Keser, Claudia; Peterle, Emmanuel; Schnitzler, Cornelius

Working Paper

Money talks: Paying physicians for performance

cege Discussion Papers, No. 173 [rev.]

Provided in Cooperation with:
Georg August University of Göttingen, Department of Economics

Suggested Citation: Keser, Claudia; Peterle, Emmanuel; Schnitzler, Cornelius (2014) : Money talks: Paying physicians for performance, cege Discussion Papers, No. 173 [rev.], University of Göttingen, Center for European, Governance and Economic Development Research (cege), Göttingen

This Version is available at:
https://hdl.handle.net/10419/103877

Center for European, Governance and Economic Development Research
cege

Discussion Papers

Number 173 – September 2014

MONEY TALKS - PAYING PHYSICIANS FOR PERFORMANCE

Claudia Keser, Emmanuel Peterle , Cornelius Schnitzler

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

ISSN: 1439-2305

Money talks - Paying physicians for performance

Claudia Keser, Emmanuel Peterle and Cornelius Schnitzler

October 01, 2014

Abstract: Pay-for-performance attempts to tie physician payment to quality of care. In a controlled laboratory experiment, we investigate the effect of pay-for-performance on physician provision behavior and patient benefit. For that purpose, we compare a traditional fee-for-service payment system to a hybrid system that blends fee-for-service and pay-for-performance incentives. Physicians are found to respond to pay-for-performance incentives. Approximately 89 percent of the participants qualify for a pay-for-performance bonus payment in the experiment. It follows that a patient treated under the hybrid payment system is significantly more likely to receive optimal treatment than a similar fee-for-service patient. Pay-for-performance generally tends to alleviate over- and under-provision of medical treatment relative to fee-for-service. Irrespective of the payment system, we observe unethical treatment behavior, i.e., the provision of medical services with zero benefit to the patient.

Keywords: Experimental economics; physician remuneration; pay-for-performance (*P4P*).

Acknowledgements: We thank Martin Schmidt for his support in programming the experiment software. Ren Ohinata gave valuable advice for the regression analysis. Thanks are also due to the organizers and participants of the 1st Workshop on Behavioral and Experimental Health Economics in Oslo for their helpful comments and suggestions.

1. Introduction

In healthcare, pay-for-performance (*P4P*) denotes payment systems, which attempt to tie provider payment to quality of care. Intended to improve medical outcomes, *P4P* schemes offer explicit financial incentives to healthcare providers (individuals, organizations or institutions) for meeting predetermined quality performance criteria for selected medical services (e.g., Emmert et al. 2011; Li et al. 2011; Maynard 2012). *P4P* is a payment concept for the reimbursement of healthcare providers, not a specific payment design (such as, for example, fee-for-service or capitation). It represents an attempt to make quality a direct component of the provider's financial compensation. In consequence, physicians are expected to increasingly shift resources toward quality improvement (Mullen et al. 2009).

Thus, *P4P* is primarily envisioned to improve quality of care. As a direct consequence of linking a portion of the provider payment to quality scores, increased adherence to evidence-based medicine could reduce the gap between the care provided and the care recommended (e.g., Institute of Medicine 2001; McGlynn et al. 2003; Hennig-Schmidt et al. 2011). Furthermore, *P4P* may, through improved quality of care, also reduce (the growth of) healthcare spending (Emmert et al.2011).

However, *P4P* could also lead to a number of unintended consequences. For instance, financial incentives (i.e. extrinsic motivation) for quality improvement could crowd out intrinsic motivation (Frey 1994). Moreover, incentivized dimensions of quality could receive increased attention at the expense of non-incentivized dimensions (the so-called multitasking problem, see Holmstrom & Milgrom 1991), which in turn could lead to a disruption of care.

Provider gaming presents another major concern (e.g., Hutchison 2008; Richards 2009). Clinical measures could be misreported in borderline cases in order to meet performance targets (Carey et al. 2009). Furthermore, patient selection may occur: providers could select patients based on their probability to meet *P4P* quality measures (e.g., Shen 2003, Rosenthal & Frank 2006; Casalino & Elster 2007), a practice that could result in diminished access to care (Epstein et al. 2004) and growing health disparities for already disadvantaged patients (Hong et al. 2010). In addition, *P4P* incentives could lead to another form of patient selection, the prioritization of patients for whom *P4P* incentives are available (Pines 2006).

The existing empirical literature on the effects of *P4P* is inconclusive. Several authors do not find any significant effect of *P4P* incentives on physician behavior and quality of care (e.g., Hutchison et al. 1996; Hillman et al. 1998; Strong et al. 2009). The bulk of the existing research, however, documents partial effects in that *P4P* incentives may lead to modest improvements in some but not all of the

evaluated quality measures (e.g., Rosenthal et al. 2005; Young et al. 2007; Glickman et al. 2007; Campbell et al. 2009; Li et al. 2011).

In our paper, we present a controlled laboratory experiment, in which we investigate the effect of *P4P* on physician provision behavior and patient benefit. We use the experimental framework introduced by Hennig-Schmidt et al. (2011) and extended by Keser et al. (2013) for the comparison of physician remuneration under *FFS* or *CAP*. For our research purpose, we implement two payment systems, a traditional fee-for-service payment system (*FFS* hereafter) and a hybrid payment system combining fee-for-service and pay-for-performance incentives (for simplicity, *P4P* hereafter). Under the payment system *FFS*, participants are paid separately for each unit of medical services provided. The hybrid payment system *P4P* blends traditional *FFS* payment with a tiered *P4P* bonus for reaching predetermined performance thresholds. The design of the *P4P* component implies that the bonus payment increases as the level of performance (measured against fixed targets) rises. In our experimental framework, a participant can earn a tiered *P4P* bonus under the hybrid payment system based on the number of patients who receive optimal care.

In the experiment, participants in the role of physicians decide on the quantity of medical services to be provided to each of the 40 virtual patients of a given patient list. Each physician treats 20 patients under *FFS* and 20 under *P4P* (within-subject design). The payment system is alternated once during the experiment, after 20 consecutive treatment decisions under a single payment system. To study potential ordering effects, we alternate the order in which the two payment systems are presented. In two of our experimental sessions, participants first face (20 consecutive patients under) *FFS* followed by (another 20 consecutive patients under) *P4P*. In two other sessions, the ordering of the payment systems is reversed.

Our experimental results show that, in our experimental model, physicians respond to the *P4P* incentives embedded in a hybrid payment system. Approximately 89 percent of the participants qualify for a *P4P* bonus payment in the experiment. The physicians' relative share of optimal treatment decisions is observed to be significantly larger under *P4P* than under *FFS*. A *P4P* patient is thus significantly more likely to receive optimal treatment than a similar *FFS* patient. *P4P* tends to alleviate over- and under-provision relative to traditional *FFS*. In addition, we observe unethical physician behavior (the provision of medical services with no benefit to the patient), irrespective of the payment system. This was impossible by design in the experiments by Hennig-Schmidt et al. (2011) and Keser et al. (2013).

This essay is organized as follows. Section 2 provides a brief literature review on the empirical evidence on the effectiveness of *P4P* incentives. Section 3 illustrates the design of our experiment. In Section 4 we present the experimental results. Concluding remarks are offered in Section 5.

2. Literature review

We are aware of one other study by Brosig-Koch et al. (2013b) that uses a controlled laboratory experiment to investigate the effects of introducing financial P4P incentives into *FFS* and *CAP* systems. Their experiment is also based on a variation of the experimental framework introduced by Hennig-Schmidt et al. (2011), as it has already been used in Brosig-Koch et al. (2013a). Their bonus system is different from the one that we use: physicians receive a bonus for each patient treated close to optimally, the level of this bonus depending on the severity of the patient's illness (and thus on the patient's optimal quantity of medical care). Similarly to our study, they find that physicians respond to the *P4P* incentives (both under *FFS* and *CAP*), implying a significant increase of benefit to the patients.

Empirical evidence on the effectiveness of *P4P* incentives is mixed. Most studies evaluate *P4P* programs adopted by private or public health plans in Canada, the United Kingdom or the United States. Several authors fail to detect any significant *P4P* effect on the quantity and quality of care. Hutchison et al. (1996), for example, show that Canadian physicians paid via capitation and a supplementary incentive payment for low hospital utilization rates do not have significantly lower hospital admission rates among their patients than physicians paid exclusively via *FFS*. Similarly, Hillman et al. (1998) and Hillman et al. (1999) do not observe a significant effect of *P4P* incentives coupled with performance feedback on physician compliance with cancer screening and pediatric preventive care guidelines in a U.S. Medicaid Health Maintenance Organization (HMO). A study by Strong et al. (2009) reports that the U.K. Quality and Outcomes Framework (QOF), the most comprehensive *P4P* program to date, does not lead to an improvement in the quality of ambulatory care.

On the contrary, there exist studies that give evidence that *P4P* may to some extent improve quality of care. Kouides et al. (1998) find that modest *P4P* incentives lead to a significant increase in the influenza immunization rate among the ambulatory elderly in the United States. Lindenauer et al. (2007) show that U.S. hospitals participating in a *P4P* program and public reporting show modestly larger improvements in quality of care over a two-year period than hospitals participating only in public reporting.

The bulk of the existing research documents partial effects of *P4P* incentives on provider behavior and quality of care. Fairbrother et al. (1999) and Li et al. (2011), for example, show that some of the investigated *P4P* incentives appear to work, while others do not. Fairbrother et al. (1999) report that a *P4P* bonus coupled with performance feedback leads to a significant increase in documented childhood immunization rates among New York City Medicaid enrollees, while enhanced fees for immunization services plus feedback are found to be ineffective. Observed improvements are primarily due to better documentation rather than changes in physician vaccination behavior. The authors argue that better documentation by itself is an important result of *P4P* since it constitutes a necessary first step toward improving actual immunization coverage. Analyzing a *P4P* program targeting family physicians and general practitioners in Ontario (Canada), Li et al. (2011) find that cumulative *P4P* bonus payments lead to modest increases in the utilization rates of four out of five incentivized preventive care services. In contrast, however, annual special payments for achieving minimum levels of service provision for certain medical services are found to be ineffective.

Numerous studies report partial effects in that *P4P* incentives may lead to modest improvements in some but not all of the evaluated quality measures. Beaulieu and Horrigan (2005), for example, observe significant improvement in several of the quality measures for Diabetes within a small sample of physicians who self-selected into a U.S. Managed Care Organization (MCO) *P4P* scheme. Comparing Californian physician groups participating in a *P4P* program with non-participating Pacific Northwest physician groups, Rosenthal et al. (2005) report that Californian physician groups show significantly greater improvement in only one of the three investigated clinical quality scores. *P4P* incentives lead to a significantly larger improvement in the quality score for cervical cancer screening, but not for mammography and hemoglobin A1c testing.

Similar empirical evidence is presented by Young et al. (2007), Glickman et al. (2007), Rosenthal et al. (2009) and Campbell et al. (2009). Young et al. (2007) show that a U.S. *P4P* scheme financed via withholds leads to a modest improvement in one of four quality measures for diabetic patients. In that scheme, physicians are at risk of losing a portion of their earned payments based on their performance relative to their peers. Glickman et al. (2007) observe slightly higher rates of improvement in two of six acute myocardial infarction quality measures for U.S. hospitals participating in a voluntary *P4P* scheme. The rate of improvement in in-hospital mortality, however, does not differ significantly for participating and non-participating hospitals. Rosenthal et al. (2009) document that a US$ 100 bonus for (both patients seeking and providers providing) timely and comprehensive prenatal care is found to reduce the probability of neonatal intensive care unit admissions and associated with lower healthcare spending in the first year of life. Participation in the scheme, however, is reported to have no significant effect on low birth weight. Campbell et al. (2009)

find that the U.K. QOF, in the short run, resulted in a significant acceleration in the improvement in quality for asthma and diabetes, but not for coronary heart disease. Rates of improvement, however, were found to be unsustainable in the intermediate term.

The existing literature also offers evidence of unintended consequences. Sicsic et al. (2012) report a potentially negative relationship between intrinsic and extrinsic motivation for French general practitioners noting that policy makers should be aware that *P4P* incentives could lead to a corrosion of intrinsic motivation. Campbell et al. (2009) and Doran et al. (2011) document that *P4P* incentives lead to a decline in the quality of non-incentivized dimensions of care. Hong et al. (2010) show that relative physician clinical performance rankings frequently employed in *P4P* programs can be confounded by patient panel characteristics. Evaluating relative performance rankings for physicians working for a renowned academic primary care system in the U.S. (Massachusetts General Hospital), the authors observe that physicians who treat a greater proportion of underinsured, minority and non-English speaking patients are found to receive lower relative physician performance rankings.

Several studies offer evidence for health-care providers showing gaming behavior. Carey et al. (2009) discover a reporting bias in clinical indicators, noting that patients in the U.K. QOF are significantly more likely to narrowly meet than narrowly miss a *P4P* performance target. Shen et al. (2003) observe that performance-based contracting leads to patient selection. Under performance-based contracting, the level or continuation of funding hinges on patient treatment outcomes. The authors find that non-profit providers of substance abuse treatment in the U.S. treat less severely ill patients and avoid more severe cases in order to improve overall performance. Gravelle et al. (2010) and Dalton et al. (2011) report gaming of exception reporting under the QOF. Exception reporting is an instrument that allows the exclusion of patients from certain QOF quality indicators based on broad clinical criteria. Gravelle et al. (2010) present evidence that both provider and patient characteristics affect exception reporting. Appropriate and permissible exception reporting, however, should exclusively be determined by patient characteristics. Dalton et al. (2011) discover significantly higher rates of exception reporting for already disadvantaged patient groups such as older patients and ethnic minorities. In consequence, *P4P* could lead to greater health disparities since exception reported patients are less likely to achieve treatment targets.

3. The experiment

Each participant in our experiment acts as a physician and treats 40 virtual patients (presented in succession). To study the impact of *P4P* on physician provision behavior and patient benefit, we implement two payment systems, a *FFS* system and a hybrid payment system combining *FFS* and *P4P*

incentives. Under the payment system *FFS*, participants are paid separately for each unit of medical services provided. *P4P*, the hybrid payment system, blends traditional *FFS* payment with a *P4P* incentive, which guarantees bonus payments for reaching absolute performance targets.

The patients present themselves in two sequences to the physician. Each of the two sequences comprises 20 virtual patients, who present themselves sequentially and whose treatment is paid for under the condition of one of the two implemented payment systems. In the experiment, the payment system is alternated once, after the completion of the first sequence.

To study potential ordering effects, we alternate the order in which the two sequences (and thus payment systems) are presented. Twenty-six participants first encounter a sequence of twenty virtual patients under *FFS* followed by a sequence of twenty virtual patients under *P4P*. The order of the sequences is reversed for 26 other participants.

Each participant determines the quantity q $(with\ q \in \{0, 1, \dots, 20\})$ of medical services to be provided to each of the patients in the experiment. Only entire units of medical services can be provided to individual patients. Treatment choices impact both physician profit and patient benefit.

Virtual patients are characterized by the three attributes, *payment system*, *treatment preference* and *illness*. The first attribute is the payment system, either *FFS* or *P4P.*The second attribute is treatment preference. We distinguish here between four patient types. Each type is characterized by a treatment preference expressed in a patient benefit function, which is either $B_1(q)$, $B_2(q)$, $B_3(q)$ or $B_4(q)$. The patient benefit function $B_i(q)$ describes the benefit that a patient of type i $(i \in \{0, 1, 2, 3, 4\})$ draws from treatment quantity q and is measured in monetary terms (in *Experimental Currency Unit*, *ECU*). Table 1 and Figure 1 present the four patient benefit functions. As in the seminal paper by Hennig-Schmidt et al. (2011), the four different benefit functions in the experiment reflect a heterogeneous patient population: patients respond differently to the quantity of treatment, independently of their respective illness. Note that the same benefit functions are implemented for the characterization of patients in the first and the second sequence of the experiment.

As can be seen in Figure 1, each benefit function $B_i(q)$ has an interior global optimum at a q_i^*, which determines the treatment preference, i.e., the optimal amount of medical care for patient type i. Specifying a global optimum in the interior of the action space sets a benchmark for the right quantity of medical care allowing us to observe over- and under-provision of medical care. It also enables us to pay physicians for performance. Note that while benefit functions differ across patient types, two of the four benefit functions (for types 2 and 4) exhibit the same treatment preference

($q_2^* = q_4^* = 8$). Patient type 1 has a relatively low treatment preference ($q_1^* = 4$), while patient type 3 has a relatively high treatment preference ($q_3^* = 12$).

In contrast to the experimental design by Hennig-Schmidt et al. (2011), we implement two benefit functions ($B_1(q)$ and $B_4(q)$) that allow for the possibility of patients being harmed by the over-provision of medical services. In other words, the benefit functions for patient type 1 and 4 are designed in such a way that patients suffer from excessive medical services by receiving a negative patient "benefit". Excessive medical services means more than 14 units of medical services for patient type 1 and more than 16 for patient type 4.

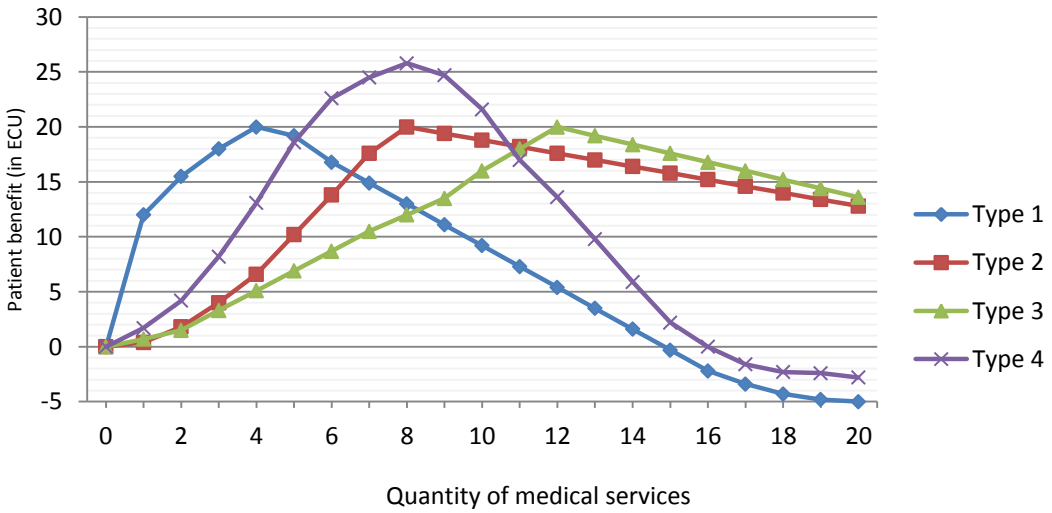

Figure 1: Patient benefit functions

The third attribute is illness. Each patient suffers from one out of five potential illnesses, A, B, C, D or E. Illness impacts the fee-for-service payment and thus physician compensation (in ECU) under both payment systems. We devise five fee-for-service fee functions, one for each illness, and implement the same functions under *FFS* and *P4P*. For each illness, payment increases with the quantity of medical services provided (see Table 2).

Each combination of the three patient attributes (payment system, treatment preference and illness) represents an individual patient in the experiment. As explained above, patients are subdivided into two sequences with different payment systems. This implies that physicians face twenty patients under the same payment system in a sequence. The two sequences consist of similar patients with respect to treatment preference and illness, and differ exclusively in the payment system used to pay physicians. Treatment choices affect patient benefit. We assume that individual patients enjoy full insurance coverage and accept any quantity of medical services.

Treatment choices also impact physician profit (in ECU). In the experiment, participants face a convex cost function (as in the experiment by Hennig-Schmidt et al. 2011) given by $c(q_j) = 0.2\, q_j^2$, where $q_j$is the quantity of medical services provided to patient j. The cost function is independent of the patient type, illness or payment system under which the patient is treated (see Table 3).

FFS fee functions are designed in such a way that four out of the five resulting profit functions exhibit a global profit maximum in the interior of the action space (see Table 4 or Figure 2). Illness B (C; D; E) shows a global profit maximum at 15 (10; 2; 18) units of medical services. Recall that the treatment comprises quantities between 0 and 20. The profit function for illness A constitutes an exception in that it exhibits a global profit maximum at 20 units of medical services. Figure 2 provides a graphical representation of the FFS profit functions.

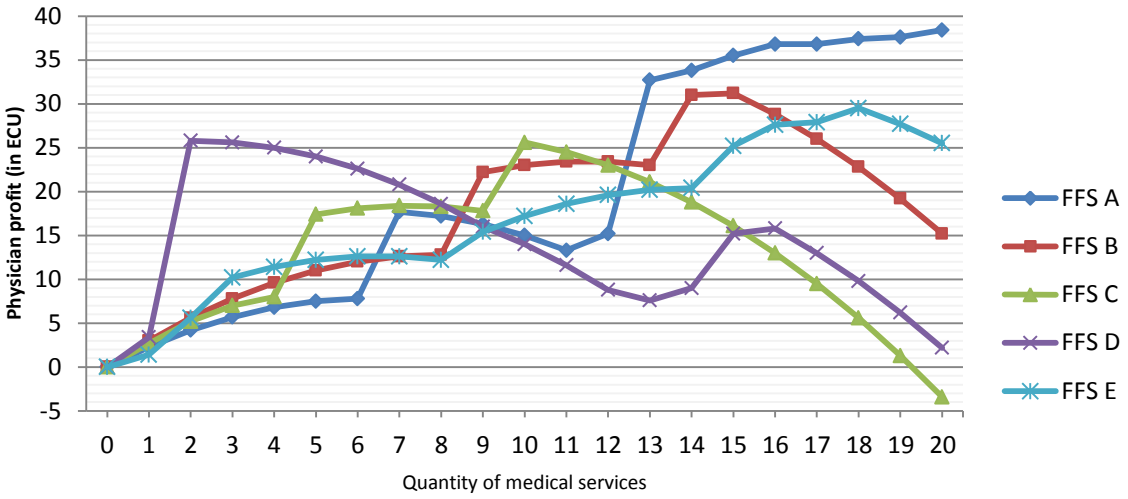

Figure 2: FFS profit functions

We noted earlier that for each illness, FFS payment increases along the quantity of medical services provided. Profits, however, do not continuously increase along the quantity of services provided, not even prior to reaching the profit maximum. For illnesses A, B, C and E, profit functions display instances of negative marginal profit for additional units of medical services before reaching the profit maximum. Only the payment function for illness D always exhibits positive marginal profit prior to reaching the global profit maximum at q = 2. The payment and patient benefit functions are designed to create a tradeoff between the physician's maximum profit and the patient's maximum benefit. The FFS profit-maximizing quantity for a patient exceeds the right amount of care for 15 of the 20 patients in a sequence. For the remaining five patients, the profit maximizing quantity undercuts the optimal amount of medical care.

Table 1: Benefit functions for patient type 1 to 4 (in ECU)

Patient type	Quantity of medical services																				
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Type 1	0.00	12.00	15.50	18.00	20.00*	19.20	16.80	14.90	13.00	11.10	9.20	7.30	5.40	3.50	1.60	-0.30	-2.20	-3.40	-4.30	-4.80	-5.00
Type 2	0.00	0.40	1.80	4.00	6.60	10.20	13.80	17.60	20.00*	19.40	18.80	18.20	17.60	17.00	16.40	15.80	15.20	14.60	14.00	13.40	12.80
Type 3	0.00	0.70	1.50	3.30	5.10	6.90	8.70	10.50	12.00	13.50	16.00	18.00	20.00*	19.20	18.40	17.60	16.80	16.00	15.20	14.40	13.60
Type 4	0.00	1.70	4.20	8.20	13.10	18.60	22.60	24.50	25.80*	24.70	21.60	17.00	13.60	9.80	5.90	2.20	0.00	-1.60	-2.30	-2.40	-2.80

* Maximum patient benefit for a patient of the given type

Table 2: FFS fee functions for illness A to E (in ECU)

Illness	Quantity of medical services																				
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	0.00	2.50	5.00	7.50	10.00	13.50	15.00	27.50	30.00	32.50	35.00	37.50	44.00	66.50	73.00	80.50	88.00	94.60	102.20	109.80	118.40
B	0.00	3.20	6.40	9.60	12.80	16.00	19.20	22.40	25.60	38.40	43.00	47.60	52.20	56.80	70.20	76.20	80.00	83.80	87.60	91.40	95.20
C	0.00	2.90	6.00	8.80	11.20	22.40	25.30	28.20	31.10	34.00	45.60	48.70	51.80	54.90	58.00	61.10	64.20	67.30	70.40	73.50	76.60
D	0.00	3.60	26.60	27.40	28.20	29.00	29.80	30.60	31.40	32.20	34.00	35.80	37.60	41.40	48.20	60.20	67.00	70.80	74.60	78.40	82.20
E	0.00	1.60	6.40	12.00	14.60	17.20	19.80	22.40	25.00	31.60	37.20	42.80	48.40	54.00	59.60	70.20	78.80	85.70	94.30	99.90	105.50

Table 3: Cost function (in ECU)

	Quantity of medical services																				
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Cost c (q)	0.00	0.20	0.80	1.80	3.20	5.00	7.20	9.80	12.80	16.20	20	24.20	28.80	33.80	39.20	45.00	51.20	57.80	64.80	72.20	80.00

Table 4: FFS physician profit (in ECU)

	Quantity of medical services																				
Illness	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	0.00	2.30	4.20	5.70	6.80	7.50	7.80	17.70	17.20	16.30	15.00	13.30	15.20	32.70	33.80	35.50	36.80	36.80	37.40	37.80	38.40*
B	0.00	3.00	5.60	7.80	9.60	11.00	12.00	12.60	12.80	22.20	23.00	23.40	23.40	23.00	31.00	31.20*	28.80	26.00	22.80	19.20	15.20
C	0.00	2.70	5.20	7.00	8.00	17.40	18.10	18.40	18.30	17.80	25.60*	24.50	23.00	21.10	18.80	16.10	13.00	9.50	5.60	1.30	-3.40
D	0.00	3.40	25.80*	25.60	25.00	24.00	22.60	20.80	18.60	16.00	14.00	11.60	8.80	7.60	9.00	15.20	15.80	13.00	9.80	6.20	2.20
E	0.00	1.40	5.60	10.20	11.40	12.20	12.60	12.60	12.20	15.40	17.20	18.60	19.60	20.20	20.40	25.20	27.60	27.90	29.50*	27.70	25.50

Table 5: Pay-for-performance bonus schedule (in ECU)

	Quantity of optimally treated patients			
	0 – 9	10 – 12	13 – 15	16 – 20
Bonus Payment	0.00	130.00	165.00	215.00

Under *P4P*, participants face a payment system that blends traditional FFS with a tiered *P4P* bonus based on physician performance. The above presented FFS fee functions are implemented in both payment systems. Physicians can additionally qualify for a tiered bonus payment (in ECU) based on the number of *P4P* patients, who receive optimal care in the experiment.

Optimal care is defined as the right amount of medical services in the experiment. Patient benefit, thus patient type but not illness, determines the optimal quantity of treatment for each individual patient. In reality, physicians can refer to widely available clinical treatment guidelines such as the indicative interdisciplinary guideline *Prevention, Diagnosis, Therapy, and Follow-up of Lung Cancer* by the German Respiratory Society and the German Cancer Society (2010) for the optimal course of treatment.

The objective of our *P4P* scheme is to increase the provision of optimal care and reduce over- and under-provision of medical services (relative to *FFS*). To qualify for a bonus payment, physicians have to provide optimal care to a threshold of patients. The amount of the bonus payment increases with the threshold.

In the experiment, we set three *P4P* performance thresholds: participants have to meet a minimum performance threshold by providing optimal care to at least ten *P4P* patients in order to qualify for a bonus payment under *P4P*. The corresponding *P4P* bonus payment is 130 ECU. Meeting higher performance thresholds leads to higher bonus payments: an experimental physician providing optimal care to at least 13 patients nets 165 ECU. A bonus payment of 215 ECU will be paid if more than 15 patients receive optimal care under the hybrid payment system. Table 5 provides an overview of the performance thresholds and the corresponding *P4P* bonus payments. The payments have been chosen such that a profit-maximizing, fully informed physician would treat those ten patients optimally, for which optimal treatment requires the lowest profit reduction to the physician (see Table B.3 for a ranked list of the necessary profit reductions).

The FFS component and the bonus of the hybrid payment system offer conflicting financial incentives: physicians face a trade-off between maximizing profit (per patient) by deviating from the optimal quantity of the individual patient and providing optimal care to at least half of the patients in order to qualify for a bonus payment. Under *P4P*, physicians may forego FFS profits to qualify for a *P4P* bonus.

The *P4P* incentive in our experiment is layered onto an existing payment system (i.e. a FFS payment system), a predominant approach for the design of P4P schemes in reality (e.g., Young et al. 2007). Moreover, we set a single type of performance standard by applying absolute performance targets.

We do not use relative performance rankings; in other words, we do not measure physician performance relative to peer performance (refer to e.g., Hahn 2006 for *P4P* typology). The bonus pool is uncapped, which implies that physicians qualifying for a *P4P* bonus are guaranteed to receive the amount for which they have met the respective performance threshold. Respective thresholds and corresponding bonus payments are known to each participant in the experiment. These design features imply that bonus payments are non-competitive in nature. A physician's individual treatment behavior determines the size of the physician's bonus payment; peer behavior has no impact on the *P4P* bonus in our design.

We conducted the experiment in the Göttingen Laboratory of Behavioral Economics at the Georg-August-University Göttingen, Germany, in July and August 2012. The software was programmed in z-Tree (Fischbacher 2007). Fifty-two (18 male and 34 female) medical students enrolled in the medical degree program of the University of Göttingen participated in the experiment. Students volunteered to take part in the experiment after having been recruited via the medical school's official mailing list. Each participant took part in the experiment once.

We conducted our experiment in several sessions and the procedure was the same in each session. Experimental sessions differed in the order in which participants encountered the sequences of patients and thus payment systems.

Before the experiment, participants and the experimenter gather in a conference room where each participant receives a white coat (to emphasize the medical context of the experiment) and a randomly assigned participation number corresponding with one of the isolated working stations featured in the laboratory. To further emphasize the medical context, the participants found posters of the Aesculap stick attached to the wall above their working stations. The isolation makes visual contact and communication between participants impossible. Copies of the instructions (including information on the payment system, which participants face in the first sequence of patients) are distributed and read out to the participants. From this moment on, participants are required to cease any communication with each other and they are instructed not to publicly raise any questions regarding the instructions.

After reading the instructions, participants get seated at their respective working stations and commence with a programmed questionnaire testing the comprehension of the experimental instructions. Participants have ample opportunity to clarify any open question concerning the instructions; the experimenter is at hand to privately resolve any question a participant might have. The experiment begins, once all participants have successfully completed the questionnaire by correctly answering all questions.

In the beginning of an experimental session, each participant chooses among three patient charities, the German Cancer Society, the German Multiple Sclerosis Society and the German Parkinson Society. As communicated in the instructions, the aggregate patient benefit created by each of the participants will be paid to the organization of her or his respective choice. Through the funding of medical research the money donated will serve real patients. This actual paying out of the patients, who participate only virtually in the experiment, should encourage the participants in the role of physicians to take the patient benefit into account.

Each participant assumes the role of a physician in the experiment. In each sequence, participants provide medical services to 20 virtual patients, characterized by illness and patient type, under a specific payment system, either *FFS* or *P4P*. Each virtual patient is presented as a table providing, for each possible quantity of medical services, physician profit and patient benefit associated with the patient's characteristics. Physicians have no information on the specifics of illnesses and patient types. They thus make each treatment decision based on the profit-benefit table for the patient under consideration. After all participants have completed their treatment decisions for the first sequence of patients, a new set of instructions is distributed detailing the payment system effective for the second sequence of patients. Participants receive plenty of time to read the new set of instructions by themselves and clarify any questions they might have in private with the experimenter before the experiment continues. The instructions for both parts of the experiment are in Appendix A.

For each participant, the sum of physician profits and the sum patient benefits (from the two sequences) are calculated separately and converted in €, applying a conversion factor of 0.01€ per ECU. Profit is paid out privately, in cash, in addition to a 3.00€ show-up fee at the end of the experiment. The monetary patient benefit which is assigned to the same charity is pooled and donated via money transfer.

According to subject availability, we conducted four experimental sessions, each lasting approximately 90 minutes. We collected 26 independent observations each for the order *FFS* / *P4P* and the order *P4P* / *FFS*. Participants earned an average amount of 12.70€. The minimum payment was 12.70€, the maximum 15.40€. Payments to the three charities summed up to 391.90€. The German Cancer Society was the charity of choice for 27 participants and received a donation in the amount of 197.90€; the German Multiple Sclerosis (MS) Society, selected by 16 participants, received 126.50€, while the German Parkinson Society was selected by nine participants and received 67.50€ (see Table 6). Digital copies of the respective receipts were provided to all participants.

Following the experiment, we collected valuable information regarding participants' personal situation through a short questionnaire. About one participant out of three states that close relatives have experienced medical emergency in the last twelve months. About three participants out of four declare severe or/and chronic illness among relatives. Personal experience appears to affect the choice made among charities, as 71.43 percent of participants declaring the occurrence of cancer among close relatives have chosen the German Cancer Society as recipient. In the same way, 66.67 percent of participants whose relatives have suffered from multiple sclerosis and 50 percent of participants whose relatives have suffered from Parkinson's disease have chosen the corresponding charity as recipient.

Table 6: Payments to charities in € (number of participants)

Order	Total	Cancer Society	MS Society	Parkinson Society
FFS / *P4P*	198.60 (26)	98.10 (13)	54.40 (7)	46.10 (6)
P4P / *FFS*	193.30 (26)	99.80 (14)	72.10 (9)	21.40 (3)
Both orders	391.90 (52)	197.90 (27)	126.50 (16)	67.50 (9)

The post-experimental questionnaire also includes questions aiming at eliciting participants' personality traits. We collect this information in order to investigate individual differences in attitude towards patient's benefit and the occurrence of unethical conduct. These questions are directly inspired from the German Socio-Economic Panel (SOEP)[1] and are reported in Table C.1 (Appendix C).

4. Results

In the following, we present our experimental results. The non-parametric analysis is based on 52 independent observations, 26 for each order, in which the sequences are presented. All test results, generated with the statistical analysis software *Stata*, are two-sided. To lead off the analysis, we investigate any ordering effects that may arise from the order in which the sequences are presented (Subsection 4.1). Since we observe no significant ordering effect, we may pool our data. Descriptive statistics of the combined data are presented in Subsection 4.2. In Subsection 4.3, we investigate

[1] Individual questionnaire 2005/2010, see http://www.diw.de/en/diw_02.c.222729.en/questionnaires.html.

over- and under-provision under *FFS* and *P4P*. Differences in physician provision behavior across the two payment systems are presented in Subsection 4.4. In Subsection 4.5, we analyze the effect of the payment systems on patient benefit. We present evidence of unethical treatment behavior in Subsection 4.6. In conclusion, in Subsection 4.7 we investigate the impact of experimental participants' personal characteristics on service provision.

4.1 Potential ordering effects

We are particularly interested in whether the order, in which physicians encounter the two payment systems under consideration, affects treatment behavior. Table 7 provides, for each order (*FFS*/*P4P* and *P4P*/*FFS*), the average quantity of medical services provided to patients of the same payment system. We find average treatment quantities to be practically unaffected by the order, in which physicians face the payment systems.

To gain statistical evidence on potential ordering effects, we consider the quantities that individual physicians provide, on average, to *FFS* and *P4P* patients in the two orders, in which the payment systems are presented. A Wilcoxon-Mann-Whitney-U test comparing average quantities provided to *FFS* patients in *FFS*/*P4P* with the average quantities provided to *FFS* patients in *P4P*/*FFS* reveals no significant difference in treatment behavior (p=0.2798). The same is true for *P4P* patients (p=0.7693).

We also conduct a statistical analysis based on the individual treatment decisions for *FFS* and *P4P* patients in each order. A Wilcoxon-Mann-Whitney-U test comparing, for each patient, the individual treatment decisions in *FFS*/*P4P* with those in *P4P*/*FFS* shows no significant ordering effect for 19 of the 20 *FFS* patients ($p \geq 0.0563$) and for 18 of the 20 *P4P* patients ($p \geq 0.0925$; refer to Table B.1 - Appendix B for the respective p-values).[2] In summary, we do not observe a significant ordering effect for 37 out of 40 patients in the experiment. This is in keeping with the above result on the overall physician behavior.

Table 7: Average quantity of treatment

	FFS/*P4P*		*P4P*/*FFS*	
	FFS	*P4P*	*P4P*	*FFS*
Average	9.77	9.06	9.05	9.74
Median	9.00	8.00	8.00	9.00
SD	3.90	4.05	4.01	4.36

[2]We identify, however, a significant ordering effect for one *FFS* (4D; p = 0.0020; Wilcoxon-Mann-Whitney-U test) and two *P4P* patients (2A; p = 0.0284; 3A; p = 0.0278; Wilcoxon-Mann-Whitney-U test)

A major concern of our research is the potential impact of *P4P* incentives on the provision of optimal care. We are thus particularly interested in whether participants who first face *P4P* followed by *FFS* provide optimal care to more *FFS* patients in the experiment than those who face the reverse order of payment systems. Consequently, we compare the individual physicians' relative shares of optimal treatment decisions for *FFS* patients observed in *FFS*/*P4P* with those in *P4P*/*FFS*. A Wilcoxon-Mann-Whitney-U test reveals no significant ordering effect ($p = 0.0893$). Similarly, we do not find a significant difference in the relative share of optimal treatment decisions for *P4P* patients across orders ($p = 0.5782$).

Conclusion 4.1: *We do not observe any significant ordering effect in our experiment. Consequently, we shall pool the data for further statistical evaluation. Subsequent results are thus based on the pooled data of 52 independent observations.*

4.2 Descriptive statistics of the pooled data

Physicians provide, on average, 9.75 (median 9.00; SD 4.14) units of medical services to *FFS* and 9.05 (median 8.00; SD 4.03) units to *P4P* patients. The average optimal quantity is eight units of medical services, irrespective of the payment system. This suggests that, on the aggregate, patients are over-served and that they are less over-served under *P4P* than under *FFS*. The latter is also reflected in that we observe an average patient benefit of 18.17 (median 20.00; SD 5.78) ECU for *FFS* patients and 19.28 (median 20.00; SD 5.99) ECU for *P4P* patients. Physicians show a higher frequency of optimal treatment decisions under *P4P*: approximately 73.5 percent of the treatment decisions for *P4P* patients coincide with the right amount of care compared to 41.2 percent of the treatment decisions for *FFS* patients.

On average, physicians earn 21.11 ECU (median 20.80; SD 7.40) per *FFS* patient and 18.50 ECU (median 18.30; SD 7.67) per *P4P* patient under the FFS component of the hybrid payment system. At the same time, we observe that the average total physician profit under *P4P* exceeds the average total physician profit under *FFS* (by 29.5 percent) because of the additional bonus payment under *P4P*. Physicians earn on average 547.15 ECU (bonus included) from treating *P4P* patients and 422.24 ECU from treating *FFS* patients. Table 8 provides an overview of these summary statistics.

In the experiment, approximately 88.5 percent of the physicians qualify for a bonus payment under *P4P*. Thirty-six physicians earn a bonus in the amount of 215.00 ECU (for providing optimal care to at least 16 *P4P* patients), five physicians qualify for a bonus in the amount of 165.00 ECU (for 13 to 15

optimally treated *P4P* patients) and five physicians earn a bonus of 130.00 ECU (for 10 to 12 optimally treated *P4P* patients). Six physicians do not qualify for any bonus payment under the hybrid payment system. *P4P* bonus payments in the experiment add up to 9215.00 ECU.

Table 8: Summary statistics

	Quantity		Patient benefit[1]		Profit per patient[1]		Profit per physician[1;2]	
	FFS	*P4P*	*FFS*	*P4P*	*FFS*	*P4P*	*FFS*	*P4P*
Average	9.75	9.05	18.17	19.28	21.11	18.50	422.24	547.15
Median	9.00	8.00	20.00	20.00	20.80	18.30	420.95	559.4
SD	4.14	4.03	5.78	5.99	7.40	7.67	83.79	59.36
Percent[3]			41.2	73.5	15.5	8.6		

[1] in ECU.
[2]Total profits per physician, resulting from the treatment of patients of the same payment system (including bonus payments)
[3]Percent of benefit-maximizing individual treatment decisions in the case of patient benefit; percent of profit maximizing individual treatment decisions in the case of profit per patient.

In total, physicians bill services in the amount of 45,195.70 ECU under *FFS* and 39,637.60 ECU under (the *FFS* component of) the hybrid payment system *P4P*. The observed difference in the total amounts billed covers approximately 60.3 percent (or 5558.10 ECU) of the total bonus payout under *P4P*. New funds have to pay for the remainder. Total spending on physician services under *P4P* (48852.60) is approximately 8.1 percent higher than under *FFS*.

4.3 Over- and under-provision

Figure 3 shows for each individual patient, characterized by a combination of treatment preference and illness, the average quantity of medical services provided by the experimental physicians under FFS on the one hand and under P4P on the other. Moreover, it indicates, for each patient, the *FFS* profit-maximizing quantity and the optimal treatment quantity. Figure 3 suggests that, on the aggregate, physicians over-serve most of the patients in the experiment. Over- or under-provision appears to be affected by the patient's illness (see also Appendix B, Table B.2, exhibiting, for each patient, the average quantity of medical services provided along with the mean deviation from the respective optimal quantity). Note that a tendency to over-provide FFS patients has also been observed by Hennig-Schmidt et al. (2011) and Keser et al. (2013).

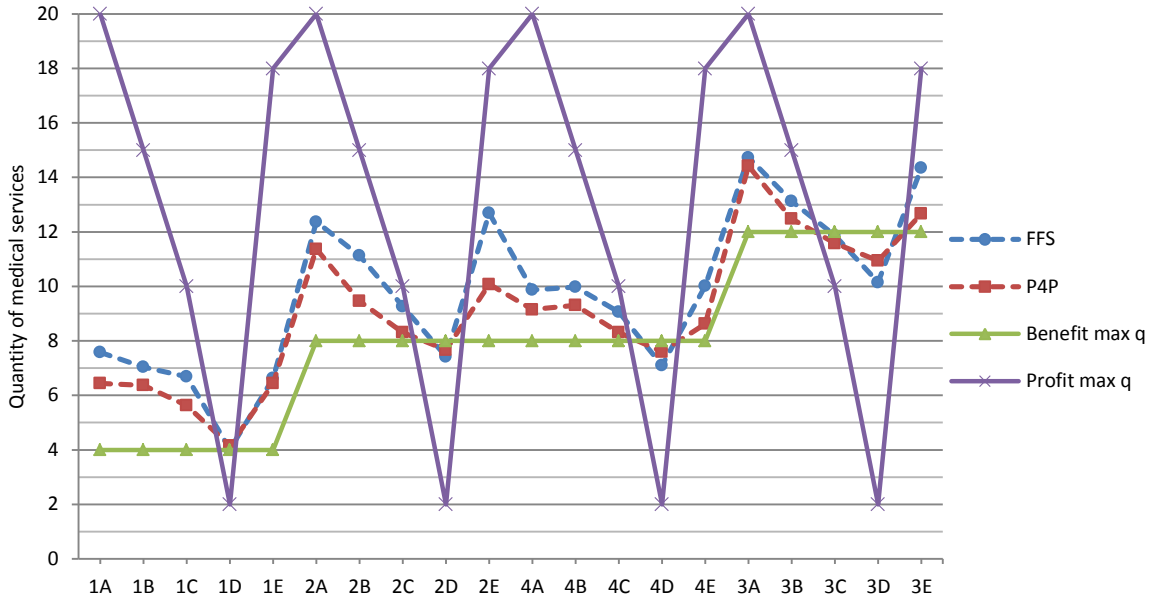

Figure 3: Average quantity of medical services per individual patient characterized by patient type [1-4] and illness [A – E]

***FFS*:** We observe that 16 of the 20 *FFS* patients receive, on average, more medical services than optimal. For 15 of these patients, the profit-maximizing quantity exceeds the patient's optimal quantity of treatment, while for one (patient 1D) the profit maximizing quantity (two units) lies below the patient's optimal treatment quantity (four units). The observed over-provision for *FFS* patient 1D is only minimal (the average quantity provided is 4.02 units) and only 13.5 percent of the individual treatment decisions deviate from the optimal quantity of treatment, which can be explained by a financial impact of deviating from the profit-maximizing quantity and providing the optimal quantity of treatment that is hardly noticeable (approximately 0.01€ in foregone physician profits).

Four (of the 20) *FFS* patients (characterized by the patient attributes 2D, 4D, 3C and 3D) in the experiment receive, on average, fewer than optimal services. For these patients, the optimal quantity of treatment lies above the profit maximizing quantity. *FFS* patient 3C is only marginally under-served (the mean deviation from the optimal quantity of treatment is – 0.13 units) and only 19.2 percent of the individual treatment decisions deviate from the optimal quantity of treatment. For this patient, the difference in physician profit resulting from the provision of the optimal quantity of treatment (twelve units) and the profit-maximizing quantity (ten units) is small (2.60 ECU or less than 0.03€).

***P4P*:** Even though the provision of optimal care is incentivized, we do find over- and under-provision under *P4P*. We observe that for 16 of the 20 *P4P* patients, the average quantity of treatment exceeds the corresponding optimal quantity of care. These patients are similar to those over-served under *FFS* (with respect to treatment preference and illness). For three of the 16 over-served *P4P* patients, the extent of the over-provision is observed to be marginal: over-provision is hardly detectable for *P4P* patient 1D and negligible for *P4P* patients 2C and 4C. For these patients, the quantity of medical services provided deviates from the corresponding optimal quantity of treatment in only 11.5, 15.4 and 11.5 percent of the treatment decisions, respectively.

Four *P4P* patients in the experiment receive, on average, fewer medical services than optimal. These are also similar to the under-served *FFS* patients (with respect to treatment preference and illness). While *P4P* patients 2D, 4D and 3C are marginally under-served, *P4P* patient 3D receives considerably fewer services (1.06 units less) than optimal.

To gain statistical evidence on the over- or under-provision of medical services, we consider physicians' treatment decisions for each individual patient[3]. A Wilcoxon signed-ranks test comparing individual treatment decisions for *FFS* patients with the patients' respective optimal amount of medical services reveals that individual treatment decisions for 18 of the 20 patients[4] differ significantly from the optimal amount of medical care (refer to Appendix B, Table B.2, for the respective p-values). Fifteen *FFS* patients are found to be significantly over-served with the profit-maximizing quantity for each of these patients exceeding the optimal amount of care. Three *FFS* patients, characterized by the attributes 2D, 4D and 3D, receive significantly fewer services than optimal. For these patients, the profit-maximizing quantity lies below the optimal quantity of services.

Individual treatment decisions for two *FFS* patients, patients 1D and 3C, do not significantly differ from the respective optimal quantity of medical services. These patients tend to receive optimal care under *FFS*. The corresponding p-values for the Wilcoxon signed ranks test are 0.3525 for *FFS* patient 1D and 0.0745 for *FFS* patient 3C. Note that the profit maximizing quantity for each lies below the optimal quantity of treatment. Yet, forgone profits (from providing optimal care instead of the profit maximizing quantity) are minimal as noted above.

[3]We also compared, for each physician, the average quantity provided to fee-for-service patients with the average optimal quantity. A Wilcoxon signed ranks test reveals that physicians provide significantly more services to *FFS* patients than optimal (p= 0.0000). The average optimal quantity for patients is eight units of medical care as noted above.

[4]These patients are characterized by the attributes 1A-C, 1E, 2A-E, 4A-E, 4E, 3A, 3B, 3D and 3E.

A similar analysis[5] of the individual treatment decisions provided to each of the *P4P* patients shows that quantities of care for 14 patients[6] differ significantly from the respective optimal amounts of care (see Appendix B, Table B.2). Thirteen patients treated under *P4P* are significantly over-served while one patient, characterized by the attributes 3D, is significantly under-served.

The remaining six *P4P* patients (characterized by the attributes 1D, 2C, 2D, 4D, 3B and 3C) tend to receive optimal care. For these patients, individual treatment decisions do not differ significantly from the respective optimal quantities. The corresponding p-values values for the Wilcoxon signed ranks tests are 0.4631 for patient 1D, 0.1235 for patient 2C, 0.2622 for patient 2D, 0.0747 for patient 4D, 0.0995 for patient 3B and 0.1282 for patient 3C. These six patients appear to be obvious choices for receiving optimal care under *P4P*. They represent six of the seven patients with the smallest difference in *FFS* physician profit for providing the optimal quantity and the profit-maximizing quantity of treatment. TableB.3 in Appendix B provides a ranking of patients with respect to the difference in profits for the respective profit maximizing and patient benefit maximizing quantities (ascending).

Conclusion 4.3: *We observe over- and under-provision behavior under both payment systems. While individual treatment decisions significantly differ from the optimal quantity of treatment for 18 of the 20 patients under FFS, the same holds true for only 14 of the 20 patients treated under P4P.*

4.4 Response to pay-for-performance

Figure 3 above provides evidence on the differences in treatment behavior across payment systems. It shows that, for the majority of pairs of similar *P4P* and *FFS* patients, the average quantity provided to the *P4P* patient is noticeably closer to the optimal quantity than the average quantity provided to the comparable *FFS* patient. This suggests that *P4P* tends to alleviate over- and under-provision relative to *FFS*.

Figure 4 exhibits, for each pair of matching *FFS* and *P4P* patients, the mean deviation from the optimal quantity of medical care. Note that the mean deviation for *FFS* patient 1D is minimal (with 0.02 units) and thus invisible in the figure (see Appendix B, Table B.2). We observe that for each pair of matching *FFS* and *P4P* patients, the sign (positive of negative) of the mean deviation (implying either over - or under-provision) remains the same across payment systems. In other words, similar patients who tend to be over-served (under-served) under one payment system also tend to be over-

[5]A Wilcoxon signed ranks test, comparing for each physician the average quantity provided to *P4P* patients with the average optimal quantity, reveals that *P4P* patients receive significantly more services than optimal (p= 0.000000).

[6]These *P4P* patients are characterized by the attributes 1A-C, 1E, 2A, 2B, 2E, 4A-C, 4E, 3A, 3D and 3E.

served (under-served) under the other payment system. See also Section 4.3 above. We also observe that for 18 of the 20 pairs of similar *FFS* and *P4P* patients the extent of the over- or under-provision is smaller for the *P4P* than for the corresponding FFS patient, while the reverse is true for the remaining two pairs.

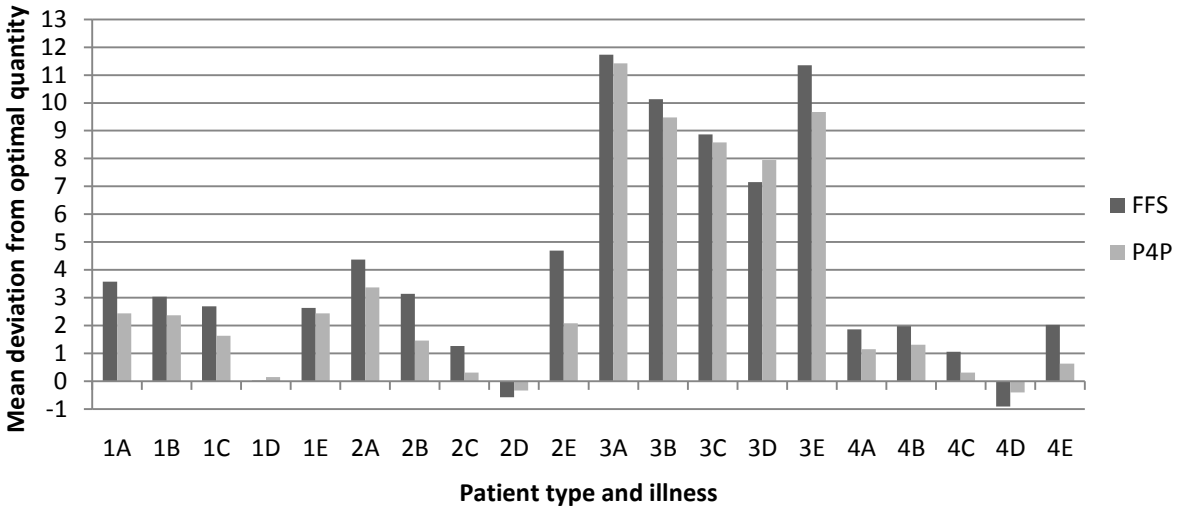

Figure 4: Mean deviation (over-/ underprovision) per individual patient

We count 16 pairs of matching patients, in which both the *FFS* and the *P4P* patient are over-served. In 15 of these pairs, the extent of the over-provision is smaller for the *P4P* patient. The reverse is true for the remaining pair characterized by the patient attributes 1D. For that pair, we report that the mean deviation, although minimal under both payment systems (0.02 units under *FFS* and 0.15 units under *P4P*), is smaller for the *FFS* than the *P4P* patient (see Appendix, Table B.2).

For three of the four pairs, in which both the *FFS* and the *P4P* patient are under-served, we observe the extent of under-provision to be smaller for the *P4P* patient. These matching pairs consist of patients characterized by the attributes 2D, 4D and 3D. The opposite is true for the remaining pair characterized by the attributes 3C: the mean deviation from the optimal quantity of care is minor under both payment systems but the extent of under-provision is larger for the *P4P* than for the *FFS patient*.

To gain statistical evidence, for each patient, on the difference in treatment behavior across the two payment systems, we compare the absolute deviations of the quantity of care provided to a patient from the patient's optimal amount of care based on a Wilcoxon signed-ranks test. We find significant

deviation for 13 of the 20 pairs of matching patients (requiring significance at the 10-percent level).[7] The corresponding p-values for the Wilcoxon signed-ranks tests can be found in Table B.2 in Appendix B. For eleven of these pairs, the extent of over-provision is significantly smaller under *P4P*.[8] For the remaining two pairs (characterized by the attributes 4D and 3D), the extent of under-provision is significantly smaller for the *P4P* patient than for the matching *FFS* patient. To conclude, for 13 of our 20 patient, we observe a significantly smaller deviation from the optimal care for *P4P* than for *FFS* patients, while seven pairs show no significant deviation and none of the pairs shows a larger deviation. A χ^2 test based on the null hypothesis of an equal probability of a smaller, and a zero or larger deviation allows us to reject the null hypothesis at the 1-percent level (Siegel 1987).We thus conclude that there is a significant tendency toward smaller deviations under *P4P* than under *FFS*.

We are particularly interested in whether the tiered *P4P* bonus leads to an increase in the provision of optimal quantities of medical care. We observe that physicians show a higher incidence of optimal treatment decisions under the hybrid payment system: Table 8 above reports that approximately 73.5 percent of the individual treatment decisions for *P4P* patients lead to optimal patient benefit relative to 41.2 percent of the decisions for *FFS* patients. Comparing, for each physician, the relative share of optimal treatment decisions across payment systems, we report the physician's relative share of optimal treatment decisions to be significantly larger under *P4P* ($p = 0.0000$, Wilcoxon signed-ranks test). Significantly more physicians (41 out of 52) are observed to have a larger relative share of optimal treatment decisions under *P4P* ($p < 0.0001$, binomial test).

Similar evidence comes from the analysis of the relative share of optimal treatment decisions for *FFS* and *P4P* patients. We find that each *P4P* patient receives more optimal treatment than the matching *FFS* counterpart (see Figure 5). Comparing, for each pair of matching *FFS* and *P4P* patients, the relative share of optimal treatment decisions across payment systems, we report that *P4P* patients are significantly more likely to receive optimal care than *FFS* patients ($p= 0.0001$; Wilcoxon signed-ranks test).

Figure 5 exhibits the relative shares of optimal treatment decisions for an individual patient, under *FFS* and under *P4P*. The patients in this figure are sorted with respect to the absolute difference in physician profit for providing the optimal quantity and the profit-maximizing quantity (in an ascending order). In each payment system, the relative share of optimal treatment decisions appears to fall as the profit difference increases. To provide statistical support for this observation, Table 9

[7] These pairs consist of patients characterized by the attributes 1A-C, 2B, 2C, 2E, 4A-E, 3D and 3E. No significant differences in the absolute deviations are reported for the remaining pairs 1D, 1E, 2A, 2D, and 3A-C.
[8] These pairs are characterized by the patient attributes 1A-1C, 2B, 2C, 2E, 4A-C, 4E and 3E.

shows the result of a random intercept logit model to explain the conditional probability of optimal treatment based on a *P4P* dummy variable, *profit difference* and a combined effect. We observe that *P4P* has a significantly positive coefficient, while the coefficient of *profit difference* is significantly negative.

Table 9: Results of a random effects logit regression on the probability of optimal treatment

Variable	Coefficient	Std. Error
P4P (dummy)	1.828***	(0.272)
Profit difference	-0.105***	(0.011)
P4P×*profit difference*	0.021	(0.016)
Intercept	1.043***	(0.322)
N	2080	
Log-likelihood	-970.815	
$\chi^2_{(3)}$	331.794***	

Conclusion 4.4: *P4P shows a tendency to alleviate over- or under-provision relative to FFS. In 13 of the 20 pairs of similar FFS and P4P patients, we observe the extent of the over- or under-provision to be significantly smaller for the P4P patient. Physicians are significantly more likely to provide optimal care under the hybrid payment system than under FFS. A P4P patient is significantly more likely to receive optimal care than a similar FFS patient of matching type and illness.*

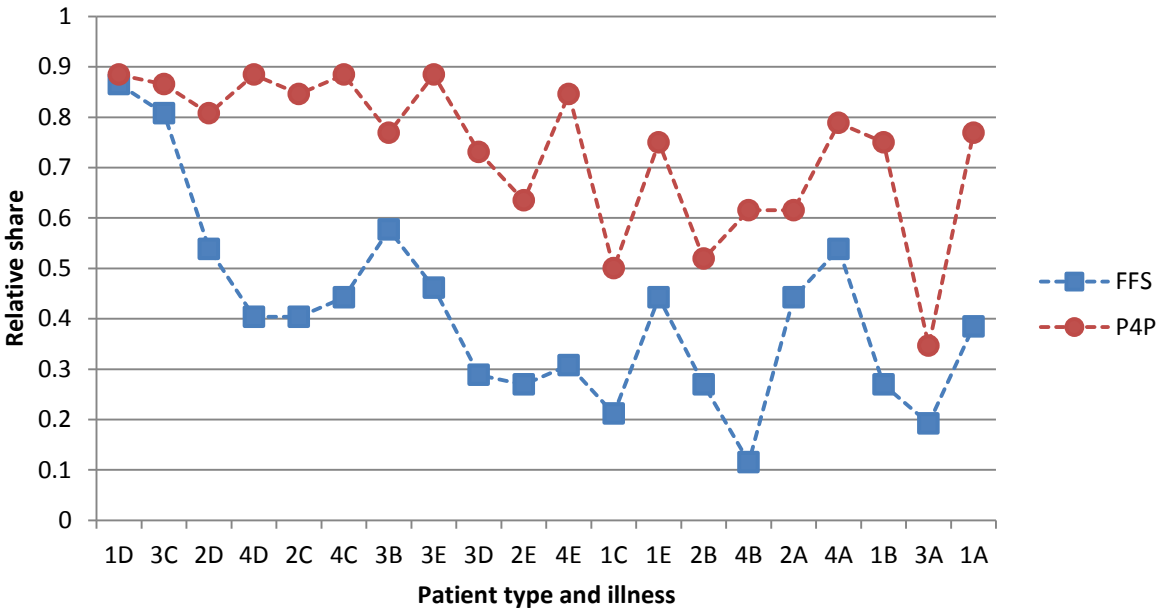

Figure 5: Relative share of optimal treatment decisions
(patients sorted with respect to the difference in profits for providing the optimal quantity and the profit max. quantity; ascending)

4.5 The effect of the payment system on patient benefit

A major concern of our research is to analyze the effect of the payment system on patients' health benefit. We have seen above that a *P4P* patient is significantly more likely to receive optimal care than a similar FFS patient of matching type and illness. Summery statistics on the patient benefit (seeTable8) show an average patient benefit of 18.17 ECU under *FFS* and 19.28 ECU under *P4P*. The average maximum attainable patient benefit in the experiment is 21.45 ECU. Consequently, average observed patient benefits for *FFS* and *P4P* patients are 15.3 and 10.1 percent lower than the average maximum attainable patient benefit.

To evaluate the effect of the payment system on patient benefit, we investigate the benefit loss (difference between the maximum attainable and the recorded patient benefit) relative to the maximum attainable benefit (see Hennig-Schmidt et al. 2011). This *relative benefit loss* can range between 0 and 1.25 in our analysis. A relative patient benefit loss of zero implies optimal treatment, while a value of one indicates a quantity of treatment with no benefit to the patient. A relative benefit loss that is greater than one indicates that the quantity of treatment is injurious to health. To begin the investigation, we analyze, for each physician, the relative benefit loss, averaged over all treatment decisions for patients of the same payment system. A comparison across payment systems

shows that *P4P* patients in the experiment fare significantly better than *FFS* patients. The corresponding p-vale for the Wilcoxon signed-ranks test is 0.000105.

Distinguishing among patient types with respect to treatment preferences (i.e., type 1 to 4), a comparison across payment systems of the relative benefit loss, averaged over patients of the same payment system and patient type, reveals that *P4P* patients of each type fare better than their *FFS* counterparts (Table 10). The largest difference in the relative benefit loss across payment systems is reported for patients of type 4. For each patient type, a Wilcoxon signed-ranks test comparing, for each physician, the average relative patient loss across the two payment systems provides statistical evidence. We find that *P4P* patients of each type do significantly better than their *FFS* counterparts. The corresponding p-values for types 1, 2, 3 and 4 are 0.015498, 0.000497, 0.000027 and 0.000007, respectively.

Table 10: Relative patient benefit loss per patient type

	Type 1	Type 2	Type 3	Type 4
FFS	0.208	0.112	0.111	0.175
P4P	0.166	0.077	0.071	0.093

In the following, we analyze the difference in health benefit for pairs of matching *FFS* and *P4P*patients across payment systems. Figure 6 contrasts the average relative patient benefit loss for matching pairs of *FFS* and *P4P* patients. We observe the relative benefit loss, averaged over physicians' individual treatment decisions for a patient, to be smaller for 19 *P4P* patients relative to their respective *FFS* counterparts.[9] On the individual patient level, the difference is statistically significant for twelve of the 19 pairs of matching patients (characterized by the attributes 1A-C, 2C-E, 3D; 3E, and 4B-E), requiring significance at the 10-percent level (Wilcoxon signed-ranks tests). For the remaining eight pairs (including the one pair in which the *FFS* patient fares, on average, better than the *P4P* match), the Wilcoxon signed-ranks tests reveal no significant difference in the relative

[9]The one patient, who fares better under *FFS* than under *P4P*, is characterized by the attributes 3C. The observed difference in the average relative patient benefit loss is relatively small.

patient benefit loss across payment systems (refer to Appendix B, Table B.4 for the respective p-values).

Conclusion.4.5: *In terms of relative benefit loss, P4P patients tend to fare better than their FFS counterparts. This holds for each of the four patient types.*

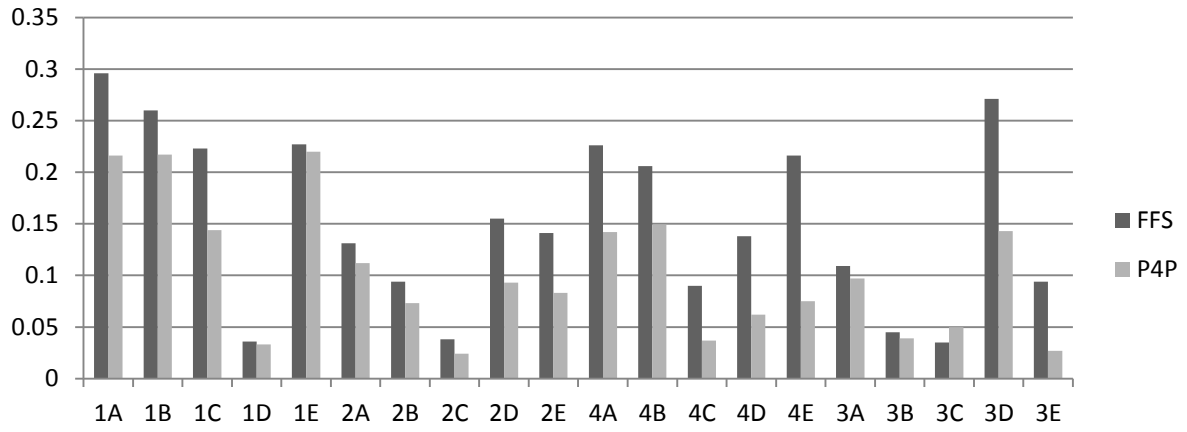

Figure 6: Average relative benefit loss per patient

4.6 Physician unethical conduct

Medical professional societies deem the prescription, provision and billing of medical services with no benefit to the patient as unethical (refer, for example, to the *Code of Professional Ethics of the American College of Obstetricians and Gynecologists,* 2011). We use this definition as a criterion to evaluate physician behavior in our experiment. Evidence of unethical conduct would be highly controversial given the supposedly special physician-patient-relationship. In contrast to the related experiments by Hennig-Schmidt et al. (2011) and Keser et al. (2013), our experimental design "allows" physicians to provide unethical treatment to half of their virtual patients in the experiment. Revisit the benefit functions for patient types 1 and 4 presented in Table 1. For patients of type 1, the provision of more than 14 units of medical care is injurious to health (resulting in a negative patient benefit). Patients of type 4 draw zero benefit from 16 units of medical services (hence, 16 units result in the same patient benefit as no treatment); seventeen or more units are injurious to health. In the experiment, ten of the 20 patients treated under *FFS* are either of type 1 or type 4. The same holds true for half of the patients treated under the hybrid payment system *P4P*.

The data shows that almost one in three physicians in our experiment provide unethical treatment to at least one virtual patient. For eight of the 16 physicians caught red-handed, this is a one-time slip-up in the experiment. Another three physicians provide unethical quantities in two instances. Five[10]physicians (or 9.6 percent of the physicians in the experiment) are observed to be repeat offenders with at least five unethical treatment decisions in the experiment each. For repeat offenders, between 25 and 45 percent of their individual treatment decisions for *FFS* and *P4P* patients of type 1 and 4 are unethical.

Six of the sixteen physicians provide unethical treatment under both payment systems. For the remaining ten physicians, we find unethical behavior only under one of the payment systems. Three physicians provide unethical care exclusively under *FFS* while seven show unethical behavior exclusively under *P4P*.

Unethical behavior, however, does not differ significantly across payment systems. Roughly 4.2 percent of the treatment decisions for *FFS* patients of type 1 and type 4 are unethical. The same is true for 5 percent of the treatment decisions for the same group of *P4P* patients. In addition, we do not find a significant difference in the individual physicians' relative share of unethical treatment decisions across payment systems ($p = 0.414447$; Wilcoxon-signed ranks test).

Under *FFS*, roughly 86.4 percent of the unethical treatment decisions are provided to five patients, characterized by 1A, 1B, 1E, 4A and 4E. The same holds true for 91.3 percent of the unethical treatment decisions under *P4P*. The explanation for the observed pattern is straightforward: The *FFS* physician profit functions and the patient benefit functions reveal that, for patients of type 1 and 4, the *FFS* profit maximizing quantities for illnesses A and E (20 units; 18 units) lie within the range of unethical treatment quantities. The same is true for the *FFS* profit maximizing quantity for illness B (15 units) in the case of patients of type 1.

Roughly 9.6 percent, 5.8 percent and 5.8 percent of the treatment decisions for *FFS* patients 1A, 1B and 1E are injurious to health. The corresponding percentages for the matching *P4P* patients are 11.5, 7.7 and 11.5. For each of the *FFS* patients 4A and 4E, we report 7.7 percent of the treatment decisions to be unethical. The same applies to 9.6 and 5.8 percent of the individual treatment decisions for the matching *P4P* patients.

Conclusion 4.6: *We observe unethical treatment behavior in 4.2 and 5 percent of the treatment decisions for FFS and P4P patients of type 1 and 4, respectively. Approximately 21.2 percent of the*

[10] The physicians #2 #16, #46, #48 and #49 are identified to be repeat offenders; 40, 25, 30, 45 and 30 percent of their individual treatment decisions for *FFS* and *P4P* patients of Type 1 and 4 are unethical, respectively.

physicians provide unethical treatment in only one or two instances, an additional 9.6 percent of the physicians are found to be repeat offenders with at least five unethical treatment decisions.

4.7 Individual characteristics and attitude towards patient's benefit

Our previous results have shown that there exists a large heterogeneity among experimental physicians in attitude towards patient's benefit. Although we find that a large proportion of participants are willing to maximize patient's benefit, we also observe evidence of unethical behavior in both remuneration schemes. In this section, we investigate if individual differences in service provision and frequency of unethical conduct could be (at least partly) explained by differences in individual characteristics. In a questionnaire following the experiment, participants answer several questions regarding their personality traits. This provides us with valuable information regarding a large range of aspects, such as sociability, reciprocity, attitude toward risk and pressure or self-esteem. The questions we use to elicit participants' personal characteristics are displayed in Table C.1 (Appendix C).

The size of our sample does not allow us to include all of this information in econometric regressions. For that reason, we run a principal component analysis to identify the most relevant factors to characterize participants' personality traits. Results suggest that 57.78 percent of the variations in could be explained by three factors. The corresponding factor loadings are presented in Table C.2 (Appendix C).

We label the first factor "self-confidence". An individual having a high and positive value on this factor is less prone to stress, is more likely to take risks, to be impulsive and to trust others. Furthermore, this factor is also associated to a high self-esteem. The second factor is labeled "antisocial". Individuals who score high on this factor are less likely to be sociable and to reciprocate positive actions. They are also more prone to negative reciprocity (envy, resentfulness) and to risk seeking. The third factor is associated to a high sociability, a high patience and a low trust. Since we do not have an intuitive consistent interpretation of this factor, we will then concentrate on the first two factors in the following of the analysis.

Table 11 reports estimates of econometric regressions aiming at identifying the determinants of physician's attitude towards patients' benefit. Results suggest that participants' answers to the post-experimental questionnaire provide us with relevant explanation of their decisions in the experiment. The first column of Table 11 addresses the determinants of physician's profits (excluding bonus in the P4P treatment). Consistent with our previous results, we observe that participants are willing to forgo a substantial part of profit when P4P bonuses are at stake. We also observe that

female physicians tend to earn less than males. Interestingly, individuals scoring higher in the "antisocial" factor tend to earn more, suggesting a behavior oriented toward profit rather than patient's benefit. This interpretation is confirmed in the second regression reported in Table 11. Our results suggest that patients' overall benefits tend to be lower when treated by a physician that scores high on the "antisocial" factor. Here again, econometric analysis confirms our previous finding that patients' benefit increase with the introduction of the P4P payment scheme. The last column of Table 11 reports that personality features have a clear impact on the occurrence of unethical conduct. Individuals scoring high on the "self-confidence" factor (characterized notably by a higher self-esteem and a higher trust) are less likely to behave unethically. On the opposite, participants who score higher on the "antisocial" factor are more prone to provide services that have no, or even negative effect on patient's benefit.

Table 11: Determinants of physician attitude towards patient's benefit – OLS regressions

Interest variable	Total profit (excluding P4P bonus)	Total patients' benefit	Unethical physician (Frequency of unethical conduct)
P4P treatment	-52.304***	22.269*	0.077
	(10.274)	(13.180)	(0.148)
Female	-22.073*	-0.594	-0.188
	(11.810)	(15.149)	(0.170)
Age	-1.397	-1.327	-0.009
	(1.592)	(2.042)	(0.023)
Factor 1 : self-confidence	-1.093	1.999	-0.120**
	(3.540)	(4.541)	(0.051)
Factor 2 : antisocial	7.975*	-15.698***	0.126**
	(4.274)	(5.482)	(0.061)
Intercept	470.505***	395.85***	0.775
	(39.876)	(51.151)	(0.573)
N	104	104	104
R^2	0.1571	0.1146	0.0620

Conclusion 4.7: *Individual personality characteristics, such as self-esteem and other-regarding preferences, directly impact the conduct of experimental physicians. Physicians who show larger concern regarding others are more likely to behave ethically.*

5. Conclusion

P4P has been enjoying a growing popularity among healthcare policy makers in spite of the lack of convincing empirical evidence on its effectiveness (e.g., Emmert et al. 2011). Our study adds to the existing empirical *P4P* literature and the sparse experimental literature on physician behavior. The presented paper provides the first experimental investigation of *P4P* in healthcare analyzing a hybrid physician payment system that blends traditional *FFS* with *P4P* incentives. The implemented non-competitive *P4P* incentive structure consists of a tiered bonus payment with absolute performance thresholds and an uncapped bonus pool.

The results show that, in our experimental model, physicians respond to *P4P* incentives embedded in a hybrid payment system. Approximately nine out of ten participants in the experiment qualify for a *P4P* bonus. The experimental physicians' relative share of optimal treatment decisions is significantly larger under the hybrid payment system than under *FFS*. A *P4P* patient tends to receive significantly more optimal treatment than a patient of matching type and illness under *FFS*. *P4P* in many cases alleviates over- and under-provision relative to traditional *FFS*. In both payment systems, we do observe the occurrence of unethical treatment behavior, i.e. the provision of medical services with no benefit to the patient, irrespective of the payment system. The conduct of experimental physicians appears to be influenced by individual personality characteristics, such as self-esteem and other-regarding preferences.

Future experimental research could extend the presented study in several directions by altering, for example, the design of the payment structure. Rewarding relative performance rather than target attainment (absolute performance) as in our experiment would present a very compelling extension of our work as it creates a competitive *P4P* incentive structure. In a tournament approach, physicians could be ranked based on their performance relative to their peers with top performers earning a bonus payment (see, e.g., Falk & Fehr 2003; Lindenauer et al. 2007; Pope 2012). Another extension of our work could evaluate the effects of financial penalties for substandard performance on healthcare provider behavior (see e.g., Lindenauer et al. 2007) rather than offering financial rewards for target attainment (as in our experiment).

References

Beaulieu, N.D. & D. R. Horrigan (2005): "Putting smart money to work for quality improvement," *Health Services Research*, 40 (5), 1318-1334.

Brosig-Koch, J.; Hennig-Schmidt, H.; Kairies, N. & D. Wiesen (2013a): "How to improve patient care? An analysis of capitation, fee-for-service, and mixed incentive schemes for physicians," Ruhr Economic Papers No. 412.

Brosig-Koch, J.; Hennig-Schmidt, H.; Kairies, N. & D. Wiesen (2013b): "How effective are pay-for-performance incentives for physicians? A laboratory experiment," Ruhr Economic Papers No. 413.

Campbell, S.; Reeves, D.; Kontopantelis, E.; Sibbald, B. & D. M. Roland (2009): "Effects of Pay for Performance on the Quality of Primary Care in England,"*New England Journal of Medicine*; 361, 368-378.

Carey, I. M.; Nightingale, C. M.; DeWilde, S.; Harris, T.; Whincup, P. H. & D. G. Cook (2009): "Blood pressure recording bias during a period when the Quality and Outcomes Framework was introduced," *Journal of Human Hypertension*, 23, 764 - 770.

Casalino, L. P. & A. Elster (2007): "Will pay-for-performance and quality reporting affect health care disparities?" *Health Affairs,* 26 (3), 405 – 414.

Dalton, A. R.; Alshamsan, R.; Majeed, A. & C. Millett (2011): "Exclusion of patients from quality measurement of diabetes care in the UK pay-for-performance programme," *Diabetic Medicine*, 28 (5), 525-531.

Doran, T.; Kontopantelis, E.; Valderas, J. M.; Campbell, S.; Roland, M.; Salisbury, C. & D. Reeves (2011): "Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework," British Medical Journal, BMJ 2011; 342:d3590doi: 10.1136/bmj.d3590.

Emmert, M.; Eijkenaar, F.; Kemter, H.; Esslinger, A. S. & O. Schöffski (2011): "Economic evaluation of pay-for-performance in healthcare – a systematic review," European Journal of Health Economics; DOI 10.1007/s10198-011-0329-8.

Epstein, A. M., Lee, T. H. &M. B. Hamel (2004): "Paying physicians for high quality care," *New England Journal of Medicine, 350* (4), 406−410.

Fairbrother, G.; Hanson, K. L.; Friedman, S., & G. C. Butts (1999): "The impact of physician bonuses, enhanced fees, and feedback on childhood immunization coverage rates," *American Journal of Public Health*, 89 (2): 171-175.

Falk, A. & E. Fehr (2003): "Why labor market experiments?" *Labour Economics*, 10, 399 - 406.

Fischbacher, U. (2007): "z-Tree: Zurich Toolbox for Ready-made Economic Experiments," *Experimental Economics*, 10 (2), 171 - 178.

Frey, B. S. (1994):“How intrinsic motivation is crowded out and in,” *Rationality and Society*, 6, 334 – 352.

Goeckenjan, G., Sitter, H., Thomas, M. et al. (2011): “Prevention, diagnosis, therapy, and follow-up of lung cancer: interdisciplinary guideline of the German Respiratory Society and the German Cancer Society,” Pneumologie, 65, 39-59.

Glickman, S. W.; Ou, F.-S.; DeLong, E. R.; Roe, M. T.; Lytle, B. L.; Mulgund, J.; Rumsfeld, J.S.; Gibler, W. B.; Ohman, E.M., Schulman, K. A. & E. D. Peterson (2007): “Pay for Performance, Quality of Care, and Outcomes in Acute Myocardial Infarction,” *Journal of the American Medical Association*, 297 (21), 2373 - 2380.

Gravelle, H.; Sutton, M. & A. Ma (2010): “Doctor behavior under a pay for performance contract: treating cheating and case finding,” *The Economic Journal*, 120, 129 - 156.

Hahn, J. (2006). “Pay-for-performance in healthcare,” Congressional Research Service Report for Congress, Order Code RL 33713.

Hennig-Schmidt, H., Selten, R. & D. Wiesen (2011): “How Payment Systems Affect Physicians’ Provision Behavior – An Experimental Investigation,” *Journal of Health Economics*, 30 (4), 637 - 646.

Hillman, A. L.; Ripley, K.; Goldfarb, N.; Nuamah, I.; Weiner, J. & E. Lusk (1998): “Physician financial incentives and feedback: failure to increase cancer screening in Medicaid managed care,” *American Journal of Public Health*, 88 (11), 1699-1701.

Hillman, A. L.; Ripley, K.; Goldfarb, N.; Weiner, J.; Nuamah, I. & E. Lusk (1999): “The use of physician financial incentives and feedback to improve pediatric preventive care in Medicaid managed care,” *Pediatrics*, 104 (4 Pt 1): 931-935.

Holmstrom, B. & P. Milgrom (1991): “Multi-task principle-agent analyses: incentive contracts, asset ownership, and job design,” *Journal of Law, Economics and Organization*, 7, 24 - 52.

Hong, C. S.; Atlas, S. J.; Chang, Y.; Ashburner, J. M.; Barry, M.J. & R. W. Grant (2010): “Relationship between patient panel characteristics and primary care physician clinical performance rankings,” *Journal of the American Medical Association*, 304 (10), 1107 - 1113.

Hutchison, B. (2008): “Pay for performance in primary care: proceed with caution, pitfalls ahead,” *Healthcare Policy*, 4 (1), 10 - 15.

Hutchison, B.; Birch, S.; Hurley, J.; Lomas, J. & F. Stratford-Devai (1996): “Do physician-payment mechanisms affect hospital utilization? A study of Health Service Organizations in Ontario,” *Canadian Medical Association Journal*, 154 (5): 653-661.

Institute of Medicine of the National Academies (2001): “Crossing the Quality Chasm. A New Health System for the 21st Century,” *National Academy Press*, Washington, D.C.

Keser, C.; Montmarquette, C.; Schmidt, M. & C. Schnitzler (2013): “Custom-made healthcare –An experimental investigation”, CIRANO Scientific Series 2013s-15.

Kouides, R. W.; Bennett, N.M.; Lewis, B.; Cappuccio, J.D.; Barker, W.H.; & F. M. LaForce (1998): "Performance-based physician reimbursement and influenza immunization rates in the elderly. The Primary-Care Physicians of Monroe County," *American Journal of Preventive Medicine*, 14 (2), 89-95.

Li, J.; Hurley, J.; DeCicca, P. & G. Buckley (2011): "Physician response to pay-for-performance: evidence from a natural experiment," *NBER Working Paper Series*, w16909.

Lindenauer, P. K.; Remus, D.; Roman, S.; Rothberg, M. B.; Benjamin, E. M.; Ma, A. & D. W. Bratzler (2007): "Public reporting and pay for performance in hospital quality improvement," *New England Journal of Medicine*, 356 (5): 486-496.

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

McGlynn, E.; Steven M. Asch, S. M.; Adams, J.; Keesey, J.; Hicks, J.; DeCristofaro, A.; & E. A. Kerr (2003): "The Quality of Health Care Delivered to Adults in the United States," *New England Journal of Medicine,* 348(26):2635 - 2646.

Mullen, K. J.; Frank, R. G. & 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 (1), 64 – 91.

Pines, J. M. (2006): "Profiles in Patient Safety: Antibiotic Timing in Pneumonia and Pay-for performance," *Journal of Academic Emergency Medicine*, 13 (7), 787-790.

Pope, G. C. (2012). "Overview of Pay-for-Performance Models and Issues," in: *Pay for Performance in Healthcare*: *Methods and Approaches*, edited by Cromwell, J., Trisolini, M. G., Pope, G.C., Mitchell, J.B. & L. M. Greenwald, Research Triangle Institute.

Richards, J. (2009): "Is there an elephant in the room?" British Journal of General Practice, 59, 376 – 377.

Rosenthal, M., Frank, R., Li, Z. & A. Epstein (2005): "Early experience with pay-for-performance: from concept to practice," *Journal of the American Medical Association*, 294, 1788 - 93.

Rosenthal, M. B. & R. G. Frank (2006): "What is the Empirical Basis for Paying for Quality in Health Care?" *Medical Care Research and Review*, 63 (2), 135 – 157.

Rosenthal, M. B.; Li, Z.; Robertson, A. D. & A. Milstein (2009): "Impact of Financial Incentives for Prenatal Care on Birth Outcomes and Spending," *Health Services Research*, 44 (5 PI), 1465 – 1479.

Shen, Y. (2003): "Selection incentives in a performance-based contracting system," *Health Services Research*, 38 (2), 535-552.

Sicsic, J.; Le Vaillant, M. & C. Franc (2012): "Intrinsic and extrinsic motivations in primary care: An explanatory study among French general practitioners," *Health Policy* (in print) http://dx.doi.org/10.1016/j.healthpol.2012.08.020 .

Siegel, S (1987): Nicht-parametrische Methoden, Fachbuchhandlung für Psychologie, Verlagsabteilung, Eschborn bei Frankfurt am Main.

Strong, M.; South, G. & R. Carlisle (2009): "The UK Quality and Outcomes Framework pay-for-performance scheme and spirometry: rewarding quality of just quantity? A cross-sectional study in Rotherham, UK," *BMC Health Services Research* 9: 108; doi:10.1186/1472-6963-9-108.

The American College of Obstetricians and Gynecologists (2011): "Code of Professional Ethics," http://www.acog.org/About_ACOG/~/media/Departments/National%20Officer%20Nominations%20Process/ACOGcode.pdf ; accessed October 12, 2012.

Young, G. J.; Meterko, M.; Beckman, H.; Baker, E.; White, B.; Sautter, K. M.; Greene, R.; Curtin, K.; Bokhour, B. G.; Berlowitz, D. & J.F. Burgess Jr. (2007): "Effects of paying physicians based on their relative performance for quality," *Journal of General Internal Medicine*, 22 (6), 872-876.

Appendix A: Instructions (for the order *FFS/P4P*)

EXPERIMENT INSTRUCTIONS [PART A]

You are participating in an experiment in which you will make independent and anonymous decisions. Depending on these decisions you can earn money. In the experiment you will make a total of 40 decisions.

The experiment consists of two parts, Part A and B, each involving 20 decisions. After you will have successfully completed Part A, you will receive (during a short interruption of the experiment) new experiment instructions for Part B.

All amounts in the experiment are denoted in ECU (Experimental Currency Units).The ECU that you earn in the experiment will be converted into € with a factor 1 ECU = 0,01€ and paid to you in addition to a show-up fee of 3,00€ in cash at the end of the experiment.

YOUR DECISIONS

In the experiment you will be in the role of a physician making medical decisions for virtual patients. These decisions will impact your profit as a physician as well as the patient benefit.

You will be responsible for the medical treatment of 20 virtual patients and decide for each individual patient on the number of medical services that you want to provide to this patient. The treatment can consist of an amount between zero (including) and twenty (including) units of medical services.

The virtual patients will be presented to you one patient after the other. Each patient suffers from one out of five potential illnesses and belongs to one out of four patient types. We shall specify neither the illnesses nor the patient types in more detail. You won't know the illness or type of a patient; you will only see numbers associated with illness and type related to possible treatments for the specific patient. These numbers include your remuneration, your costs, your profit and the patient benefit

MONETARY PATIENT BENEFIT

Your decision on the number of medical services that you want to provide to a patient also determines the benefit that this patient gets from your treatment. This benefit depends on the patient type and the number of services but not on the illness.

YOUR REMUNERATION

Your medical treatment will be remunerated based on a Fee-for-Services (FFS) system. In this remuneration system, each unit of service that you provide will be paid separately. Your remuneration thus increases with the number of services. In addition, your remuneration depends on the patient's illness.

YOUR COSTS AND PROFIT

With your decision on the number of medical services that you want to provide to a patient you also determine your costs of treating this patient. The treatment costs increase with the number of services.

Your profit per patient is determined by your remuneration minus your treatment costs for this patient.

YOUR INFORMATION

In the experiment we shall confront you with decision situations as in the following example. In the experiment we shall present you a sequence of 40 such decision situations.

You have to make a decision on the number of services for this patient. The table shows you for each potential number of services that you provide to this patient your respective remuneration, your treatment costs, your profit (remuneration minus costs), and the monetary patient benefit. You will be asked to enter your decision on the number of services units in the box below the table. Please choose an integer number between zero and 20. To confirm your decision, please click on "OK".

Example:

Number of Services	Remuneration	Costs	Your Profit	Patient Benefit
0	0.00	0.00	0.00	0.00
1	2.50	0.20	2.30	12.00
2	5.00	0.80	4.20	15.50
3	7.50	1.80	5.70	18.00
4	10.00	3.20	6.80	20.00
5	12.50	5.00	7.50	19.20
6	15.00	7.20	7.80	16.80
7	27.50	9.80	17.70	14.90
8	30.00	12.80	17.20	13.00
9	32.50	16.20	16.30	11.10
10	35.00	20.00	15.00	9.20
11	37.50	24.20	13.30	7.30
12	44.00	28.80	15.20	5.40
13	66.50	33.80	32.70	3.50
14	73.00	39.20	33.80	1.60
15	80.50	45.00	35.50	-0.30
16	88.00	51.20	36.80	-2.20
17	94.60	57.80	36.80	-3.40
18	102.20	64.80	37.40	-4.30
19	109.80	72.20	37.60	-4.80
20	118.40	80.00	38.50	-5.00

Please enter the number of services that you want to provide to this patient:

OK

PAYMENT

At the end of the experiment, your individual profits of Part A (resulting from the treatment of all 20 patients) and Part B of the experiment will be summed up, converted into € [1 ECU = 0.01€] and paid to you in addition to your show-up fee [3.00€] in cash.

Since there are no real patients participating in this experiment, we shall donate the sum of patient benefits to a charitable healthcare organization. In this way your treatment decisions create benefit to real patients.

At the beginning of the experiment, you may decide on the charitable healthcare organization to which you want to donate. You can choose among:

- DIE DEUTSCHE KREBSGESELLSCHAFT
- DIE DEUTSCHE MULTIPLE SKLEROSE GESELLSCHAFT (DMSG)
- DIE DEUTSCHE PARKINSON GESELLSCHAFT

The patient benefit resulting from your decisions will be added up for all Patients, converted into € [1 GE = 0.01€] and paid to the organization of your choice. The total patient benefit that has been created by all participants having chosen the same charitable healthcare organization will be donated online to the respective organization. Within 14 days you will receive a copy of the receipt by Email.

Please turn now to the computer with your participation number and click on “Start”. You will be requested to answer a number of questions related to the understanding of these instructions. If you should have remaining questions, we will come to your workplace and answer them individually. As soon as all participants will have correctly answered all questions, the experiment can start.

EXPERIMENT INSTRUCTIONS [PART B]

In Part B of the experiment you will be in the same decision situation but under a new remuneration rule:

YOUR REMUNERATION

Similarly to Part A of the experiment, Your medical treatment will be remunerated based on a Fee-for-Services (FFS) system. Each unit of service that you provide will thus be paid separately.

In addition to the FFS you may receive an extra payment at the end of the experiment. The amount of this payment depends on the number of patients that you have treated optimally in Part B of the experiment. An optimally treated patient receives from you the number of medical services that leads to the maximal patient benefit. To receive an extra payment it is necessary that you treat a minimum number of patients in part B optimally. If you treat a larger number of patients optimally, your extra payment increases. This is presented in Table 1. If you treat a minimum of 10 patients optimally you receive an extra payment in the amount of 130.00 ECU. In the case of 13 to 15 optimally treated patients you receive 165.00 ECU. If you boast more than 15 optimally treated patients in Part B, then you receive an extra payment of 215.00 ECU.

Table 1: Amounts of extra payment for optimally treated patients in experiment Part B

Number of optimally treated patients (out of 20)	Extra payment (in ECU)
0-9 Patients	0.00
10 -12 Patients	130.00
13 -15 Patients	165.00
16 -20 Patients	215.00

PAYMENT

At the end of the experiment, your individual profits of Part A and Part B (resulting from the treatment of all 20 patients plus a potential extra payment) will be summed up, converted into € [1 ECU = 0.01€] and paid to you in addition to your show-up fee [3.00€] in cash. Donations of the monetary patient benefits to the charitable healthcare organizations shall be effectuated in the same way as in Part A

Please continue with Part B of the experiment.

Appendix B: Data tables

Table B.1: The ordering effect: Quantity q and p-values

	FFS							*P4P*						
	Order *FFS/P4P*			Order *P4P/FFS*				Order *FFS/ P4P*			Order *P4P/FFS*			
Patient	Mean	Median	SD	Mean	Median	SD	p-value[a]	Mean	Median	SD	Mean	Median	SD	p-value[b]
1A	7.08	7.00	3.87	8.08	5.50	5.80	0.731491	6.50	4.00	5.14	6.38	4.00	5.58	0.637338
1B	7.04	5.50	2.93	7.04	5.00	4.10	0.161539	6.38	4.00	4.45	6.35	4.00	4.31	0.856411
1C	7.08	6.00	3.32	6.31	5.00	2.56	0.236694	5.23	4.00	2.18	6.04	5.00	3.62	0.449331
1D	3.88	4.00	0.43	4.15	4.00	1.93	0.781176	3.85	4.00	0.54	4.46	4.00	1.70	0.102486
1E	6.35	5.00	3.36	6.92	4.00	4.75	0.247234	5.65	4.00	4.22	7.23	4.00	5.67	0.323391
2A	12.46	13.00	3.70	12.27	8.00	5.10	0.736015	12.69	13.00	4.89	10.04	8.00	4.65	0.028405
2B	11.08	9.00	2.88	11.19	9.00	3.42	0.984971	9.58	8.00	2.59	9.35	8.00	3.07	> 0.99999
2C	9.50	10.00	0.86	9.04	8.00	1.40	0.056324	8.23	8.00	0.65	8.38	8.00	1.88	0.941802
2D	7.69	7.50	2.09	7.15	8.00	3.47	0.754680	7.38	8.00	1.68	7.96	8.00	2.58	0.236342
2E	13.04	12.00	3.87	12.35	9.50	4.72	0.276135	10.19	8.00	3.96	9.96	8.00	3.83	0.625411
4A	9.15	8.00	2.29	10.58	8.00	4.21	0.276101	9.23	8.00	3.54	9.08	8.00	3.63	0.768101
4B	9.85	9.00	1.85	10.12	9.00	2.55	0.793248	9.15	8.00	2.11	9.46	8.00	2.72	0.528675
4C	9.19	10.00	0.98	8.92	8.00	1.20	0.297479	8.04	8.00	0.45	8.58	8.00	1.45	0.092468
4D	6.58	7.00	1.33	7.62	8.00	2.73	0.001990	7.54	8.00	1.63	7.65	8.00	1.47	0.986837
4E	10.08	9.00	2.48	9.96	8.50	3.00	0.302360	8.88	8.00	2.53	8.38	8.00	2.02	0.466088
3A	14.65	13.00	2.53	14.81	13.00	3.15	0.727380	14.85	13.00	3.08	14.00	12.00	3.21	0.027848
3B	13.15	13.00	1.22	13.15	12.00	1.90	0.496883	12.69	12.00	1.76	12.27	12.00	1.59	0.619481
3C	12.12	12.00	1.70	11.62	12.00	0.75	0.196147	11.92	12.00	0.39	11.23	12.00	2.57	0.643561
3D	10.50	11.00	3.87	9.81	12.00	4.72	0.992578	10.42	12.00	3.79	11.46	12.00	3.36	0.526545
3E	14.85	14.50	2.77	13.85	12.00	2.46	0.128989	12.69	12.00	1.95	12.65	12.00	1.85	0.973649

[a] p-values for Wilcoxon-Mann-Whitney-U tests, comparing individual treatment decisions for individual *FFS* patients in *FFS*/*P4P* with individual decisions in *P4P*/*FFS*.

[b] p-values for Wilcoxon-Mann-Whitney-U tests, comparing individual treatment decisions for individual *P4P* patients in *FFS*/*P4P* with individual decisions in *P4P*/*FFS*.

Table B.2: Quantity q, mean deviation from the optimal quantity q* and p-values

	FFS									*P4P*									*Comp.*
	Quantity q			Deviation from q*			Rel. Share opt. Dec.	p-values		Quantity q			Deviation from q*			Rel. Share opt. Dec.	p-values		p-value
Patient	Mean	Median	SD	Mean	Median	SD		p^a	p^b	Mean	Median	SD	Mean	Median	SD		p^a	p^b	p^c
1A	7.58	7.00	4.90	3.58	3.00	4.90	0.385	0.000001	0.000000	6.44	4.00	5.31	2.44	0.00	5.31	0.769	0.003264	0.000000	0.018427
1B	7.04	5.00	3.53	3.04	1.00	3.53	0.269	0.000000	0.000000	6.37	4.00	4.33	2.37	0.00	4.33	0.75	0.001474	0.000000	0.027291
1C	6.69	5.00	2.96	2.69	1.00	2.96	0.212	0.000000	0.000001	5.63	4.00	2.98	1.63	0.00	2.98	0.5	0.000067	0.000000	0.018444
1D	4.02	4.00	1.39	0.02	0.00	1.39	0.865	0.352543	0.000000	4.15	4.00	1.29	0.15	0.00	1.29	0.885	0.463072	0.000000	0.779435
1E	6.63	5.00	4.08	2.64	1.00	4.08	0.442	0.000005	0.000000	6.44	4.00	5.01	2.44	0.00	5.01	0.75	0.001872	0.000000	0.350496
2A	12.37	13.00	4.41	4.37	5.00	4.41	0.442	0.000003	0.000000	11.37	8.00	4.91	3.37	0.00	4.91	0.615	0.000103	0.000000	0.182897
2B	11.13	9.00	3.13	3.13	1.00	3.13	0.269	0.000000	0.000000	9.46	8.00	2.82	1.46	0.00	2.82	0.519	0.000808	0.000000	0.006452
2C	9.27	10.00	1.17	1.27	2.00	1.17	0.404	0.000001	0.001183	8.31	8.00	1.39	0.31	0.00	1.39	0.846	0.123486	0.000000	0.000563
2D	7.42	8.00	2.85	-0.58	0.00	2.85	0.538	0.003252	0.000000	7.67	8.00	2.18	-0.33	0.00	2.18	0.808	0.262193	0.000000	0.204122
2E	12.69	11.50	4.29	4.69	3.50	4.29	0.269	0.000000	0.000000	10.08	8.00	3.86	2.08	0.00	3.86	0.635	0.000539	0.000000	0.000911
4A	9.87	8.00	3.43	1.87	0.00	3.43	0.538	0.000441	0.000000	9.15	8.00	3.55	1.15	0.00	3.55	0.788	0.029383	0.000000	0.097374
4B	9.98	9.00	2.21	1.98	1.00	2.21	0.115	0.000000	0.000000	9.31	8.00	2.41	1.31	0.00	2.41	0.615	0.000089	0.000000	0.005929
4C	9.06	9.50	1.09	1.06	1.50	1.09	0.442	0.000000	0.000028	8.31	8.00	1.09	0.31	0.00	1.09	0.885	0.046400	0.000000	0.003379
4D	7.10	7.00	2.19	-0.90	-1.00	2.19	0.404	0.000674	0.000000	7.60	8.00	1.54	-0.40	0.00	1.54	0.885	0.074736	0.000000	0.000318
4E	10.02	9.00	2.73	2.02	1.00	2.73	0.308	0.000001	0.000000	8.63	8.00	2.28	0.63	0.00	2.28	0.846	0.049951	0.000000	0.000003
3A	14.73	13.00	2.83	2.73	1.00	2.83	0.192	0.000000	0.000000	14.42	13.00	3.15	2.42	1.00	3.15	0.346	0.000000	0.000000	0.520317
3B	13.13	12.00	1.58	1.13	0.00	1.58	0.577	0.000040	0.000000	12.48	12.00	1.67	0.48	0.00	1.67	0.769	0.099482	0.000000	0.112780
3C	11.87	12.00	1.33	-0.13	0.00	1.33	0.808	0.074463	0.000000	11.58	12.00	1.85	-0.42	0.00	1.85	0.865	0.128191	0.000002	0.833936
3D	10.15	11.50	4.29	-1.85	-0.50	4.29	0.288	0.009887	0.000000	10.94	12.00	3.58	-1.06	0.00	3.58	0.731	0.028009	0.000000	0.000233
3E	14.35	13.00	2.64	2.35	1.00	2.64	0.462	0.000004	0.000000	12.67	12.00	1.89	0.67	0.00	1.89	0.885	0.027709	0.000000	0.000040

[a] p-values for Wilcoxon signed-ranks tests, comparing individual treatment decisions with the respective right amount of care

[b] p-values for Wilcoxon signed-ranks tests, comparing individual treatment decisions with the *FFS* profit maximizing quantity.

[c] p-values for Wilcoxon signed-ranks tests, comparing absolute deviations resulting from individual treatment decisions for matching patients across payment systems

Table B.3: Difference in physician profit per patient for providing the profit maximizing quantity relative to the benefit maximizing quantity q* (ascending; in ECU)

Patient type and illness	$\prod^{pmaxq1}$	$\prod^{bmaxq2}$	Difference	# of q* under *FFS*	# of q* under *P4P*
1D	25.80	25.00	0.80	45	46
3C	25.60	23.00	2.60	42	45
2D	25.80	18.60	7.20	28	42
4D	25.80	18.60	7.20	21	46
2C	25.60	18.30	7.30	21	44
4C	25.60	18.30	7.30	23	46
3B	31.20	23.40	7.80	30	40
3E	29.50	19.60	9.90	24	46
3D	25.80	8.80	17.00	15	38
2E	29.50	12.20	17.30	14	33
4E	29.50	12.20	17.30	16	44
1C	25.60	8.00	17.60	11	26
1E	29.50	11.40	18.10	23	39
2B	31.20	12.80	18.40	14	27
4B	31.20	12.80	18.40	6	32
2A	38.40	17.20	21.20	23	32
4A	38.40	17.20	21.20	28	41
1B	31.20	9.60	21.60	14	39
3A	38.40	15.20	23.20	10	18
1A	38.40	6.80	31.60	20	40

[1] profit per patient in case of providing the profit maximizing quantity to the patient
[2] profit per patient in case of providing the benefit maximizing quantity to the patient

Table B.4: P-values for Wilcoxon signed-ranks tests, comparing relative benefit losses from individual treatment decisions for matching pairs of patients, across payment systems

	Patient																			
	1A	1B	1C	1D	1E	2A	2B	2C	2D	2E	4A	4B	4C	4D	4E	3A	3B	3C	3D	3Ee
p-value	0.031437	0.041115	0.026577	0.833635	0.368402	0.353259	0.126382	0.002958	0.081902	0.003841	0.107470	0.022320	0.007271	0.000318	0.000004	0.475051	0.722563	0.972185	0.000757	0.000052

Appendix C: Post-experimental questionnaire analysis

Table C.1: Variables considered in the dimension reduction (Principal Component Analysis)

Sociability	Average values (from 1 to 7) attributed to the following statements : *I see myself as someone who is communicative.* *I see myself as someone who is conciliatory.* *I see myself as someone who can sometimes be rude to others. (reverted)* *I see myself as someone who is outgoing, sociable.* *I see myself as someone who is attentive and nice to others.* Cronbach's alpha = 0.6326
Stress	Average values (from 1 to 7) attributed to the following statements : *I see myself as someone who often worries.* *I see myself as someone who gets easily nervous.* *I see myself as someone who is cautious.* *I see myself as someone who is relaxed, can handle stress. (reverted)* Cronbach's alpha = 0.7154
Positive reciprocity	Average values (from 1 to 7) attributed to the following statements : *If someone makes me a favor, I am most likely to return it.* *I am particularly committed to help people who helped me in the past.* *I am ready to incur cost in order to help someone who helped me in the past.* Cronbach's alpha = 0.6245
Negative reciprocity	Average values attributed to the following statements : *If I suffer from injustice, I will take revenge at any cost and at the first occasion.* *I someone puts me in a difficult position, I will do the same to her.* *If someone insults me, I will be insulting toward this person.* *If someone has done me wrong, I do not forget easily.* *I tend to be resentful.* *When someone does me wrong, I try to forgive and forget. (reverted).* Cronbach's alpha = 0.8204
Trust	Binary choice between the two following statements : *Would you say that most people could be trusted or that one has always to be cautious?*
Risk Seeking	Value (from 0 to 10) attributed to the statement : *I am generally willing to take risks, I generally do not avoid risk.*
Patience	Value (from 0 to 10) attributed to the statement : *I am generally patient.*
Impulsivity	Value (from 0 to 10) attributed to the statement : *I do not take too much time to think before acting, I am rather impulsive.*
Self-esteem	Value (from 0 to 10) attributed to the statement : *I have a positive opinion of myself*

Table C.2: Factor loading – Principal component analysis

Variables	Factor 1 *Self-Confidence*	Factor 2 *Antisocial*	Factor 3 -	Uniqueness
Sociability	-	-0.5096	0.3412	0.2044
Stress	-0.5618	-	-	0.2168
Positive reciprocity	-	-0.3217	-	0.7531
Negative reciprocity	-	0.5442	-	0.2946
Trust	0.3264	-	-0.4916	0.4102
Risk Seeking	0.2879	0.4625	-	0.4488
Patience	-	-	0.6195	0.4821
Impulsivity	0.3522	-	-	0.4980
Self-esteem	0.3916	-	0.3562	0.4913
Proportion explained	0.2514	0.1873	0.1391	