Physician performance pay: Experimental evidence

Jeannette Brosig-Koch, Heike Hennig-Schmidt, Nadja Kairies-Schwarz,
Johanna Kokot, Daniel Wiesen

London School of Economics
Department of Health Policy Seminar

1 June 2021

supported by the German Research Foundation (DFG) and the *Zentralinstitut der Kassenärztlichen Bundesvereinigung* (Zi)

Outline

Why do we care?

- Understanding how physicians respond to incentives is important for policy-makers and researchers alike
- The traditional payment system: fee-for-service may incentivize "too many" services; overtreatment (e.g., Ellis and McGuire 1986, JHE)
- A prominent attempt to control costs: lump-sum capitation (CAP) payments (e.g., in managed care)
- CAP may lead to underprovision of medical services (e.g., Cutler 1995, ECMA)
- Pay for performance (P4P) programs are frequently suggested to improve the quality of health care (e.g., UK, USA)
- Ongoing health policy debate on the introduction and design of P4P

Mixed empirical evidence

- Inconclusive evidence on the effect of performance pay on the quality of care (e.g., Epstein 2012, NEJM; Witter et al. 2012, Cochrane Rev.; Eijkenaar et al. 2013, EJHE; Milstein and Schreyögg, 2016, HP)
- If at all, moderate effects (e.g., Mullen et al. 2010, RAND; Li et al. 2014, HE; Scott et al. 2018, MCRR)
- Possible reasons:
  - Biased or difficult to observe health outcomes (e.g., Campbell et al. 2009, NEJM; Gravelle et al. 2010, EJ; Roland and Olesen 2016, BMJ)
  - Simultaneous interventions (e.g., Cutler et al. 2004, AER; Kolstad 2013, AER)
  - Effects of P4P-design elements not well understood (e.g., Scott et al., 2018)
  - Self selection into payment schemes (e.g., Cadena and Smith, 2021)
  - Heterogeneity in physicians' responses typically not considered (e.g., Donato et al. 2017, AER)

Mixed empirical evidence

- Inconclusive evidence on the effect of performance pay on the quality of care (e.g., Epstein 2012, NEJM; Witter et al. 2012, Cochrane Rev.; Eijkenaar et al. 2013, EJHE; Milstein and Schreyögg, 2016, HP)
- If at all, moderate effects (e.g., Mullen et al. 2010, RAND; Li et al. 2014, HE; Scott et al. 2018, MCRR)
- Possible reasons:
  - Biased or difficult to observe health outcomes (e.g., Campbell et al. 2009, NEJM; Gravelle et al. 2010, EJ; Roland and Olesen 2016, BMJ)
  - Simultaneous interventions (e.g., Cutler et al. 2004, AER; Kolstad 2013, AER)
  - Effects of P4P-design elements not well understood (e.g., Scott et al., 2018)
  - Self selection into payment schemes (e.g., Cadena and Smith, 2021)
  - Heterogeneity in physicians' responses typically not considered (e.g., Donato et al. 2017, AER)

▷ Causal effect of performance pay on physicians' behavior and the quality of health care is difficult to infer using field data

Design of P4P: Size of bonus and unintended consequences

- How the size of the performance bonus affects physicians' medical service provision not well understood
- Unintended effects like a crowding-out of physicians' altruistic (patient-regarding) behavior and motivation might occur
- Other-regarding motivations are fundamental in public service provision (e.g., Besley and Ghatak 2005, AER; Prendergast 2007, AER; Delfgaauw and Dur 2008, EJ) particularly in health (Arrow 1963, AER)
- Financial incentives might lead to crowding-out of intrinsic motivation (e.g., Deci 1971; Frey et al. 1996, JPE; Frey 1997, EJ; Maynard 2012, HE)
- Some experimental evidence for motivation crowding-out (e.g., Gneezy and Rustichini 2000, QJE; Ariely et al. 2009, REStud; Huffman and Bognanno 2018, MS)

Design of P4P: Size of bonus and unintended consequences

- How the size of the performance bonus affects physicians' medical service provision not well understood
- Unintended effects like a crowding-out of physicians' altruistic (patient-regarding) behavior and motivation might occur
- Other-regarding motivations are fundamental in public service provision (e.g., Besley and Ghatak 2005, AER; Prendergast 2007, AER; Delfgaauw and Dur 2008, EJ) particularly in health (Arrow 1963, AER)
- Financial incentives might lead to crowding-out of intrinsic motivation (e.g., Deci 1971; Frey et al. 1996, JPE; Frey 1997, EJ; Maynard 2012, HE)
- Some experimental evidence for motivation crowding-out (e.g., Gneezy and Rustichini 2000, QJE; Ariely et al. 2009, REStud; Huffman and Bognanno 2018, MS)

▷ No causal evidence on the behavioral effect of bonus levels and on whether P4P crowds-out physicians' altruistic behavior

This paper

- Artefactual field experiment (Harrison and List 2004, JEL) with primary care physicians from a representative sample of resident physicians in Germany
- 'Clean' performance measure tied to the patient-optimal quality of medical care
- Within-subjects: Exogenous variation from CAP to blended CAP + P4P
- Between-subjects comparison of different bonus levels
- Random selection of subjects in experimental treatments
- Link of behavioral data to physicians' practice characteristics such as location and annual profit

Why an experiment?

Behavioral experiments: A complementary approach in health economics and health policy research (Galizzi and Wiesen 2018, ORE)

- Lab and artefactual field experiments are well suited to testing explicit predictions of simple theoretical models under controlled conditions.
- No patients are harmed due to unintended effects of an intervention.
- Experiments often provide unique opportunities to study behavior that is hidden or prohibited in the field.
- Experimental data, combined with field studies and social surveys, can help us understand sources of heterogeneity in behaviors.
- Experiments are highly replicable and scalable.
- Experiments are a good way to pre-test designs and behavioral mechanisms for more expensive and cumbersome field experiments and RCTs.

Why an experiment?

Behavioral experiments: A complementary approach in health economics and health policy research (Galizzi and Wiesen 2018, ORE)

- Lab and artefactual field experiments are well suited to testing explicit predictions of simple theoretical models under controlled conditions.
- No patients are harmed due to unintended effects of an intervention.
- Experiments often provide unique opportunities to study behavior that is hidden or prohibited in the field.
- Experimental data, combined with field studies and social surveys, can help us understand sources of heterogeneity in behaviors.
- Experiments are highly replicable and scalable.
- Experiments are a good way to pre-test designs and behavioral mechanisms for more expensive and cumbersome field experiments and RCTs.
  Lab and artefactual field experiments could be seen as the health economist's equivalent of animal trials in medical research.

Related behavioral experiments in health

- Fee-for-service, capitation, and salary:

  Hennig-Schmidt et al. (2011, JHE), Green (2014 JEBO), Hennig-Schmidt and Wiesen (2014, SSM), Brosig-Koch et al. (2016, JEBO); Lagarde and Blauuw (2017, SSM), Green et al. (2017, JEBO), Di Guida et al. (2019, HE); Reif et al. (2020, IJERPH); Wang et al. (2020, EER); Waibel and Wiesen (2021, EER)

- Mixed payment systems:

  Brosig-Koch et al. (2017, HE)

- P4P:

  Oxholm et al. (2021, SSM); Green et al. (2020, BMJ Quality and Safety) Brosig-Koch et al. (2021)

Research questions

1. How does performance pay affect physicians' behavior?
2. Does the bonus level affect physicians' behavior (Low bonus of 5% vs. High bonus of 20% on top of baseline CAP)?
3. How do physicians' practice characteristics relate to their medical service provision?
4. Does performance pay crowd-out physicians' patient-regarding (altruistic) behavior?

Our physician sample

- Overall, 104 primary care physicians (PCPs) participated in our artefactual field experiment
- Sub-sample ($\sim$10%) of PCPs enrolled in the Zi practice panel (ZiPP) which comprises a representative sample of resident physicians in Germany
- ZiPP is run annually with about 5,000 resident physicians
- In Germany, around 54,000 resident PCPs contract with the statutory health insurance (GKV), about 1,000 PCPs participate in the ZiPP

Sample characteristics

- Average age: 56 years (ZiPP: 54, German PCPs: ∼53 years)
- Share of female PCPs: 35% (ZiPP: 39% German PCPs: ∼44%)
- Distribution of locations similar to ZiPP
  - City: ∼30%; ZiPP: ∼34%
  - Outer conurbation: ∼36%; ZiPP: ∼37%
  - Rural: ∼34%; ZiPP: ∼29%
- Annual profit: ∅150,383 EUR (ZiPP: ∅158,733 EUR)
- Our sample is not significantly different from non-participating PCPs of the ZiPP

Experimental design

- Within-subject design: Introduction of P4P with two different bonus levels

Experimental condition	First payment system	Second payment system	# Sub. (# pat.)
Low bonus (5%)	CAP	CAP+P4P-5%	53 (954)
High bonus (20%)	CAP	CAP+P4P-20%	51 (918)

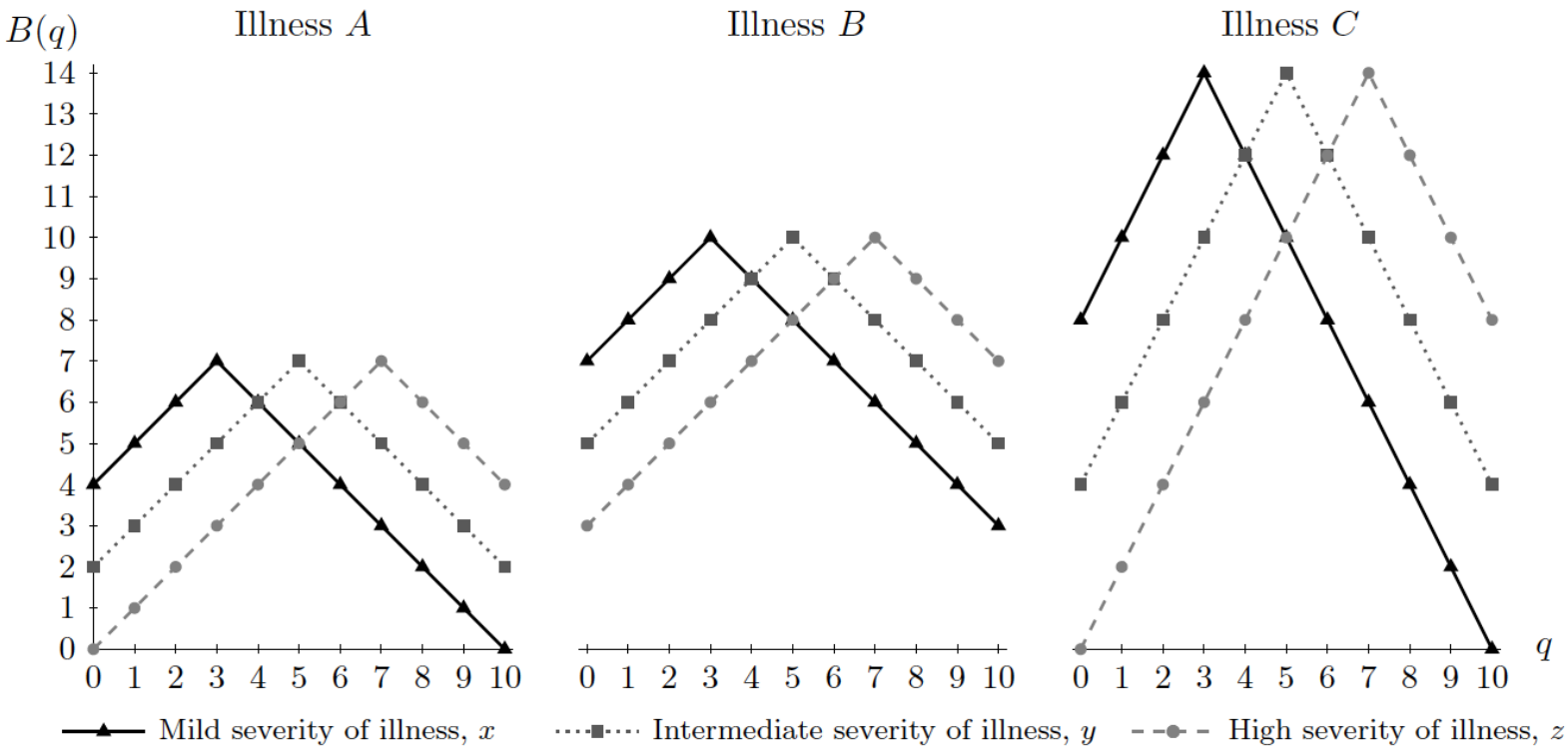
- Between-subject comparison for performance-pay systems
- Control treatments with medical students

Decision situation

- Framed physician decision-making experiment
- Physicians decide on the quantity of medical services q
- Individual decisions on $q \in \{0, 1, \ldots 10\}$ for 9 abstract patients
- Subjects simultaneously determine profit and the patient's health benefit (measured in monetary terms)
- Framing and setting are the same for all payment systems

Patients' health benefit

- Systematic variation of health benefits; constant for all payment systems
- Illnesses A, B, C with three severities x (mild), y (interm.), z (high)



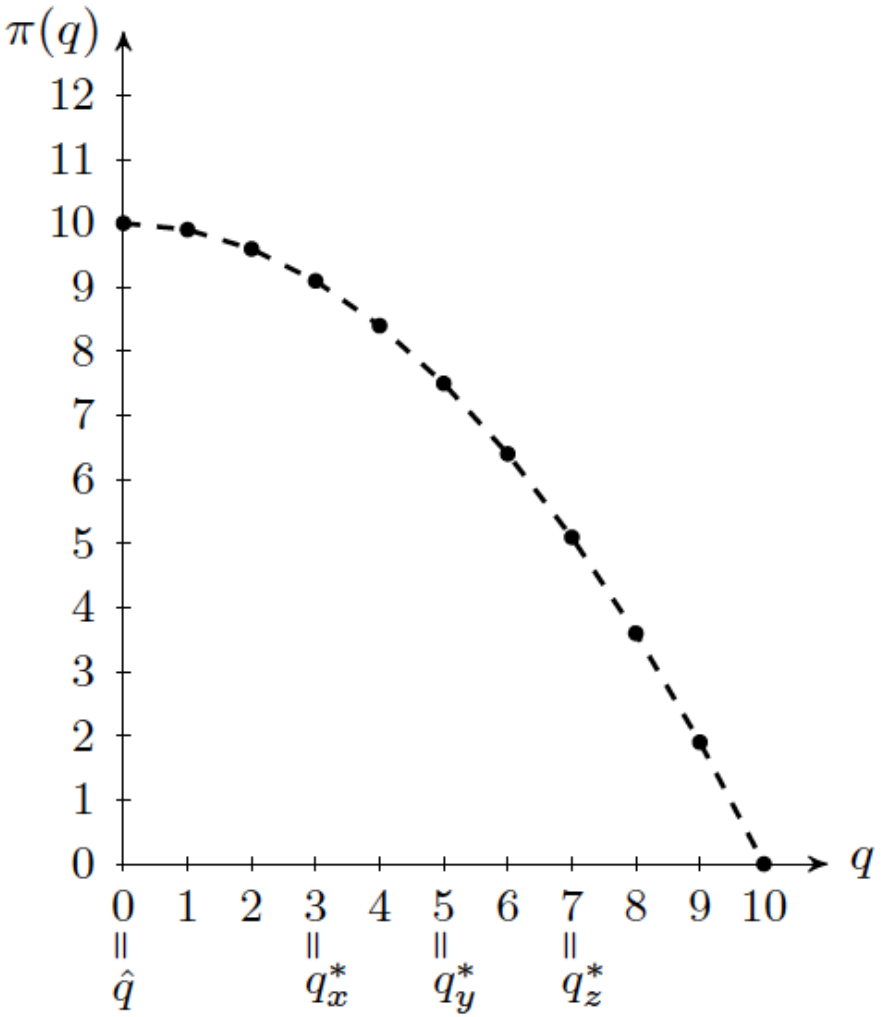
- Salient incentive: Patients' health benefit measured in monetary terms, benefits real patients' health outside the lab

Payment systems

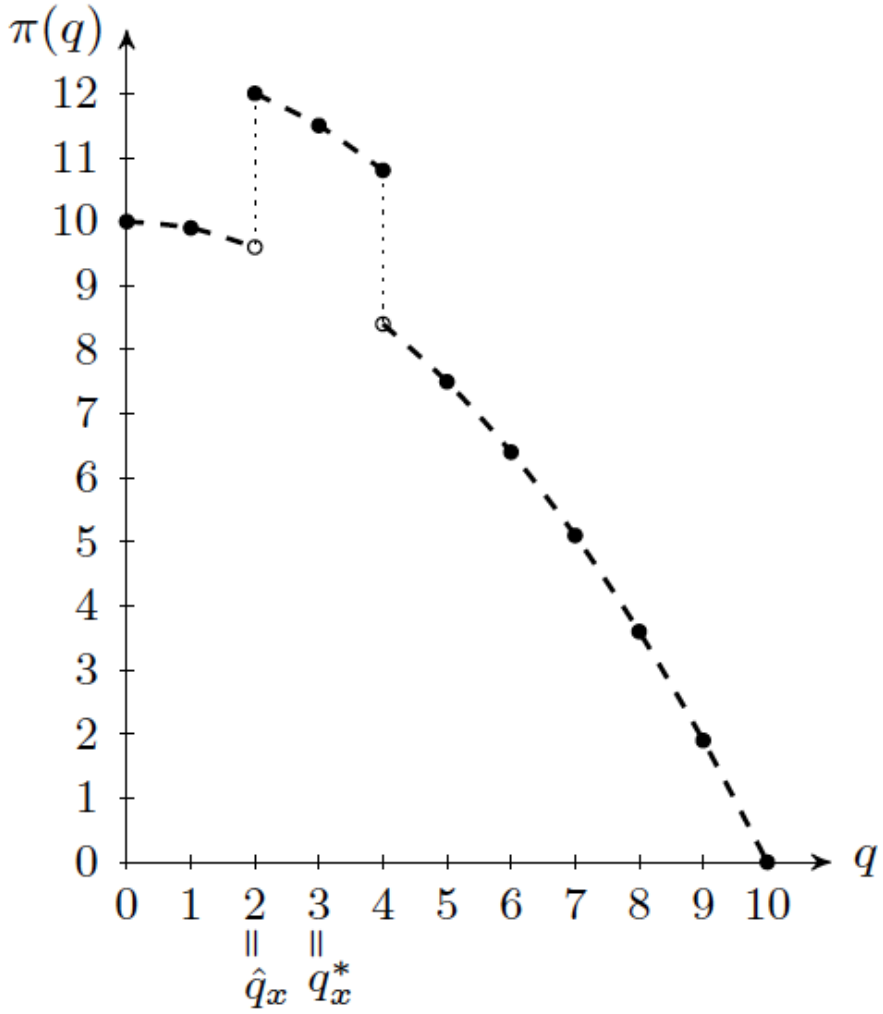
- CAP: lump-sum payment of 25 EUR for physicians
- Performance pay linked to patients' benefit (health outcome) and adjusted for severities of illness
- Discrete bonus is granted if quality threshold is reached $|q - q^*| \leq 1$
- Reflects asymmetric information between payer and physician
- Cost are convex $c(q) = q^2/10$

Parameters: Illustration of physicians' profits

CAP



CAP+P4P



Sample decision screen

Patient with illness B, mild severity (x)

Round 1: Patient 1 **Link to instructions**

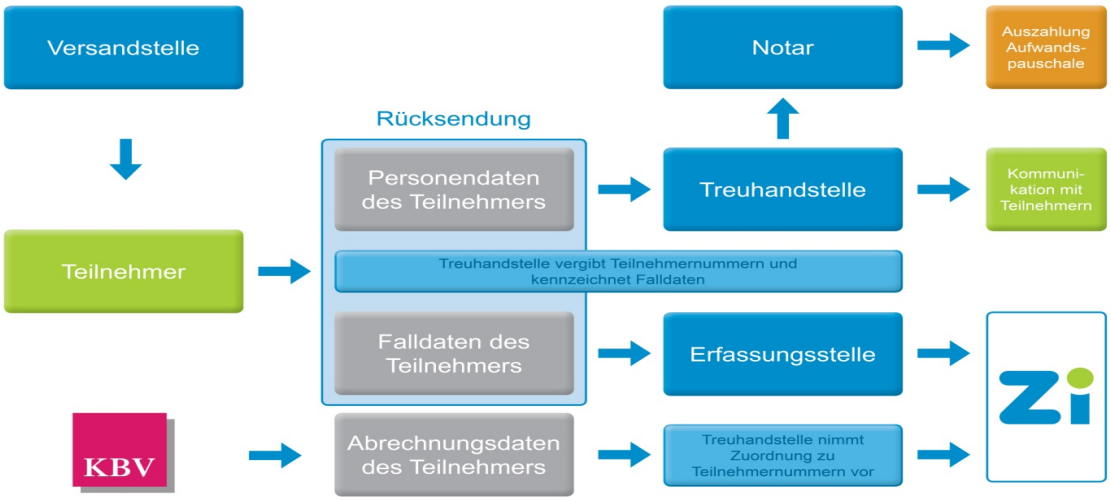
Quantity of medical services	Your lump-sum remuneration (in Euro)	Your bonus payment (in Euro)	Your costs (in Euro)	Your payoff = remuneration + bonus – costs (in Euro)	Benefit of the patient with illness B and severity x (in Euro)
0	25	0.00	0.00	25.00	17.5
1	25	0.00	0.25	24.75	20.0
2	25	2.25	1.00	26.25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

Which quantity of medical services do you want to provide?

send...

Facilitation of the artefactual field experiment

ZiPP: Data collection procedure



- Double-blind procedure
- Anonymity of subjects ensured
- Experiment followed the data security guidelines of the ZiPP
- Payment procedure via notary office

Experimental protocol

- Experiments with physicians were run in March 2016; average duration of about 30 minutes
- Post experimental questionnaire (e.g., risk attitudes, altruism)
- Random payment technique: One decision is randomly selected for payment in each part
- Average payment per subject: 45.93 EUR (total: 4,823 EUR)
- Average payment per patient: 47.64 EUR (total: 5,003 EUR)
- Behavioral data linkage: Administrative data on practice characteristics (e.g., annual profit, location) are provided by Zi

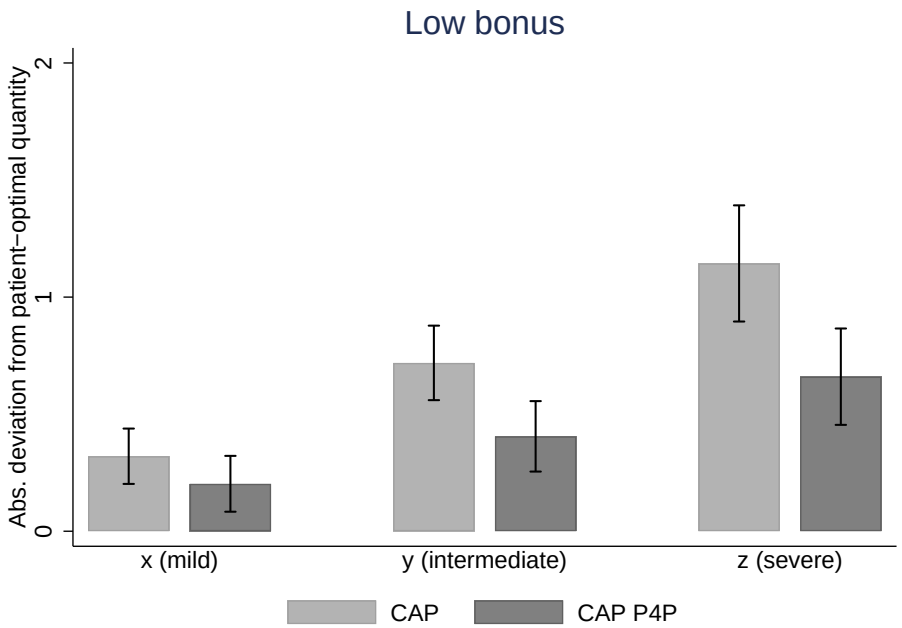
Behavioral results

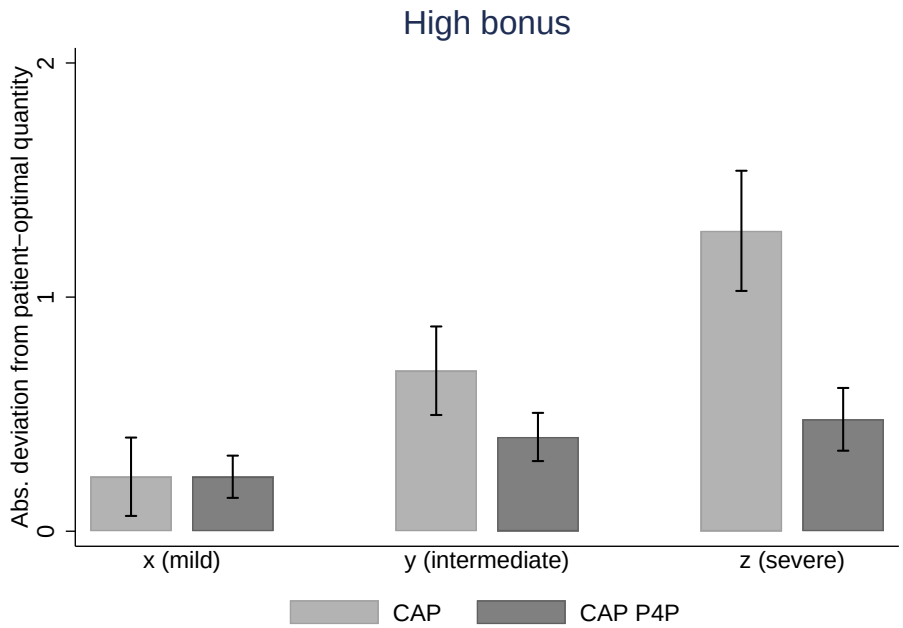
Physicians' medical service provision in CAP (first part of the experiment)

- Physicians significantly underprovide medical services in CAP for patients with intermediate and high severity of illness ($p \leq 0.014$, Wilcoxon signed-rank test; comparison with q^* for all illnesses)
- Underprovision increases in patients' severity of illness, patients' marginal benefit does not significantly affect behavior
- Consistent with findings in the experimental literature
  (e.g., Hennig-Schmidt et al. 2011, JHE; Brosig-Koch et al. 2017, HE)

How performance pay affects physicians' behavior

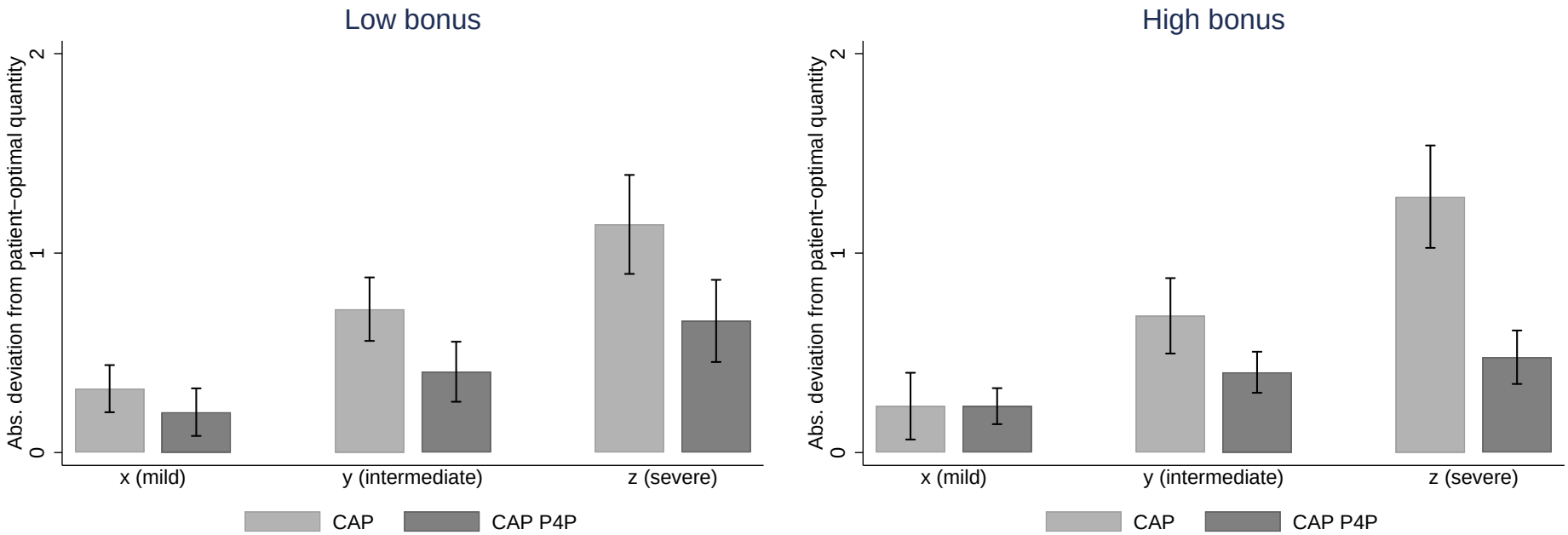
Deviation from the patient-optimal quantity





How performance pay affects physicians' behavior

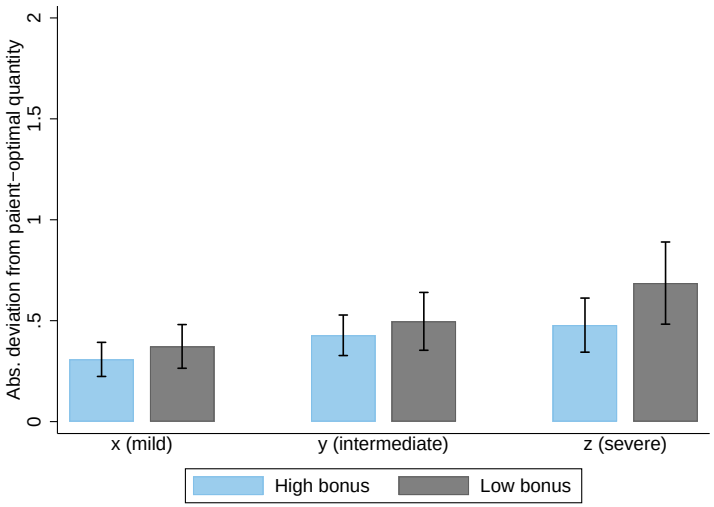
Deviation from the patient-optimal quantity



- Underprovision is significantly reduced for intermediately (y) and severely ill (z) patients in CAP+P4P-20% and CAP+P4P-5% ($p \leq 0.094$, Wilcoxon signed-rank test)
- For mild severity patients (x), the reduction in underprovision is not significant ($p > 0.162$)

Does the size of the bonus affect behavior?

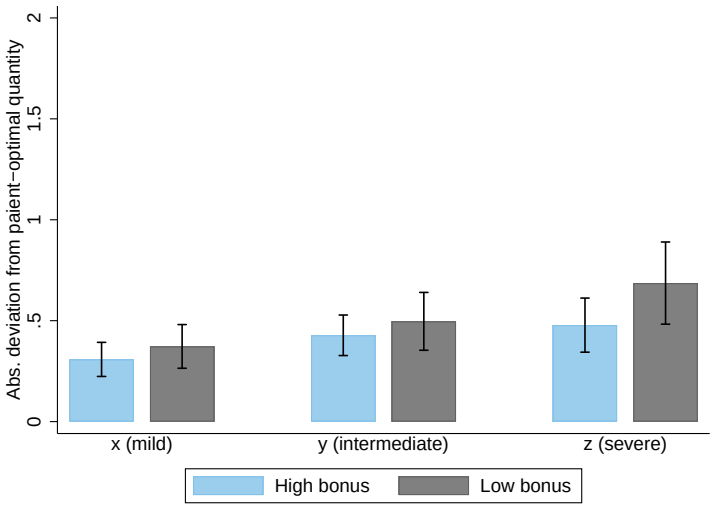
Absolute deviation from the patient-optimal quantity (second part of the experiment)



- Very similar behavioral responses for the two different bonus levels
- No statistically significant differences ($p > 0.4964$, Mann-Whitney U-Test)

Does the size of the bonus affect behavior?

Absolute deviation from the patient-optimal quantity (second part of the experiment)



- Very similar behavioral responses for the two different bonus levels
- No statistically significant differences ($p > 0.4964$, Mann-Whitney U-Test)

▷ The bonus level does not significantly affect physicians' behavior.

Physicians' characteristics and the quality of care

Multilevel mixed effects regressions on the relative quality of care

Model:	(1)	(2)	(3)	(4)	(5)
Performance pay (P4P)	0.068***	0.068***	0.068***	0.055***	0.072***
	(0.007)	(0.007)	(0.007)	(0.009)	(0.012)
High annual profit ($>$ 147k EUR)	-0.050*		-0.058**	-0.072**	-0.058**
	(0.027)		(0.028)	(0.029)	(0.028)
City		-0.030	-0.042	-0.042	-0.042
		(0.036)	(0.037)	(0.037)	(0.038)
Outer conurbation		0.005	-0.003	-0.003	0.003
		(0.033)	(0.034)	(0.034)	(0.035)
P4P x High annual profit				0.029**	
				(0.014)	
P4P x City					-0.001
					(0.017)
P4P x Outer conurbation					-0.012
					(0.017)
Constant	0.815***	0.784***	0.812***	0.819***	0.810***
	(0.057)	(0.059)	(0.066)	(0.066)	(0.066)
Observations	1764	1764	1764	1764	1764
Physicians	98	98	98	98	98

Notes. This table shows parameter estimates (fixed effects) from multilevel mixed-effects REML regressions. All models include subject-specific random effects and controls for gender, years of practice and bonus size. Standard errors are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

So far...

- P4P significantly increases the quality of care
- Quality in the experiment is lower for physicians with high annual practice profit
- Physicians with high annual profits respond significantly stronger to P4P incentives
- Physicians' location does not significantly affect the quality of care

Unintended consequences: Crowding-out of patient-regarding behavior

Descriptive analysis

- Analysis is based on how (104x9) individual patients are treated in both parts
- Behavioral patterns:
  - Profit maximization (PM)
  - Benefit maximization (BM)
  - Trade-off (TO)
- Behavioral patterns by part of the experiment:
  - 1st part (CAP): PM: 1%; BM: 54%; TO: 42%; Other: 3%
  - 2nd part (CAP+P4P): PM: 30%; BM: 64%; TO: 0%; Other: 6%
- Transitions:
  - Crowding out: BM $\longrightarrow$PM: 7% ($\sim$ 14% of BM); TO $\longrightarrow$ PM: 22%
  - Crowding in: PM $\longrightarrow$ BM: 1%; TO $\longrightarrow$ BM: 17%

Main takeaways

- Controlled artefactual field experiments to test the effect of introducing performance pay on physicians' behavior
- Underprovision in CAP is significantly reduced under performance pay
- Patients' severities of illness affect physicians' behavior
- Surprisingly, the level of the bonus pay does *not* significantly affect physicians' behavior
- Physicians with higher practice profits respond significantly stronger to P4P
- Non-negligible evidence for crowding-out of patient-regarding behavior

Some policy implications...

...within the confines of our experiment

Gains in patient benefit and additional remuneration cost

- Increase of health benefit:
  - Low bonus: 8%
  - High bonus: 7.5%

Some policy implications...

...within the confines of our experiment

Gains in patient benefit and additional remuneration cost

- Increase of health benefit:
  - Low bonus: 8%
  - High bonus: 7.5%

- Arc-elasticity of patient benefit with respect to remuneration (similar to Brot-Goldberg et al. 2017, QJE):
  - Low bonus: 0.18
  - High bonus: 0.08

Some policy implications...

...within the confines of our experiment

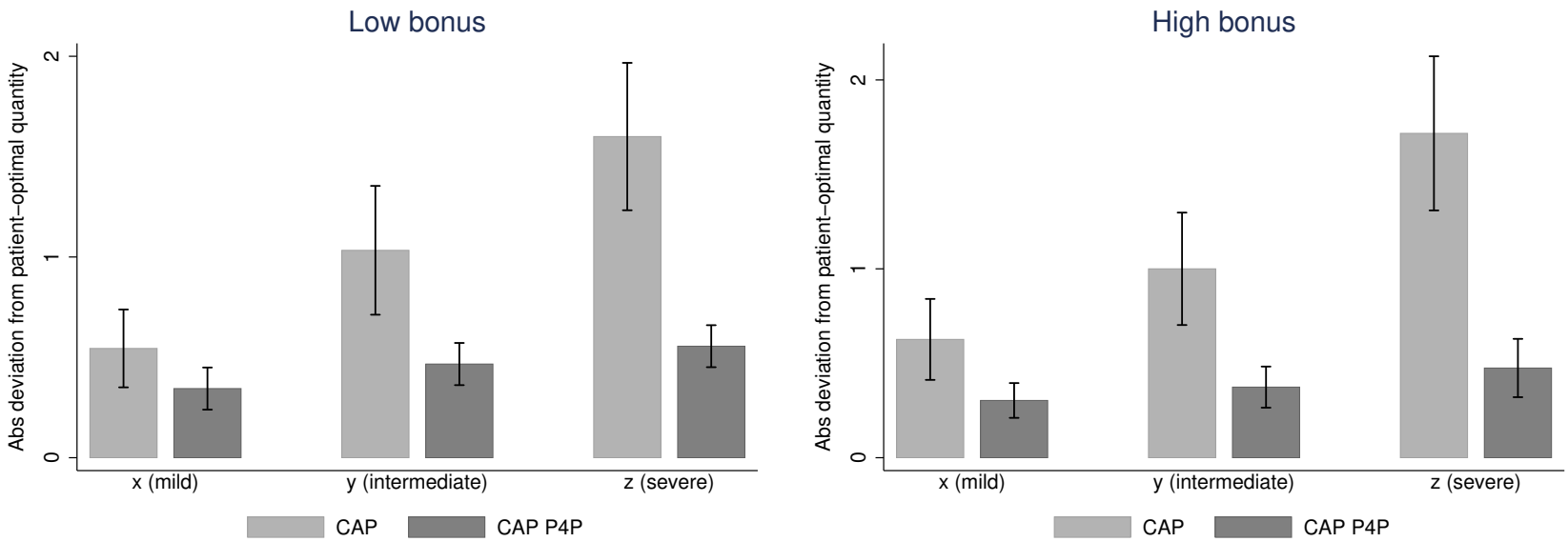
Gains in patient benefit and additional remuneration cost

- Increase of health benefit:
  - Low bonus: 8%
  - High bonus: 7.5%
- Arc-elasticity of patient benefit with respect to remuneration (similar to Brot-Goldberg et al. 2017, QJE):
  - Low bonus: 0.18
  - High bonus: 0.08
- Low bonus sufficient to change behaviors and more cost efficient

THANK YOU!

APPENDIX

Does the behavior of physicians and med. students differ?



- Within-subjects: Underprovision in CAP is significantly reduced under performance pay
- Level of bonus pay does not significantly affect students either

▷ Performance pay affects students' behavior very similarly.

Robustness of results: Evidence from control treatments with medical students

- "Taking performance pay away" (reverse order) does not affect medical students behavior in a significant way compared to introducing performance pay
- No significant differences under constant maximum incentives (increased capitation)

▷ Findings are robust across subject pools and towards order of payment systems as well as levels of incentives.