ($\leq$ 100 inhabitants/km$^2$), rural ($<$ 100 inhabitants/km$^2$). Source: Bundesinstitut für Bau-, Stadt-, und Raumforschung. |
| *Physician characteristics* | |
| Age | Age of the physician. Source: own questionnaire |
| Gender | Sex of physician. Source: Zi-Praxis-Panel, wave 2015. |
| Practice years | Number of years practicing as employee, in own practice, in hospital. Source: own questionnaire. |
| Risk attitudes | Self-assessed risk attitudes on a scale from 1 (not at all willing to take risks) to 10 (very willing to take risks). Risk attitudes are given as general risk attitudes, risk attitudes regarding own health, and risk attitudes regarding health of patients. Source: own questionnaire; question is based on Socio-economic Panel (Dohmen et al., 2011). |
| Altruism attitudes | On a scale from 1 to 10: [1] "Most of the time people are mostly just looking out for themselves." [10] "Most of the time people try to be helpful." Source: own questionnaire. The question is based on the European Value Study (European Values Study, 2016). |
| Competition attitudes | On a scale from 1 to 10: [1] "Competition is harmful. It brings out the worst in the people." [10] "Competition is good. It stimulates people to work hard and develop new ideas." Source: own questionnaire. The question is based on the European values study (European Values Study, 2016). |

# B Additional analyses of the experimental data

Table B.1: Comparison between health care provision and patient-optimal care

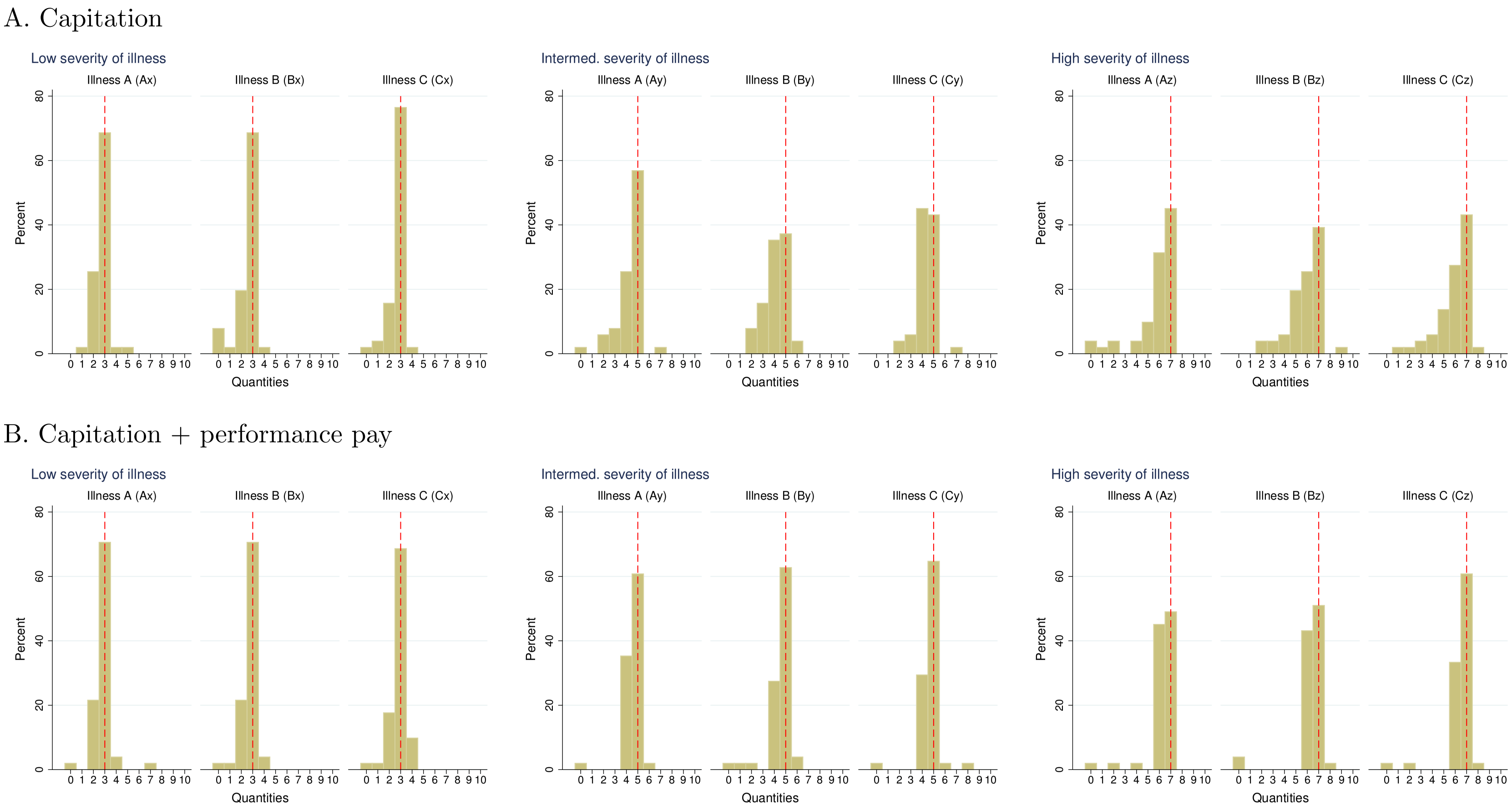
| | $p$-values WSR | $p$-values FPPT |
|---|---|---|
| **CAP** | (High bonus, Low bonus) | (High bonus, Low bonus) |
| Mild severity of illness ($l = x$) | | |
| Illness $A$ | 0.2074, 0.0034 | 0.9989, 0.0158 |
| Illness $B$ | 0.0014, 0.0034 | 0.0015, 0.0004 |
| Illness $C$ | 0.0138, 0.0037 | 0.0354, 0.0045 |
| Intermediate severity of illness ($l = y$) | | |
| Illness $A$ | 0.0000, 0.0000 | 0.0001, 0.0001 |
| Illness $B$ | 0.0000, 0.0000 | 0.0009, 0.0000 |
| Illness $C$ | 0.0000, 0.0000 | 0.0000, 0.0000 |
| High severity of illness ($l = z$) | | |
| Illness $A$ | 0.0000, 0.0000 | 0.0000, 0.0000 |
| Illness $B$ | 0.0000, 0.0000 | 0.0000, 0.0000 |
| Illness $C$ | 0.0000, 0.0000 | 0.0000, 0.0000 |

*Notes:* Comparison between health care provision and patient-optimal care for both the Low bonus and High bonus condition. Two-sided p-values are shown for Wilcoxon sigend rank tests for matched samples and for Fisher-Pitman permutation tests for paired samples.

Table B.2: Physicians' health care service provision under capitation

| | Quantity, $q$ | | |
|---|---|---|---|
| Model: | (1) | (2) | (3) |
| High bonus (= 1 if 20%-Bonus) | -0.006 | 0.006 | 0.044 |
| | (0.163) | (0.158) | (0.158) |
| Interm. severity of illness (= 1 if $l = y$) | 1.574*** | 1.574*** | 1.574*** |
| | (0.088) | (0.088) | (0.088) |
| High severity of illness (= 1 if $l = z$) | 3.061*** | 3.061*** | 3.061*** |
| | (0.141) | (0.141) | (0.141) |
| High marginal health benefit (= 1 if $\theta = 2$, Illness $C$) | -0.029 | -0.029 | -0.029 |
| | (0.063) | (0.063) | (0.063) |
| Female | | 0.107 | 0.077 |
| | | (0.176) | (0.172) |
| Years in practice | | 0.010 | 0.009 |
| | | (0.010) | (0.010) |
| Other physician characteristics | No | No | Yes |
| Constant | 2.737*** | 2.500*** | 1.825*** |
| | (0.110) | (0.287) | (0.454) |
| Wald-test: | | | |
| $H_0$: Interm. severity = High severity ($p$-value) | 0.000 | 0.000 | 0.000 |
| $R^2$ | 0.5042 | 0.5071 | 0.5181 |
| Observations | 936 | 936 | 936 |
| Physicians | 104 | 104 | 104 |

*Notes:* Ordinary Least Square (OLS) estimates are reported with robust standard errors clustered for subjects (in brackets). The reference category is the 'low severity of illness', $l = x$. 'Other physician characteristics' comprise a question each for the attitude towards altruism and competition from the European Values Study (European Values Study, 2016), and risk attitudes according to the German Socio-Economic Panel on the willingness to take risk in general, related to health (Dohmen et al., 2011) and one questions eliciting attitudes related to a patient's health. *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$.


A. Capitation

B. Capitation + performance pay

Figure B.1: Distributions of physicians' quantity choice by severity of illness under Low bonus (5%)

*Notes:* This figure shows the distribution of physicians' chosen quantities of medical services for the three severities of illness and for the illnesses $A$, $B$, and $C$. The red dashed vertical line indicates the patient-optimal quantity of medical services being $q_x^* = 3$, $q_y^* = 5$, and $q_z^* = 7$ for the low ($x$), the intermediate ($y$), and the high severity ($z$) of illness, respectively. Also recall that the profit-maximizing quantity choices is 0 under capitation; under capitation + performance pay, the profit-maximizing quantities are 2, 4, and 6 for patients with the low ($x$), the intermediate ($y$), and the high severity ($z$) of illness, respectively. For each patient ($Ax$ to $Cz$), we observe the quantity choices of 51 primary care physicians.

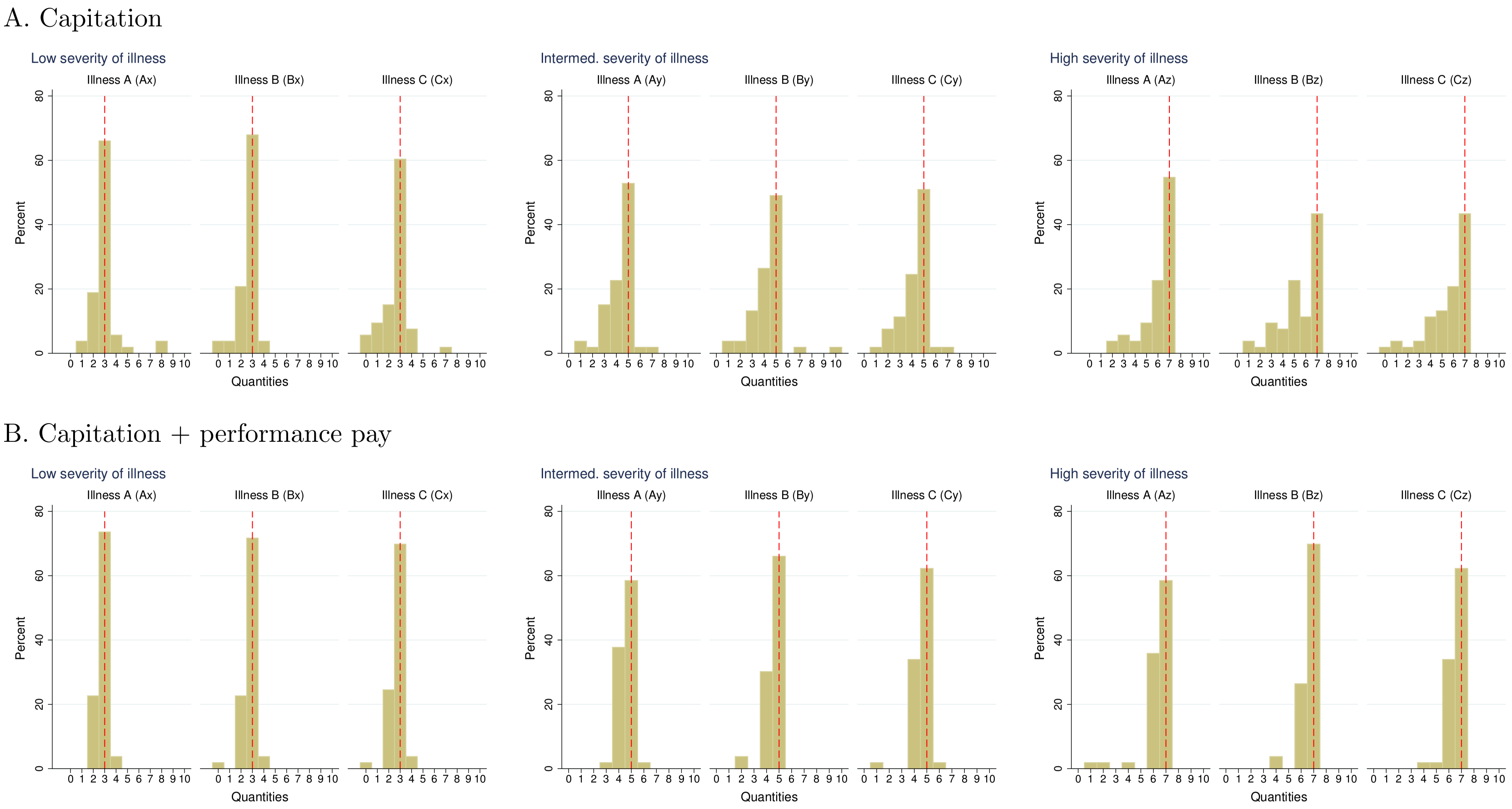


Figure B.2: Distributions of physicians' quantity choice by severity of illness under High bonus (20%)

*Notes:* This figure shows the distribution of physicians' chosen quantities of medical services for the three severities of illness and for the illnesses $A$, $B$, and $C$. The red dashed vertical line indicates the patient-optimal quantity of medical services being $q_x^* = 3$, $q_y^* = 5$, and $q_z^* = 7$ for the low ($x$), the intermediate ($y$), and the high severity ($z$) of illness, respectively. Also recall that the profit-maximizing quantity choices is 0 under capitation; under capitation + performance pay, the profit-maximizing quantities are 2, 4, and 6 for patients with the low ($x$), the intermediate ($y$), and the high severity ($z$) of illness, respectively. For each patient ($Ax$ to $Cz$), we observe the quantity choices of 53 primary care physicians.

Table B.3: Multilevel mixed-effects regressions on the effect of marginal health benefit and performance pay on health care provision

| | A. Quantity $q_{kl}$ | | | |
|---|---|---|---|---|
| Model: | (1) | (2) | (3) | (4) |
| **Fixed effects** | | | | |
| Performance pay | 0.304*** | 0.304*** | 0.028 | 0.028 |
| | (0.075) | (0.075) | (0.090) | (0.090) |
| High bonus (= 1 if 20%-Bonus) | 0.036 | 0.106 | 0.036 | 0.106 |
| | (0.125) | (0.131) | (0.125) | (0.131) |
| Interm. severity of illness (= 1 if $l = y$) | 1.694*** | 1.694*** | 1.574*** | 1.574*** |
| | (0.049) | 0.049 | (0.065) | (0.065) |
| High severity of illness (= 1 if $l = z$) | 3.356*** | 3.356*** | 3.061*** | 3.061*** |
| | (0.049) | (0.049) | (0.065) | (0.065) |
| High marginal health benefit (= 1 if $\theta = 2$) | -0.029 | -0.029 | -0.029 | -0.029 |
| | (0.057) | (0.057) | (0.057) | (0.057) |
| Performance pay× Interm. severity | | | 0.240** | 0.240** |
| | | | (0.086) | (0.086) |
| Performance pay× High severity | | | 0.590*** | 0.590*** |
| | | | (0.086) | (0.086) |
| Performance pay× High marginal health benefit | 0.090 | 0.090 | 0.090 | 0.090 |
| | (0.076) | (0.076) | (0.074) | (0.074) |
| Physicians' characteristics | No | Yes | No | Yes |
| Constant | 2.577*** | 1.773*** | 2.715*** | 1.911*** |
| | (0.109) | (0.384) | (0.112) | (0.385) |
| **Random effects** | | | | |
| Subject level | | | | |
| Var(Performance pay) | 0.389*** | 0.389*** | 0.396*** | 0.396*** |
| | (0.073) | (0.073) | (0.073) | (0.073) |
| Var(Constant) | 1.643*** | 1.627*** | 1.659*** | 1.642*** |
| | (0.278) | (0.277) | (0.278) | (0.277) |
| Cov(Performance pay, Constant) | -0.074*** | -0.074*** | -0.719*** | -0.716*** |
| | (0.135) | (0.135) | (0.135) | (0.135) |
| Patient level | | | | |
| Var(Constant) | 0.074*** | 0.074*** | 0.090*** | 0.090*** |
| | (0.024) | (0.024) | (0.023) | (0.023) |
| Var(Residual) | 0.607*** | 0.607*** | 0.575*** | 0.575*** |
| | (0.030) | (0.030) | (0.028) | (0.028) |
| Observations | 1.872 | 1.872 | 1.872 | 1.872 |
| Physicians | 104 | 104 | 104 | 104 |

*Notes:* This table shows parameter estimates from multilevel mixed-effects REML regressions. Standard errors are shown in parentheses. Models 1-4 include subject and patient-specific random effects with stage at level 1, patients at level 2 and subjects at level 3. The reference category is the 'low severity of illness', $l = x$. Physicians' characteristics comprise gender, practice years, a question each for the attitude towards altruism and competition from the European Values Study (European Values Study, 2016), and on the willingness to take risk in general, related to health according to the German SOEP (Dohmen et al., 2011) and one questions eliciting attitudes related to a patient's health. *** $p < 0.001$, ** $p < 0.01$, and *$p < 0.05$.

# C Control experiments with medical students

In this section, we provide details on the procedure of the online experiments with medical students (Subsection C.1) and present results from these experiments (Subsection C.2).

## C.1 Procedure in the control experiments

In our control experiments, we employed nearly the same double-blind procedure as for primary care physicians. Our procedure was also in accordance with the data protection guidelines of the Medical Faculty of the University of Cologne, on which medical students were informed. Invitations to medical students were sent out via email by a trustee at the Medical Faculty. Before logging into the experiment, each participant created a personal ID. All decisions in the control experiments were made using these IDs, and we can relate medical students' choices to these IDs only.

The payment to the participants was made in cash about one week after login for the experiment had expired. Participants were informed at the end of the experiment where and when to get their payment. The trustee provided us with a list containing the participants' IDs and with the anonymized data. For each ID, we computed the respective payment and put the money and a receipt into an envelope that was marked with the ID. Participants were handed out the respective envelope when providing their ID. They signed the receipt, which they then confidentially put into a box and left. This payment procedure does not allow us to trace any individual subjects' decisions.

To verify that the money corresponding to the sum of the patient benefits was actually transferred, we applied a procedure similar to Hennig-Schmidt et al. (2011) and Eckel and Grossman (1996). To this end, one of the participants was randomly chosen to be a monitor. When getting his/her money, the monitor verified that a payment order on the amount of the aggregate benefit was written to the financial department of the University of Cologne to transfer the money to

Christoffel Blindenmission. 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.

The online experiment was programmed using the software ILIAS (https://www.ilias.uni-koeln.de/) and was conducted in May/June 2018. The experimental procedure was nearly identical to that of the main conditions. All monetary amounts from the main experiment were multiplied by 0.32 for the medical students to have comparable financial incentives for physicians and students. After having finished the second part of the experiment, medical students were asked to complete a questionnaire on social demographics (age and gender), risk and time preferences, the social traits altruism and competitiveness, and their general attitude towards pay for performance.

Medical students earned, on average, €15.22 for completing the experiment and the questionnaire, which took on average 40 minutes[27]. In total, €1,270.40 were transferred to Christoffel Blindenmission. The control experiments, thus, allowed to treat another 40 patients. In total, cataract operations of 206 patients were financed by our study.

Table C.1 characterizes our student sample in terms of their demographics and self-reported attitudes.

## C.2 Results from control experiments with medical students

In this section, we check for the robustness of our results towards (i) subject pool effects, (ii) order effects, and (iii) income effects. We conducted all control experiments with medical students, enrolled at the University of Cologne, Germany, who are supposed to become physicians in the future. To ensure that the student

[27] The payment is equivalent to an hourly wage of €22.83 and is more than twice as high as the gross hourly wage of €9.21 a student helper is paid at the University of Cologne (https://verwaltung.uni-koeln.de/abteilung41/content/e143023/e143137/e143150/e143209/Hilfskraftrichtlinie_ger.pdf, retrieved 01.08.2018).

Table C.1: Descriptives of the student sample

| Condition | C–CAP–High | | C–Low–<br>bonus (5%) | | C–High–bonus<br>(20%)–First | | C–High–)<br>bonus (20%) | |
|---|---|---|---|---|---|---|---|---|
| | Mean | s.d. | Mean | s.d. | Mean | s.d. | Mean | s.d. |
| Share of female | 0.62 | 0.49 | 0.80 | 0.41 | 0.59 | 0.50 | 0.67 | 0.48 |
| Age | 24.27 | 3.54 | 24.23 | 2.87 | 24.96 | 5.73 | 24.67 | 6.77 |
| Risk attitudes: General | 4.24 | 2.25 | 4.00 | 2.55 | 3.33 | 2.02 | 3.94 | 1.90 |
| Altruism attitudes | 4.51 | 1.89 | 4.40 | 1.75 | 4.33 | 2.11 | 4.58 | 1.95 |
| Competition attitude | 4.11 | 1.74 | 4.10 | 1.97 | 4.33 | 1.75 | 3.79 | 1.32 |
| Risk attitudes: Own health | 4.03 | 2.13 | 4.13 | 2.13 | 3.59 | 1.76 | 3.45 | 1.54 |
| Risk attitudes: Patient's health | 3.46 | 2.06 | 2.63 | 1.19 | 3.15 | 2.01 | 2.70 | 1.55 |
| *N* | 37 | | 30 | | 27 | | 33 | |

This table presents summary statistics of individual students characteristics and attitudes for all control experiments

sample is a valid control for the primary care physician sample we first analyze whether the behavior of the medical students does not differ significantly from that of primary care physicians. To this end, we conducted the experimental conditions C–Low–bonus (5%) and C–High–bonus (20%) using the same experimental parameters as for the primary care physicians, the only difference being that we adapted the conversion rate to have adequate financial incentives for medical students. Second, the test of order effects allows us to analyze whether adding performance pay or taking it away affects behavior. We test this in condition C–High–bonus (20%)–First compared to C–High–bonus (20%). Finally, our second control condition is related to the fact that in CAP+P4P the maximum attainable profit is higher than in CAP. To check whether this design feature of our experiment has an effect, we raised the lump-sum reimbursement in CAP by 20%, while keeping the CAP+P4P-20% payment constant in the second part of the experiment.

As to subject pool effects, descriptive statistics in Table C.2 indicate that medical students respond to performance pay in a similar way as physicians. The descriptive statistics suggest that underprovision under CAP is reduced under CAP+P4P for both bonus levels. This patten is confirmed at an individual patient level, see Table C.3. Regression analyses in Table C.5 provide further support. Similar to the physicians, medical students increase the quantity of

Table C.2: Health care service provision in the control experiments with medical students

| | Capitation (CAP) | | | | CAP+P4P | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Condition | Mean | s.d. | Min | Max | Mean | s.d. | Min | Max | *N* |
| C–Low–bonus (5%) | 3.95 | 1.88 | 0 | 7 | 4.54 | 1.63 | 1 | 7 | 540 |
| C–High–bonus (20%) | 3.92 | 1.97 | 0 | 8 | 4.63 | 1.67 | 2 | 7 | 594 |
| C–CAP–High | 4.09 | 1.87 | 0 | 7 | 4.65 | 1.61 | 2 | 8 | 666 |
| C–High–bonus (20%)–First | 3.60 | 1.99 | 0 | 7 | 4.67 | 1.63 | 2 | 7 | 486 |

*Notes:* This table shows descriptive statistics of health care services medical students chose in our control experiments. In C–Low–bonus (5%) there are 30, in C–High–bonus (20%) 33, in C–CAP–High 37, and in C–High–bonus (20%)–First 27 participants.

health care services significantly when P4P is introduced. The effect significantly depends on patients' severity of illness. The non-optimal service provision is significantly reduced under P4P, which, again, depends on the severity of illness; see estimation results of Models (1) and (2) of Table C.5. When comparing health care service provision of physicians and medical students, estimation results show that the effect of performance pay on medical service provision and relative quality is not significantly different across subject pools; see Models (1) and (5) of Table C.6.

Interestingly, physicians respond somewhat stronger to performance pay than medical students. We find that physicians provide significantly less health care services under performance pay compared to medical students under CAP. This result suggests that physicians are even more sensitive to the introduction of performance pay; see; see Models (2) and (4) for quantity and, for quality of care, Models (6) and (8) of Table C.6.

We now analyze the robustness of our results with regard to order and income effects. Descriptive statistics in Table C.2 indicate that the order of conditions does not affect behavior as the effect of performance pay is very similar for medical students in the control conditions C–High–bonus (20%)–First compared to C–High–bonus (20%); see descriptive statistics and non-parametric test results in Table C.4 and estimation results in Models (5), (6) and (9) of Table C.5. Estimation results of regression analyses in Table C.5 provide support for the absence of income effects. The increase in the quantity of health care services under P4P is not significantly different between C–CAP-High and C–High–bonus (20%), see Models (3), (4) and (8) in Table C.5.

In sum, the effect of performance pay on health care service provision is robust (i) between physician and medical student samples, (ii) towards keeping the level of incentives constant between capitation and blended capitation plus performance pay, and (iii) concerning the order of payment systems.

Table C.3: Quantity $q$ and relative quality of health care provision $\rho$ in C-Low-bonus (5%) and C-High-bonus (20%) with medical students by payment system, illness, and severity of illness

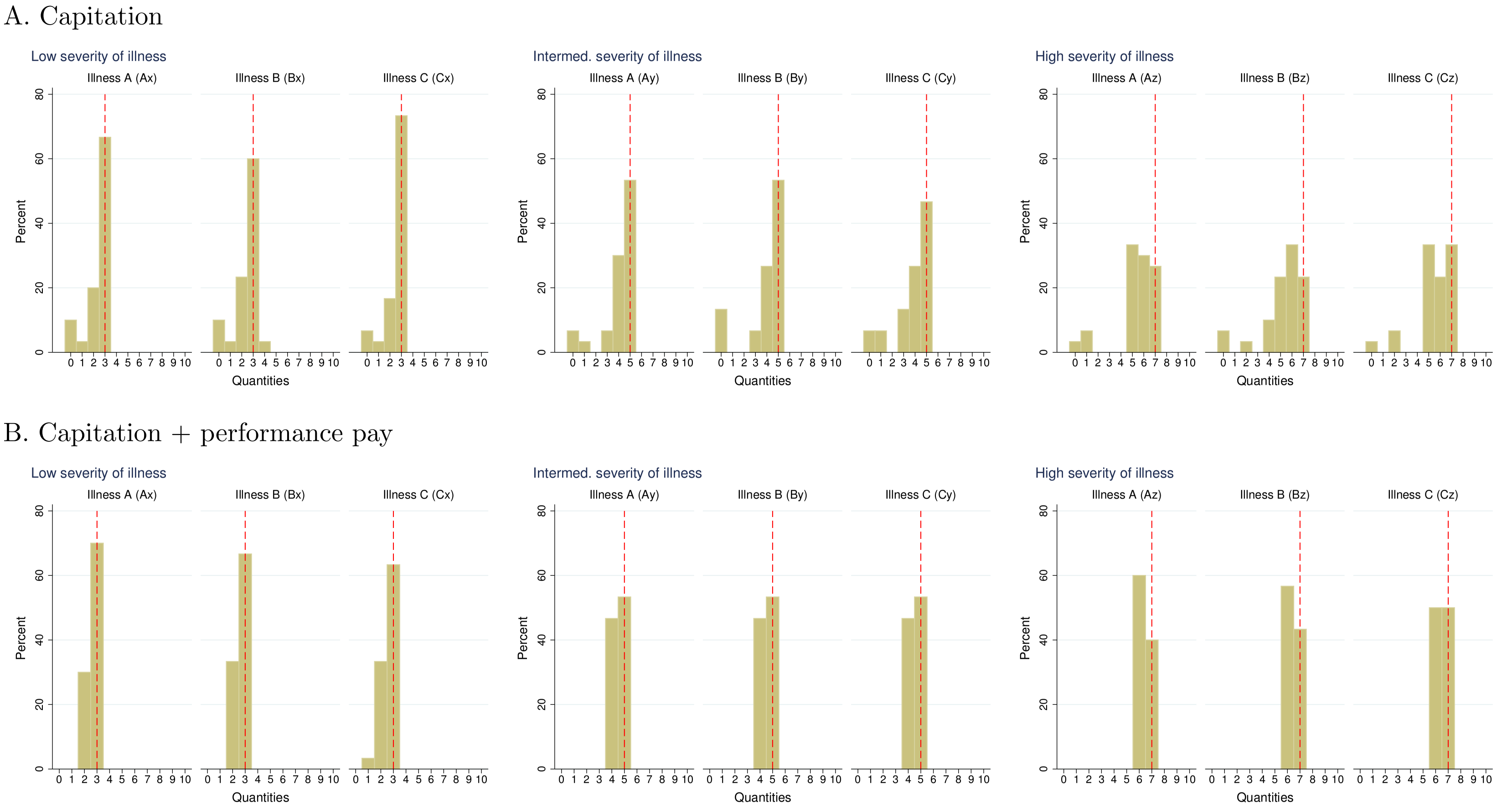
| | **A. Quantity, $q$** | | | | **B. Relative quality, $\rho$** | | | |
|---|---|---|---|---|---|---|---|---|
| | CAP | CAP+P4P | %-change | $p$-value | CAP | CAP + P4P | %-change | $p$-value |
| **C-Low-bonus (5%)** | | | | | | | | |
| Low severity of illness | | | | | | | | |
| Illness $A$ | 2.43 (0.97) | 2.70 (0.47) | 10.96 | 0.082 [0.100] | 0.81 (0.32) | 0.90 (0.16) | 10.96 | 0.078 [0.066] |
| Illness $B$ | 2.43 (1.01) | 2.66 (0.48) | 9.59 | 0.296 [0.242] | 0.79 (0.32) | 0.89 (0.16) | 12.69 | 0.100 [0.073] |
| Illness $C$ | 2.57 (0.86) | 2.60 (0.56) | 1.30 | 0.944 [1.000] | 0.86 (0.29) | 0.87 (0.19) | 1.30 | 0.879 [0.713] |
| Intermediate severity of illness | | | | | | | | |
| Illness $A$ | 4.10 (1.42) | 4.53 (0.51) | 10.57 | 0.108 [0.093] | 0.82 (0.28) | 0.91 (0.10) | 10.57 | 0.108 [0.091] |
| Illness $B$ | 3.93 (1.68) | 4.53 (0.51) | 15.25 | 0.063 [0.045] | 0.91 (0.10) | 0.47 (0.51) | 15.25 | 0.063 [0.046] |
| Illness $C$ | 3.87 (1.53) | 4.53 (0.51) | 17.24 | 0.006 [0.007] | 0.91 (0.10) | 0.47 (0.51) | 17.24 | 0.006 [0.007] |
| High severity of illness | | | | | | | | |
| Illness $A$ | 5.40 (1.79) | 6.40 (0.50) | 18.52 | 0.000 [0.000] | 0.91 (0.07) | 0.60 (0.50) | 18.52 | 0.000 [0.000] |
| Illness $B$ | 5.27 (1.84) | 6.43 (0.50) | 22.15 | 0.000 [0.000] | 0.92 (0.07) | 0.57 (0.50) | 22.15 | 0.000 [0.000] |
| Illness $C$ | 5.53 (1.68) | 6.50 (0.51) | 17.47 | 0.000 [0.000] | 0.79 (0.24) | 0.93 (0.07) | 17.47 | 0.000 [0.000] |
| Aggregated | 3.95 (1.87) | 4.54 (1.63) | 14.94 | | 0.79 (0.29) | 0.90 (0.12) | 14.02 | |
| **C-High-bonus (20%)** | | | | | | | | |
| Low severity of illness | | | | | | | | |
| Illness $A$ | 2.42 (0.94) | 2.73 (0.52) | 12.50 | 0.154 [0.099] | 0.81 (0.31) | 0.89 (0.16) | 10.00 | 0.145 [0.060] |
| Illness $B$ | 2.45 (1.03) | 2.67 (0.48) | 8.64 | 0.570 [0.296] | 0.82 (0.34) | 0.89 (0.16) | 8.64 | 0.570 [0.161] |
| Illness $C$ | 2.55 (1.39) | 2.76 (0.44) | 8.33 | 0.087 [0.440] | 0.77 (0.36) | 0.92 (0.15) | 19.74 | 0.013 [0.005] |
| Intermediate severity of illness | | | | | | | | |
| Illness $A$ | 4.15 (1.46) | 4.61 (0.70) | 10.95 | 0.190 [0.076] | 0.83 (0.29) | 0.91 (0.14) | 9.49 | 0.424 [0.111] |
| Illness $B$ | 3.82 (1.59) | 4.70 (0.47) | 23.02 | 0.000 [0.000] | 0.76 (0.32) | 0.94 (0.09) | 23.02 | 0.000 [0.000] |
| Illness $C$ | 4.03 (1.45) | 4.64 (0.49) | 15.04 | 0.009 [0.008] | 0.81 (0.29) | 0.93 (0.09) | 15.04 | 0.009 [0.007] |
| High severity of illness | | | | | | | | |
| Illness $A$ | 5.21 (2.07) | 6.48 (0.57) | 24.42 | 0.000 [0.000] | 0.74 (0.29) | 0.93 (0.08) | 24.42 | 0.000 [0.000] |
| Illness $B$ | 5.12 (2.07) | 6.61 (0.66) | 28.99 | 0.000 [0.000] | 0.73 (0.29) | 0.94 (0.09) | 28.99 | 0.000 [0.000] |
| Illness $C$ | 5.52 (2.03) | 6.48 (1.03) | 17.58 | 0.001 [0.002] | 0.79 (0.29) | 0.93 (0.15) | 17.58 | 0.001 [0.002] |
| Aggregated | 3.92 (1.97) | 4.63 (1.67) | 18.11 | | 0.78 (0.31) | 0.92 (0.13) | 17.44 | |

*Notes:* This table shows descriptive statistics on the quantities $q$ and the relative quality of health care provision $\rho$ at the level of payment systems, illnesses, severities of illness (means and standard deviations in brackets). 33 medical students decide in the C-High-bonus (20%) and 30 in the C-Low-bonus (5%) condition. Two-sided $p$-values are shown for Wilcoxon sigend rank tests for matched samples and for Fisher-Pitman permutation tests for paired samples in squared brackets.

Table C.4: Quantity $q$ and relative quality of health care provision $\rho$ in C-CAP-High and C-High-bonus (20%)-First with medical students by payment system, illness, and severity of illness

| | **A. Quantity, $q$** | | | | **B. Relative quality, $\rho$** | | | |
|---|---|---|---|---|---|---|---|---|
| | CAP | CAP+P4P | %-change | $p$-value | CAP | CAP + P4P | %-change | $p$-value |
| **C-CAP-High** | | | | | | | | |
| Low severity of illness | | | | | | | | |
| Illness *A* | 2.68 (0.88) | 2.92 (0.64) | 9.09 | 0.019 [0.032] | 0.86 (0.28) | 0.92 (0.19) | 7.37 | 0.094 [0.110] |
| Illness *B* | 2.68 (0.91) | 2.78 (0.58) | 4.04 | 0.478 [0.469] | 0.86 (0.29) | 0.89 (0.18) | 4.21 | 0.432 [0.246] |
| Illness *C* | 2.70 (0.91) | 2.76 (0.60) | 2.00 | 0.923 [0.807] | 0.86 (0.29) | 0.88 (0.18) | 2.08 | 0.846 [0.475] |
| Intermediate severity of illness | | | | | | | | |
| Illness *A* | 4.05 (1.41) | 4.59 (0.50) | 13.33 | 0.006 [0.006] | 0.81 (0.28) | 0.92 (0.09) | 13.33 | 0.006 [0.006] |
| Illness *B* | 4.16 (1.28) | 4.54 (0.56) | 9.09 | 0.057 [0.053] | 0.83 (0.26) | 0.89 (0.10) | 7.79 | 0.158 [0.102] |
| Illness *C* | 4.14 (1.40) | 4.70 (0.52) | 13.73 | 0.004 [0.003] | 0.83 (0.28) | 0.93 (0.09) | 12.42 | 0.015 [0.012] |
| High severity of illness | | | | | | | | |
| Illness *A* | 5.59 (1.92) | 6.41 (0.60) | 14.49 | 0.007 [0.004] | 0.80 (0.27) | 0.91 (0.08) | 14.49 | 0.007 [0.004] |
| Illness *B* | 5.16 (2.11) | 6.46 (0.61) | 25.13 | 0.000 [0.000] | 0.74 (0.30) | 0.91 (0.07) | 24.08 | 0.000 [0.000] |
| Illness *C* | 5.62 (1.98) | 6.68 (0.47) | 18.75 | 0.001 [0.000] | 0.80 (0.28) | 0.95 (0.07) | 18.75 | 0.001 [0.000] |
| Aggregated | 4.09 (1.87) | 4.65 (1.61) | 13.69 | | 0.82 (0.28) | 0.91 (0.13) | 11.61 | |
| **C-High-bonus (20%)- First** | | | | | | | | |
| Low severity of illness | | | | | | | | |
| Illness *A* | 2.41 (1.01) | 2.78 (0.51) | -13.33 | 0.018 [0.027] | 0.80 (0.34) | 0.90 (0.15) | -10.96 | 0.076 [0.065] |
| Illness *B* | 2.22 (1.12) | 2.81 (0.40) | -21.05 | 0.003 [0.003] | 0.74 (0.37) | 0.94 (0.13) | -21.05 | 0.003 [0.003] |
| Illness *C* | 2.37 (1.08) | 2.78 (0.42) | -14.67 | 0.017 [0.023] | 0.79 (0.35) | 0.93 (0.14) | -14.67 | 0.017 [0.023] |
| Intermediate severity of illness | | | | | | | | |
| Illness *A* | 3.63 (1.67) | 4.63 (0.49) | -21.60 | 0.002 [0.001] | 0.73 (0.33) | 0.93 (0.09) | -21.6 | 0.002 [0.001] |
| Illness *B* | 3.56 (1.58) | 4.59 (0.50) | -22.58 | 0.000 [0.000] | 0.71 (0.32) | 0.92 (0.10) | 22.58 | 0.000 [0.000] |
| Illness *C* | 3.59 (1.65) | 4.63 (0.49) | -22.40 | 0.001 [0.000] | 0.72 (0.33) | 0.93 (0.09) | -22.4 | 0.001 [0.000] |
| High severity of illness | | | | | | | | |
| Illness *A* | 4.93 (2.25)) | 6.59 (0.50) | -25.28 | 0.000 [0.000] | 0.70 (0.32) | 0.94 (0.07) | -25.28 | 0.000 [0.000] |
| Illness *B* | 4.63 (2.26) | 6.52 (0.51) | -28.98 | 0.000 [0.000] | 0.66 (0.32) | 0.93 (0.07) | -28.98 | 0.000 [0.000] |
| Illness *C* | 5.11 (2.26) | 6.70 (0.47) | -23.76 | 0.001 [0.000] | 0.73 (0.32) | 0.96 (0.07) | -23.76 | 0.000 [0.000] |
| Aggregated | 3.60 (1.99) | 4.67 (1.63) | -21.52 | | 0.73 (0.33) | 0.93 (0.11) | -21.25 | |

*Notes:* This table shows descriptive statistics on the quantities $q$ and the relative quality of health care provision $\rho$ at the level of payment systems, illnesses, severities of illness (means and standard deviations in brackets). 37 medical students decide in the C-CAP-High and 27 in the C-High-bonus (20%)-First condition. Recall, in condition C-High-bonus (20%) CAP+P4P refers to part 1 and CAP to part 2 of the experiment. The percentage change relates to the change from part 1 to part 2. Two-sided $p$-values are shown for Wilcoxon signed rank tests for matched samples and for Fisher-Pitman permutation test for paired samples in squared brackets.

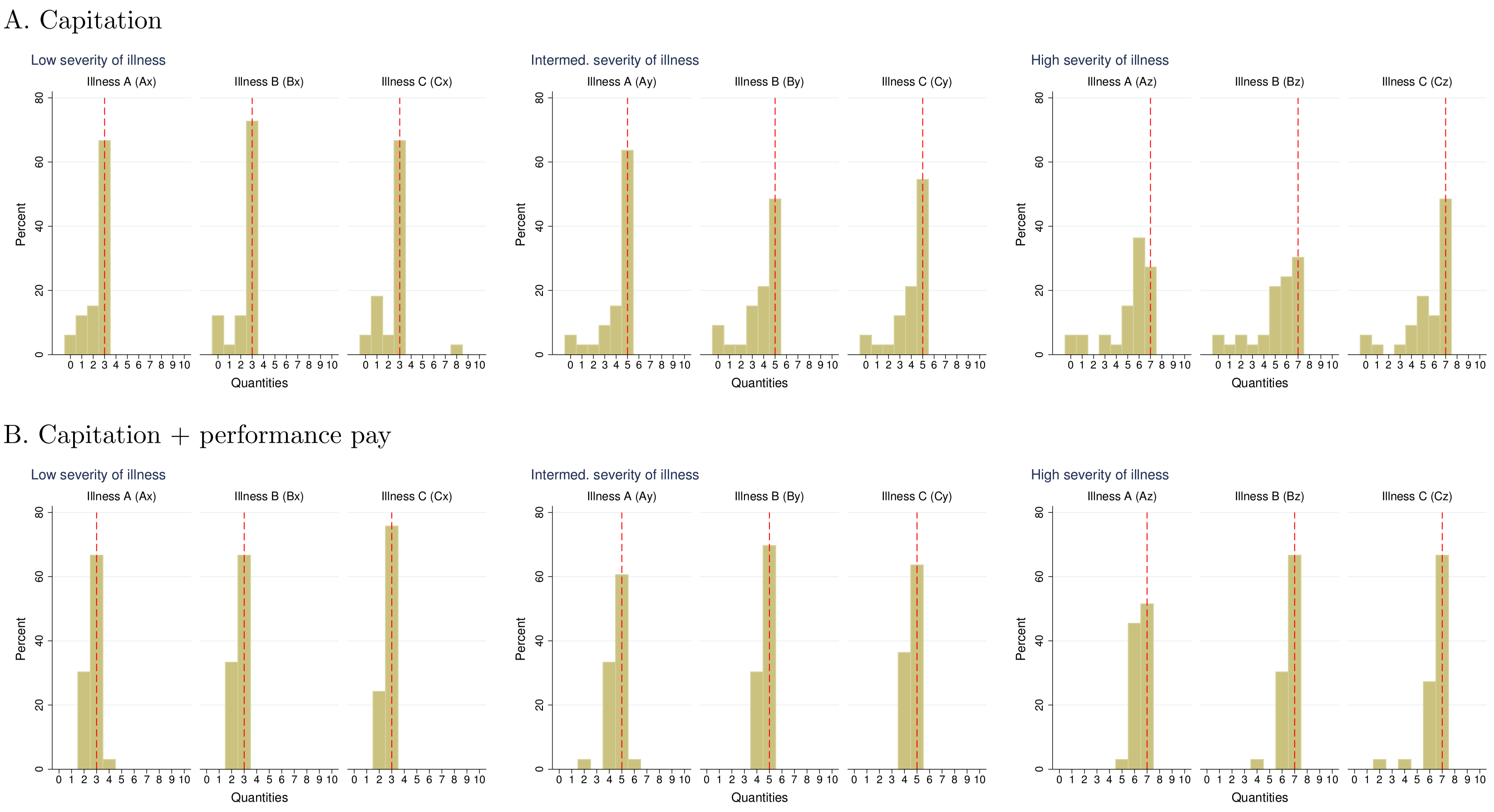


A. Capitation

B. Capitation + performance pay

Figure C.1: Distributions of medical students' quantity choice by severity of illness under C-Low-bonus (5%)

*Notes:* This figure shows the distribution of medical students' chosen quantities of medical services for the three severities of illness and for the illnesses $A$, $B$, and $C$. The red dashed vertical line indicates the patient-optimal quantity of medical services being $q_x^* = 3$, $q_y^* = 5$, and $q_z^* = 7$ for the low ($x$), the intermediate ($y$), and the high severity ($z$) of illness, respectively. Also recall that the profit-maximizing quantity choices is 0 under capitation; under capitation + performance pay, the profit-maximizing quantities are 2, 4, and 6 for patients with the low ($x$), the intermediate ($y$), and the high severity ($z$) of illness, respectively. For each patient ($Ax$ to $Cz$), we observe the quantity choices of 30 medical students.

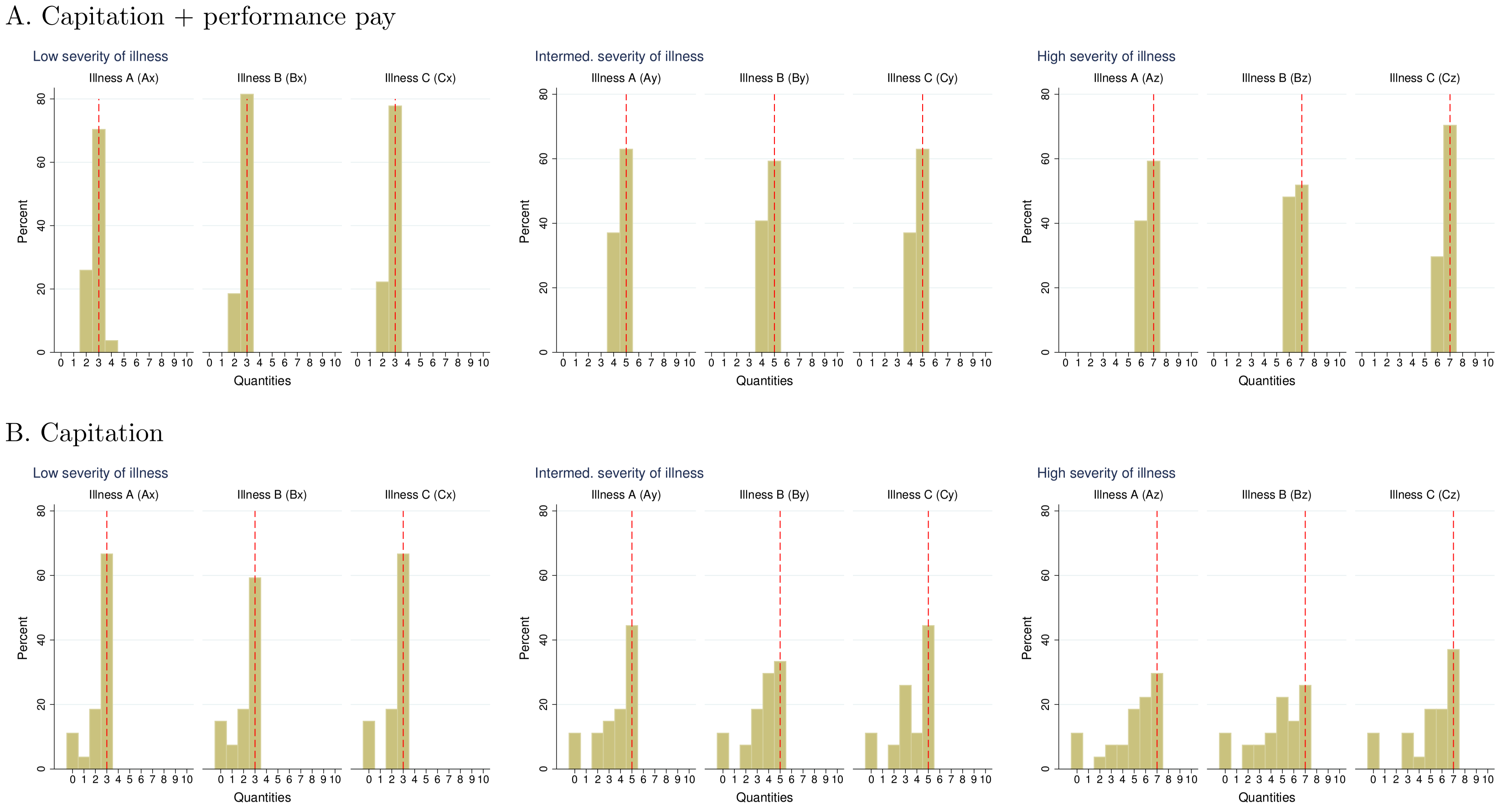


A. Capitation

B. Capitation + performance pay

Figure C.2: Distributions of medical students' quantity choice by severity of illness under C-High-bonus (20%)

*Notes:* This figure shows the distribution of medical students' chosen quantities of medical services for the three severities of illness and for the illnesses $A$, $B$, and $C$. The red dashed vertical line indicates the patient-optimal quantity of medical services being $q_x^* = 3$, $q_y^* = 5$, and $q_z^* = 7$ for the low ($x$), the intermediate ($y$), and the high severity ($z$) of illness, respectively. Also recall that the profit-maximizing quantity choices is 0 under capitation; under capitation + performance pay, the profit-maximizing quantities are 2, 4, and 6 for patients with the low ($x$), the intermediate ($y$), and the high severity ($z$) of illness, respectively. For each patient ($Ax$ to $Cz$), we observe the quantity choices of 33 medical students.

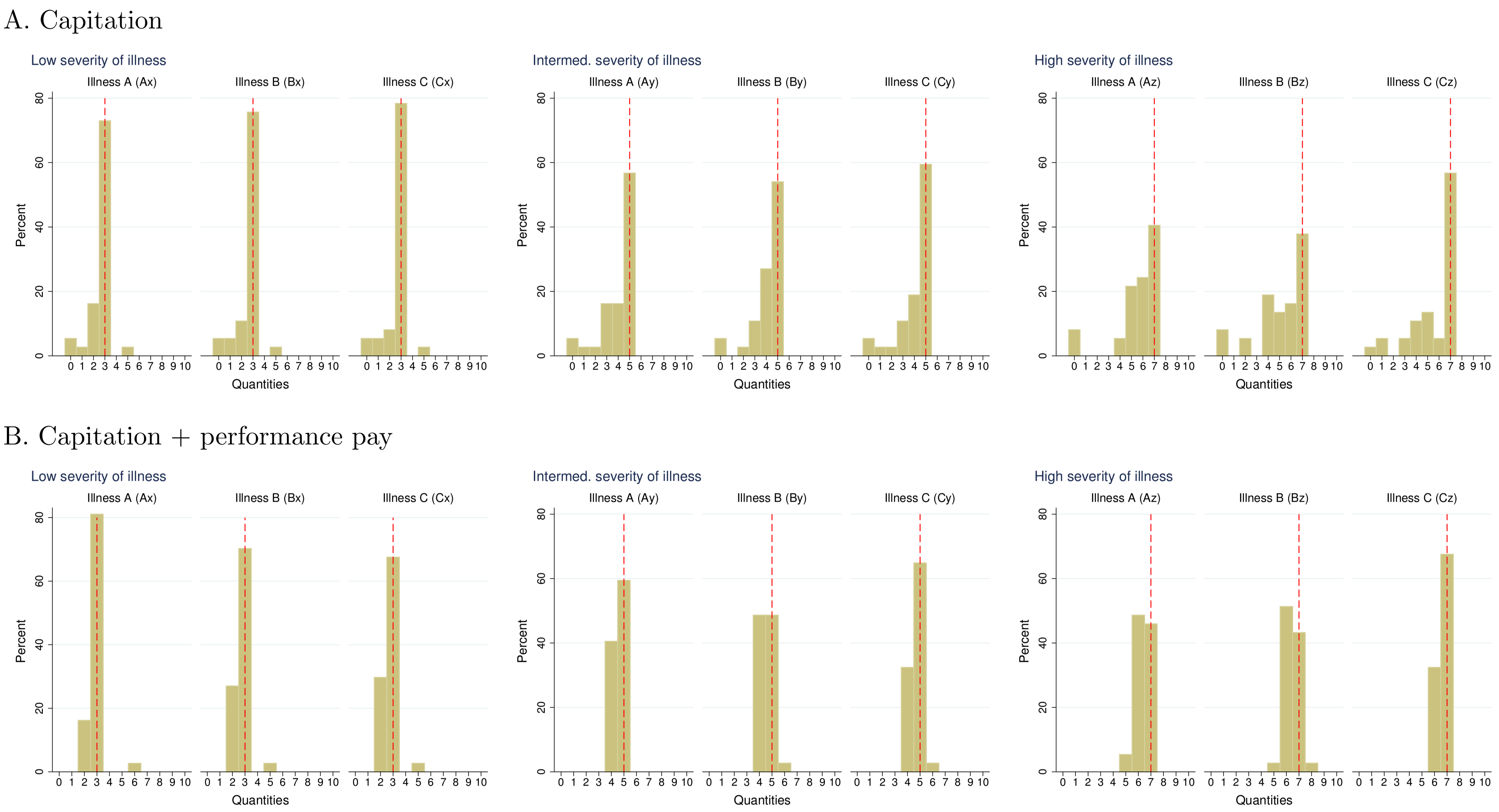


A. Capitation + performance pay

B. Capitation

Figure C.3: Distributions of medical students' quantity choice by severity of illness under C-High-bonus (20%)-First

*Notes:* This figure shows the distribution of medical students' chosen quantities of medical services for the three severities of illness and for the illnesses $A$, $B$, and $C$. The red dashed vertical line indicates the patient-optimal quantity of medical services being $q_x^* = 3$, $q_y^* = 5$, and $q_z^* = 7$ for the low ($x$), the intermediate ($y$), and the high severity ($z$) of illness, respectively. Also recall that the profit-maximizing quantity choices is 0 under capitation; under capitation + performance pay, the profit-maximizing quantities are 2, 4, and 6 for patients with the low ($x$), the intermediate ($y$), and the high severity ($z$) of illness, respectively. For each patient ($Ax$ to $Cz$), we observe the quantity choices of 27 medical students.



A. Capitation

B. Capitation + performance pay

Figure C.4: Distributions of medical students' quantity choice by severity of illness under C-CAP-High

*Notes:* This figure shows the distribution of medical students' chosen quantities of medical services for the three severities of illness and for the illnesses $A$, $B$, and $C$. The red dashed vertical line indicates the patient-optimal quantity of medical services being $q_x^* = 3$, $q_y^* = 5$, and $q_z^* = 7$ for the low ($x$), the intermediate ($y$), and the high severity ($z$) of illness, respectively. Also recall that the profit-maximizing quantity choices is 0 under capitation; under capitation + performance pay, the profit-maximizing quantities are 2, 4, and 6 for patients with the low ($x$), the intermediate ($y$), and the high severity ($z$) of illness, respectively. For each patient ($Ax$ to $Cz$), we observe the quantity choices of 37 medical students.

Table C.5: Health care service provision in experiments with medical students

| | Quantity, $q$ | | | | | | Relative quality, $\rho_{kl}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| C-Low-bonus (5%) | -0.094 | -0-094 | | | | | -0.104 | | |
| | (0.203) | (0.203) | | | | | (0.042) | | |
| C-CAP-High | | | 0.088 | 0.088 | | | | 0.021 | |
| | | | (0.202) | (0.202) | | | | (0.041) | |
| C-High-bonus (20%)-First | | | | | -0.242 | -0.242 | | | -0-042 |
| | | | | | (0.211) | (0.211) | | | (0.043) |
| Performance pay | 0.656*** | 0.212* | 0.632*** | 0.186* | 0.870*** | 0.339*** | 0.123*** | 0.113*** | 0.163*** |
| | (0.137) | (0.090) | (0.126) | (0.078) | (0.153) | (0.097) | (0.027) | (0.025) | (0.030) |
| Intermediate severity | 1.706*** | 1.508*** | 1.667*** | 1.476*** | 1.644*** | 1.406*** | | | |
| | (0.056) | (0.095) | (0.056) | (0.089) | (0.059) | (0.102) | | | |
| High severity | 3.331*** | 2.862*** | 3.269*** | 2.790*** | 3.253*** | 2.694*** | | | |
| | (0.092) | (0.162) | (0.104) | (0.170) | (0.104) | (0.186) | | | |
| Marg. health ben. (= 1 if $\theta = 2$) | 0.057 | 0.057 | 0.089* | 0.089* | 0.085* | 0.085* | | | |
| | (0.053) | (0.053) | (0.046) | (0.047) | (0.050) | (0.050) | | | |
| Perf. pay× Interm. severity | | 0.397*** | | 0.381*** | | 0.478*** | | | |
| | | (0.096) | | (0.086) | | (0.103) | | | |
| Perf. pay× High severity | | 0.937*** | | 0.957*** | | 1.117*** | | | |
| | | (0.159) | | (0.156) | | (0.177) | | | |
| Constant | 1.620*** | 1.842*** | 2.315*** | 2.538*** | 1.788*** | 2.054*** | 0.649*** | 0.818*** | 0.683*** |
| | (0.505) | (0.501) | (0.425) | (0.428) | (0.524) | (0.523) | (0.107) | (0.088) | (0.110) |
| Observations | 1,134 | 1,134 | 1,260 | 1,260 | 1,080 | 1,080 | 1,134 | 1,260 | 1,080 |
| Subjects | 63 | 63 | 70 | 70 | 60 | 60 | 63 | 70 | 60 |
| $R^2$ | 0.591 | 0.602 | 0.577 | 0.589 | 0.558 | 0.573 | 0.118 | 0.123 | 0.205 |

*Notes:* OLS, clustered standard errors at individual level in parentheses. CAP is the reference category. Comparisons of Perf. pay× High severity and Perf. pay× Int. severity indicate highly significant differences (Wald-tests $p < 0.0001$). *** p<0.001, ** p<0.01, * p<0.05.

Table C.6: Physicians versus students in the Low- and High-bonus condition

| | Quantity, $q$ | | | | Relative quality, $\rho_{kl}$ | | | |
|---|---|---|---|---|---|---|---|---|
| Model: | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Physicians | 0.176 | 0.337* | 0.149 | 0.309* | 0.020 | 0.047 | 0.012 | 0.039 |
| | (0.120) | (0.185) | (0.119) | (0.180) | (0.025) | (0.037) | (0.025) | (0.036) |
| Performance pay | 0.456*** | 0.656*** | 0.456*** | 0.656*** | 0.089*** | 0.123*** | 0.089*** | 0.123*** |
| | (0.069) | (0.136) | (0.069) | (0.136) | (0.013) | (0.027) | (0.013) | (0.027) |
| High bonus (= 1 if 20%-Bonus) | 0.025 | 0.025 | 0.091 | 0.091 | 0.000 | 0.000 | 0.017 | 0.017 |
| | (0.111) | (0.111) | (0.112) | (0.112) | (0.023) | (0.023) | (0.022) | (0.022) |
| Intermediate severity (= 1 if $l = y$) | 1.699*** | 1.699*** | 1.699*** | 1.699*** | | | | |
| | (0.040) | (0.040) | (0.040) | (0.040) | | | | |
| High severity (= 1 if $l = z$) | 3.346*** | 3.346*** | 3.346*** | 3.346*** | | | | |
| | (0.070) | (0.070) | (0.070) | (0.070) | | | | |
| High marginal health benefit (= 1 if $\theta = 2$, Illness $C$) | 0.031 | 0.031 | 0.031 | 0.031 | | | | |
| | (0.031) | (0.031) | (0.031) | (0.031) | | | | |
| Physicians × Performance pay | | -0.322* | | -0.322* | | -0.055 | | -0.055 |
| | | (0.153) | | (0.154) | | (0.030) | | (0.030) |
| Other characteristics | No | No | Yes | Yes | No | No | Yes | Yes |
| Constant | 2.328*** | 2.228*** | 1.676*** | 1.676*** | 0.806** | 0.789*** | 0.652*** | 0.652*** |
| | (0.127) | (0.163) | (0.299) | (0.299) | (0.028) | (0.035) | (0.062) | (0.066) |
| $R^2$ | 0.608 | 0.610 | 0.622 | 0.623 | 0.043 | 0.047 | 0.095 | 0.099 |
| Observations | 3,006 | 3,006 | 3,006 | 3,006 | 3,006 | 3,006 | 3,006 | 3,006 |
| Subjects | 167 | 167 | 167 | 167 | 167 | 167 | 167 | 167 |

*Notes:* Ordinary Least Square (OLS) estimates are reported with robust standard errors clustered for subjects (in brackets). The reference category is CAP and the 'low severity of illness', $l = x$. 'Other characteristics' comprise a female dummy, a question each for the attitude towards altruism and competition from the European Values Study, and risk attitudes according to the German Socio-Economic Panel on the willingness to take risk in general, related to health (Dohmen et al., 2011) and one questions eliciting attitudes related to a patient's health. *** $p < 0.001$, ** $p < 0.01$, and *$p < 0.05$.

# D Robustness checks

In this section, we further test the robustness of our behavioral results from the main experiment and on the linkage between behavior and practice characteristics.

Table D.1: Ordinary least square regressions on physicians' health care service provision and quality of care under capitation and performance pay

| | A. Quantity $q$ | B. Relative quality, $\rho_{kl}$ |
|---|---|---|
| Model: | (1) | (2) |
| Performance pay | 0.058 | 0.068*** |
| | (0.074) | (0.013) |
| High bonus (= 1 if 20%-Bonus) | 0.086 | 0.019 |
| | (0.136) | (0.025) |
| Interm. severity of illness (= 1 if $l = y$) | 1.574*** | |
| | (0.088) | |
| High severity of illness (= 1 if $l = z$) | 3.061*** | |
| | (0.141) | |
| High marginal health benefit (= 1 if $\theta = 2$) | 0.016 | |
| | (0.038) | |
| Performance pay× Interm. severity | 0.240** | |
| | (0.088) | |
| Performance pay× High severity | 0.590*** | |
| | (0.121) | |
| Constant | 1.869** | 0.642*** |
| | (0.365) | (0.080) |
| Observations | 1.872 | 1.872 |
| Physicians | 104 | 104 |
| $R^2$ | 0.637 | 0.088 |

*Notes:* This table shows parameter estimates from Ordinary Least Square (OLS) estimation. Robust standard errors are shown in parentheses. The reference category is the 'mild severity of illness', $l = x$. Models contain controls for gender, practice years, attitudes towards altruism and competition from the European Values Study (European Values Study, 2016), and risk attitudes according to the German Socio-Economic Panel on the willingness to take risk in general, related to health (Dohmen et al., 2011) and one questions eliciting attitudes related to a patient's health. *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$.

Table D.2: Fractional probit models on physicians' quality of care under capitation and performance pay, average marginal effects

| Model: | (1) | (2) |
|---|---|---|
| Performance pay | 0.068*** | 0.068*** |
| | (0.013) | (0.013) |
| High bonus (= 1 if 20%-Bonus) | 0.000 | 0.019 |
| | (0.027) | (0.024) |
| Physicians' characteristics | No | Yes |
| Observations | 1.872 | 1.872 |
| Physicians | 104 | 104 |

*Notes:* Average marginal effects of fractional probit models are reported with robust standard errors clustered for subjects (in brackets). Models contain controls for gender, practice years, attitudes towards altruism and competition from the European Values Study (European Values Study, 2016), and risk attitudes according to the German Socio-Economic Panel on the willingness to take risk in general, related to health (Dohmen et al., 2011) and one questions eliciting attitudes related to a patient's health. *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$.

Table D.3: Logit regressions on crowding-out of patient-regarding behavior, average marginal effects

| | Full sample | | Sample restricted to benefit maximizers in CAP | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| High bonus (= 1 if 20%-Bonus) | 0.010 | 0.007 | 0.010 | 0.007 |
| | (0.045) | (0.042) | (0.046) | (0.043) |
| Intermediate severity (= 1 if $l = y$) | 0.014 | 0.020 | 0.014 | 0.020 |
| | (0.037) | (0.036) | (0.037) | (0.037) |
| High severity (= 1 if $l = z$) | -0.017 | -0.012 | -0.017 | -0.012 |
| | (0.043) | (0.042) | (0.043) | (0.042) |
| High marginal health benefit (= 1 if $\theta = 2$, Illness $C$) | -0.059* | -0.056* | -0.060* | -0.056* |
| | (0.028) | (0.027) | (0.028) | (0.027) |
| Physician characteristics | No | Yes | No | Yes |
| Observations | 936 | 936 | 503 | 503 |

*Notes:* The table shows marginal effects from logit regressions with robust standard errors clustered for subjects (in parentheses). The reference category is ‘mild severity’, $l = z$. Marginal health benefit is a dummy equal to 1 if $\theta = 2$ for illness $C$ and = 0 if $\theta = 1$ for illness $A$, $B$. Logit regressions yield very similar estimation results. *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$.

Table D.4: Quality of care ($\rho_{kl}$) and physician practice characteristics

| Model | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Performance_pay | 0.055*** | 0.055*** | 0.055*** | 0.055*** |
| | (0.014) | (0.014) | (0.014) | (0.014) |
| High bonus | 0.050* | 0.049 | 0.053* | 0.049 |
| | (0.030) | (0.030) | (0.030) | (0.030) |
| High annual profit | -0.095 | -0.121 | 0.014 | -0.054 |
| | (0.051) | (0.071) | (0.064) | (0.066) |
| City | -0.143** | -0.110** | -0.102** | -0.105*** |
| | (0.054) | (0.038) | (0.037) | (0.038) |
| Outer conurbation | -0.049 | -0.050 | -0.043 | -0.049 |
| | (0.054) | (0.034) | (0.034) | (0.034) |
| Share of SHI patients | -0.009 | -0.017 | -0.007 | -0.008 |
| | (0.013) | (0.017) | (0.013) | (0.014) |
| Revenue share from SHI patients | 0.003 | 0.002 | 0.015 | 0.003 |
| | (0.012) | (0.012) | (0.015) | (0.012) |
| Group practice (=1 if no. of physicians > 1) | 0.036 | 0.034 | 0.033 | 0.035 |
| | (0.027) | (0.028) | (0.027) | (0.028) |
| Time spent on SHI patients | 0.022 | 0.024 | 0.021 | 0.025 |
| | (0.012) | (0.012) | (0.012) | (0.016) |
| City × High annual profit | 0.088 | | | |
| | (0.075) | | | |
| Outer conurbation × High annual profit | -0.011 | | | |
| | (0.070) | | | |
| Share of SHI patients × High annual profit | | 0.017 | | |
| | | (0.022) | | |
| Revenue share from SHI patients × High annual profit | | | -0.028 | |
| | | | (0.019) | |
| Time spent on SHI patients × High annual profit | | | | 0.005 |
| | | | | (0.020) |
| Constant | 0.693*** | 0.681*** | 0.614*** | 0.635*** |
| | (0.109) | (0.111) | (0.102) | (0.107) |
| Observations | 1,566 | 1,566 | 1,566 | 1,566 |
| Physicians | 87 | 87 | 87 | 87 |

*Notes:* This table shows parameter estimates (fixed effects) from multilevel mixed-effects REML regressions. Standard errors are shown in parentheses. All models include subject-specific random effects, and all models control for the physicians' characteristics which comprise gender, practice years, a question each for the attitudes towards altruism and competition from the European Values Study (European Values Study, 2016), and risk attitudes according to the German Socio-Economic Panel on the willingness to take risk in general, related to health (Dohmen et al., 2011), as well as one questions eliciting attitudes related to a patient's health. *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$.

Table D.5: Quality of care ($\rho_{kl}$), physician practice characteristics, and bonus level

| Model | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Performance_pay | 0.055*** | 0.055*** | 0.055*** | 0.055*** |
| | (0.014) | (0.014) | (0.014) | (0.014) |
| High bonus | 0.008 | 0.112 | 0.124 | 0.080 |
| | (0.050) | (0.073) | (0.067) | (0.064) |
| High annual profit | -0.070* | -0.069* | -0.062* | -0.068* |
| | (0.029) | (0.028) | (0.029) | (0.029) |
| City | -0.121* | -0.099** | | |
| -0.094* | -0.103** | | | |
| | (0.053) | (0.038) | (0.038) | (0.038) |
| Outer conurbation | | -0.079 | -0.048 | -0.048 |
| -0.051 | | | | |
| | (0.044) | (0.034) | (0.034) | (0.034) |
| Share of SHI patients | -0.007 | 0.004 | -0.008 | -0.008 |
| | (0.013) | (0.018) | (0.013) | (0.013) |
| Revenue share from SHI patients | 0.004 | 0.002 | 0.015 | 0.002 |
| | (0.012) | (0.012) | (0.015) | (0.012) |
| Group practice (=1 if no. of physicians > 1) | 0.040 | 0.038 | 0.032 | 0.035 |
| | (0.028) | (0.028) | (0.028) | (0.028) |
| Time spent on SHI patients | 0.022 | 0.022 | 0.021 | 0.028 |
| | (0.012) | (0.012) | (0.012) | (0.017) |
| City × High bonus | 0.043 | | | |
| | (0.073) | | | |
| Outer conurbation × High bonus | 0.073 | | | |
| | (0.067) | | | |
| Share of SHI patients × High bonus | -0.022 | | | |
| | | (0.023) | | |
| Revenue share from SHI patients × High bonus | | | -0.026 | |
| | | | (0.021) | |
| Time spent on SHI patients × High bonus | | | | -0.011 |
| | | | | (0.020) |
| Constant | 0.667*** | 0.639*** | 0.617*** | 0.632*** |
| | (0.108) | (0.101) | (0.103) | (0.103) |
| Observations | 1,566 | 1,566 | 1,566 | 1,566 |
| Physicians | 87 | 87 | 87 | 87 |

*Notes:* This table shows parameter estimates (fixed effects) from multilevel mixed-effects REML regressions. Standard errors are shown in parentheses. All models include subject-specific random effects, and all models control for the physicians' characteristics which comprise gender, practice years, a question each for the attitudes towards altruism and competition from the European Values Study (European Values Study, 2016), and risk attitudes according to the German Socio-Economic Panel on the willingness to take risk in general, related to health (Dohmen et al., 2011), as well as one questions eliciting attitudes related to a patient's health. *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$.

Table D.6: Quality of care in high and low profit practices

| | A. Low-profit practices | | | | | B. High-profit practices | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model: | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| City | -0.114* | | | | | -0.032 | | | | |
| | (0.057) | | | | | (0.047) | | | | |
| Outer conurbation | -0.045 | | | | | -0.043 | | | | |
| | (0.055) | | | | | (0.039) | | | | |
| Share of SHI patients | | 0.008 | | | | | 0.020 | | | |
| | | (0.016) | | | | | (0.015) | | | |
| Revenue share from SHI patients | | | 0.026 | | | | | -0.001 | | |
| | | | (0.013) | | | | | (0.014) | | |
| Group practice (no. of physicians > 1) | | | | 0.044 | | | | | 0.035 | |
| | | | | (0.044) | | | | | (0.034) | |
| Time spent on SHI patients | | | | | 0.029* | | | | | 0.025* |
| | | | | | (0.013) | | | | | (0.013) |
| Constant | 0.760*** | 0.675*** | 0.580*** | 0.723*** | 0.595*** | 0.748*** | 0.701*** | 0.727*** | 0.688*** | 0.672*** |
| | (0.091) | (0.114) | (0.107) | (0.090) | (0.099) | (0.096) | (0.092) | (0.107) | (0.098) | (0.093) |
| Observations | 810 | 810 | 810 | 810 | 810 | 756 | 756 | 756 | 756 | 756 |
| Physicians | 45 | 45 | 45 | 45 | 45 | 42 | 42 | 42 | 42 | 42 |

*Notes:* This table shows parameter estimates from multilevel mixed-effects REML regressions. Standard errors are shown in parentheses. All models include subject-specific random effects. Share of SHI patients, revenue share from SHI patients, and time spent on SHI patients is given as proportionate share relative to patients with private insurance. All models include experimental controls and physician characteristics. Experimental controls comprise dummy variables for the High bonus condition (= 1 if 20%-Bonus) und performance pay. Physicians' characteristics comprise gender, years in practice and a question each for the attitude towards altruism and competition from the European Values Study(European Values Study, 2016), and risk attitudes according to the German Socio-Economic Panel on the willingness to take risk in general, related to health (Dohmen et al., 2011) and one questions eliciting attitudes related to a patient's health. *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$.

# Table D.7: Quality of care and regional differences

| | A. Rural | | | | | B. Outer conurbation | | | | | C. City | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model: | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) | (13) | (14) | (15) |
| High profit | -0.043 | | | | | -0.096 | | | | | 0.035 | | | | |
| | (0.038) | | | | | (0.051) | | | | | (0.080) | | | | |
| Share SHI | | 0.014 | | | | | -0.003 | | | | | 0.001 | | | |
| | | (0.014) | | | | | (0.022) | | | | | (0.026) | | | |
| Rev. share SHI | | | -0.005 | | | | | -0.002 | | | | | 0.023 | | |
| | | | (0.012) | | | | | (0.017) | | | | | (0.024) | | |
| Group practice | | | | 0.001 | | | | | 0.015 | | | | | 0.034 | |
| | | | | (0.031) | | | | | (0.043) | | | | | (0.074) | |
| Time with SHI | | | | | 0.011 | | | | | 0.004 | | | | | 0.022 |
| | | | | | (0.010) | | | | | (0.018) | | | | | (0.022) |
| Constant | 0.916*** | 0.895*** | 0.931*** | 0.856*** | 0.884*** | 0.619*** | 0.426*** | 0.437*** | 0.417*** | 0.428*** | 0.837*** | 0.832*** | 0.745*** | 0.830*** | 0.718*** |
| | (0.108) | (0.108) | (0.128) | (0.114) | (0.114) | (0.175) | (0.147) | (0.175) | (0.150) | (0.147) | (0.200) | (0.248) | (0.220) | (0.200) | (0.230) |
| Observations | 522 | 522 | 522 | 522 | 522 | 558 | 558 | 558 | 558 | 558 | 486 | 486 | 486 | 486 | 486 |
| Physicians | 29 | 29 | 29 | 29 | 29 | 31 | 31 | 31 | 31 | 31 | 27 | 27 | 27 | 27 | 27 |

*Notes:* This table shows parameter estimates from multilevel mixed-effects REML regressions. Standard errors are shown in parentheses. All models include subject-specific random effects. Share of SHI patients, revenue share from SHI patients, and time spent on SHI patients is given as proportionate share relative to patients with private insurance. All models include experimental controls and physician characteristics. Experimental controls comprise dummy variables for the High bonus condition (= 1 if 20%-Bonus) and performance pay. 'Physicians' characteristics' comprise gender, years in practice and a question each for the attitude towards altruism and competition from the European Values Study(European Values Study, 2016), and risk attitudes according to the German Socio-Economic Panel on the willingness to take risk in general, related to health (Dohmen et al., 2011) and one questions eliciting attitudes related to a patient's health. *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$.

# E Theoretical considerations on the crowding-out of patient-regarding behavior

More generally, we relax our assumption of a constant weight on the bonus payment. We still assume that all components of the utility are normalized in total, i.e., $\alpha + \beta + \gamma = 1$. Given this assumption, an increased weight on the discrete bonus payment reduces,*ceteris paribus*, the relative weight the physician attaches in total to the patient's health benefit as well as profit margin, i.e., $\alpha+\beta$. In the main text, we discussed the case of $\gamma = 1$ and Hypothesis 1 predicts that introducing performance reduces underprovision. This is the case since the intermediate-type physicians (area $B$) increase medical services provision to $q^* - 1$, while the physicians in areas $A$ and $C$ provide the quantity $q^{\text{Max}}$.

In the more general case, we now compare utility levels with and without performance pay: $U(q^* - 1)$ versus $U(q^{\text{Max}})$. Also, we solve for the "minimum" weight on the performance payment such that at least some physician previously in area $A$ switch to $B$ (i.e., provide a higher medical service provision and reduce underprovision). This minimum value for $\gamma$ is given by:

$$\tilde{\gamma} = \frac{\theta\,(q^* - 1 - q^{\text{Max}})^2}{2\,b_l\,(2\,\theta + q^{\text{Max}})} > 0.$$

Intuitively, the weight on the performance payment must be sufficient to compensate for the higher medical treatment provision. If the bonus payment $b_l$ is higher, $\tilde{\gamma}$ decreases. Consistent with Hypothesis 3, a higher bonus payment reduces underprovision since a lower minimum value for $\gamma$ is necessary. A higher marginal health benefit $\theta$ also decreases $\tilde{\gamma}$. Consistent with Hypothesis 2, area $A$ decreases to the benefit of area $B$, since it is easier to provide quantity $q^* - 1$. However, $\tilde{\gamma}$ is only the threshold between areas $A$ and $B$. The argumentation regarding areas $B$ and $C$ as given in the main text still holds such that the total effect on the level of underprovision is unambiguous.

Finally, higher severities of illness $l$ increase $q^*$ and $\tilde{\gamma}$. This implies, *ceteris paribus*, an increase in $\tilde{\gamma}$ and underprovision of care. However, higher severities also increase $\tilde{\gamma}$ through $b_l$ such that a countervailing effect exists. Consistent with Hypothesis 2, if the latter effect dominates, the effect of performance pay on underprovision increases with the patients' severity of illness.

Independent of this, an import implication of a strictly positive preference for the performance payment is the relative decrease in the relative preference for the patient's health benefit, as well as the profit margin, i.e., $\alpha + \beta$. Given our experimental design, we can only observe the chosen medical services quantities under CAP and CAP+P4P. If the chosen quantities under performance pay are only theoretically consistent with a decrease in $\alpha$, we consider this decrease in altruism as a *crowding-out* of motivation.

For which types of physicians this crowding-out of altruistic behavior is of relevance will be explained in the following. Hypothesis 1 states that low and high altruism types (areas $A$ and $C$) do not change the provided medical services quantity, since their provided quality $q^{\text{Max}}$ is independent of the performance payment. If those physicians reduce the provided quantity from CAP to CAP+P4P in the experiment, this can be explained by a decrease in $\alpha$.

In a more complex way, for the medium altruism types, an increase in the medical services quantity from CAP to CAP+P4P can actually be explained by decreasing altruism. This is the case if the weight on performance pay $\gamma$ is so high, that the observed positive medical services change in the experiment is solely driven by physicians who switch from $A$ to $B$ in order to earn barely the performance pay $b_\ell$. For these intermediate types, a positive weight on the performance pay can be seen as a "devaluation" of the weight on altruism. The motivational crowding-out effect crucially depends on comparative static results, as discussed for $\tilde{\gamma}$. A crowding out of altruistic behavior is less pronounced if $\tilde{\gamma}$ is lower. Consistent with our behavioral findings (Observation 1), this is the case for a high marginal health benefit.