Empir Econ
https://doi.org/10.1007/s00181-018-1425-8

Do pay-for-performance incentives lead to a better health outcome?

Alina Peluso[1] · **Paolo Berta**[2] · **Veronica Vinciotti**[1]

Received: 14 March 2017 / Accepted: 6 February 2018

Abstract Pay-for-performance approaches have been widely adopted in order to drive improvements in the quality of healthcare provision. Previous studies evaluating the impact of these programs are limited by either the number of health outcomes or of medical conditions considered. In this paper, we evaluate the effectiveness of a pay-for-performance program on the basis of five health outcomes and across a wide range of medical conditions. The context of the study is Lombardy region in Italy, where a rewarding program was introduced in 2012. The policy evaluation is based on a difference-in-differences approach. The model includes multiple dependent outcomes, which allow quantifying the joint effect of the program, and random effects, which account for the heterogeneity of the data at the ward and hospital level. Our results show that the policy had a positive effect on the hospitals' performance in terms of those outcomes that can be more influenced by a managerial activity, namely the number of readmissions, transfers and returns to surgery room. No significant changes which can be related to the pay-for-performance introduction are observed for the number of voluntary discharges and for mortality. Moreover, our study shows evidence that the medical wards have reacted more strongly to the pay-for-performance program than the surgical ones, whereas only limited evidence is found in support of a different policy reaction across different types of hospital ownership. Finally, the evaluation found no evidence of a distortion of the hospital behavior aimed at inflating the performance evaluation, such as cream skimming behavior.

✉ Alina Peluso
alina.peluso@brunel.ac.uk

[1] Department of Mathematics, Brunel University London, London, UK

[2] Department of Statistics and Quantitative Methods, CRISP, University of Milan-Bicocca, Milan, Italy

Published online: 14 June 2018

Keywords Pay-for-performance · Difference-in-differences · Multilevel modeling · Policy evaluation · Hospital effectiveness

1 Introduction

Quality improvement is the principal strategy of any healthcare system. For this reason, there is a strong focus on assessment and redesign of the work process and of the systems themselves in order to lower the costs and to deliver care that is safer and that results in the best outcome for patients. The adoption of a pay-for-performance (P4P) approach aims to drive the hospitals in this direction. The idea behind the implementation of a P4P approach is quite simple: in order to improve the overall quality delivered, healthcare providers are given the opportunity to have their reimbursements increased when they achieve specified quality benchmarks (Eijkenaar et al. 2013; Alshamsan et al. 2010). From an economics perspective, the hospital is considered as a profit-maximizer agent which is encouraged to compete for quality in order to obtain a financial reward, rather than to attract more patients. Therefore, a P4P program is considered efficient when an improved quality of care is achieved with equal or lower costs (Emmert et al. 2012). Clearly, the evaluation of the quality delivered is a crucial part to every P4P approach. While quality in health care is a broad concept composed of different dimensions, such as efficiency, evaluation of standard, appropriateness and customer satisfaction, P4P programs refer to the healthcare system's quality mostly in terms of its effectiveness (Van Herck et al. 2010).

Due to the potential of P4P programs, in recent years there has been a growing interest in the application of these programs to the healthcare systems of different countries. These studies are collected in several systematic reviews (Van Herck et al. 2010; Eijkenaar 2012; Petersen et al. 2006), but mixed results transpire about the impact of the programs to the quality of care. The aim of the current paper is to contribute to the existing literature by providing a thorough evaluation of a P4P program and its effect on the overall quality of the healthcare system. The study discussed in this paper pertains to Lombardy region of Italy, previously identified as a suitable context for the adoption of a P4P program (Castaldi et al. 2011). In 2012, a tailored P4P program was introduced to control the amount of the annual budget provided to each hospital on the basis of their effectiveness. In line with the designs adopted by previous studies (Rosenthal et al. 2005; Lindenauer et al. 2007), 9 hospital wards covering a wide range of medical conditions were exogenously selected for the treatment group and were subjected to the P4P program, whereas the other hospital wards were not involved in the program. Data were collected both two years prior and two year post- introduction of the policy for all hospitals in the Lombardy region. The aim of this paper is then to evaluate the effect of the policy on the basis of the data collected.

As in the evaluation of any policy, a choice needs to be made about which health outcome to use for quantifying the impact of the P4P program. In many studies, a single outcome is considered. For example, Sutton et al. (2012) quantify the impact of the P4P adoption in England by analyzing the hospital overall mortality. In addition, the evaluation of P4P programmes is often confined to specific clinical conditions, such as acute myocardial infarction (AMI), coronary artery bypass graft surgery (CABG),

heart failure, pneumonia and hip/knee replacement (Jha et al. 2012; Levin-Scherz et al. 2006; Glickman et al. 2007; Shih et al. 2014; Sutton et al. 2012). In contrast to these studies, we analyze the P4P effect using five different health outcomes and based on the overall case-mix hospitalizations of the wards considered. Moreover, for the first time in a P4P study, we investigate the policy effect with regard to hospital ownership, by evaluating possible different reactions to the P4P program among the private (for-profit and not-for-profit) and public providers, and also with regard to the different wards, by evaluating whether surgical and medical wards reacted differently to the policy.

The article proceeds as follows: in Sect. 2 we describe the healthcare system in Lombardy and the adopted P4P program; in Sect. 3, we describe the chosen methodological approach; in Sect. 4, we present the data used in the analysis, and in Sect. 5, we discuss the results of the policy evaluation. Section 6 concludes the paper.

2 The healthcare system and the P4P program in Lombardy

The Italian healthcare system provides universal healthcare coverage. The state government guarantees the essential levels of assistance (LEA) over all regions of the country. Each region has administrative and executive freedom of implementation of the LEA, and citizens may freely choose the healthcare provider. The Italian NHS is funded mainly from general taxation. Financial resources for NHS are transferred from the state to a regional budget and are then managed by the local healthcare system (Martini et al. 2014). Among the 21 regions in Italy, Lombardy is one of the top-ranked for socio-demographic indicators and one of the most competitive areas in Europe according to economic indicators. Lombardy has a population of 10 million residents, equal to 16% of the total Italian population, with a density of 404 inhabitants per km^2. The Lombardy healthcare system comprises of circa 150 hospitals generating around 1.6 million discharges annually, with circa 18 billion Euro allocated for the healthcare spending (circa 75% of the regional budget) every year.

A regional reform in 1997 radically transformed the healthcare system in Lombardy into a quasi-market healthcare system in which citizens can freely choose the provider regardless of its ownership (private for-profit, private not-for-profit or public). The healthcare system in Lombardy is entirely built on a prospective payment system based on diagnosis-related groups (DRGs), with a maximum annual reimbursement defined by a budget yearly allocated by the region to each hospital (Martini et al. 2014). The 1997 reform also established that the Lombardy administration is responsible for monitoring the effectiveness of the health care provided by the hospitals belonging to the regional accreditation system (Brenna 2011).

As a consequence, the Lombardy regional healthcare directorate developed a set of measures to systematically evaluate the performance of the hospitals in terms of the quality supplied. The details of this process are given in Berta et al. (2013) and in the regional resolution (p. 4 of ACT 349 2012). The following outcome measures have been selected: overall mortality (in-hospital mortality + 30 days after discharge), number of transfers to a different hospital, number of discharges against medical advice, number of returns to the surgery room, and number of repeated hospitalizations

or readmissions. The choice of these outcomes was based both on their popularity in the scientific literature, i.e., mortality and readmissions, and on the necessity of driving hospitals toward a reduction in the number of adverse outcomes, such as voluntary discharges, return to the surgery room and transfers to a different hospital.

In 2012, a new policy was introduced, whereby the increment of the hospital annual budget is based on a weighted mean of the hospital's evaluated outcomes. The hospitals are ranked according to this measure: the first hospital in the ranking receives an increment of 2% of its annual budget, the worst one gets a penalty of 2%, whereas all the others receive an amount between the interval $[-2\%, +2\%]$, and proportional to the distance between their score and the score of the last hospital in the category's ranking (p. 84 of ACT 2633 2011; ACT 349 2012). In the first instance, the regional healthcare management decided to evaluate the weighted outcome measures only on 9 wards, i.e., cardiology, cardiosurgery, neurosurgery, neurology, oncology, general medicine, urology, orthopedic, surgery. The wards were chosen according to the coverage within the hospitals, the inclusion of both medical and surgical disciplines as well as the level of specialization (cardiosurgery and neurosurgery). Further details on the policy introduction can be found in the regional resolution (ACT 2633 2011). It is interesting to note that the incentive is provided to the hospital as a whole, as typical of P4P programmes in health care (Cashin et al. 2014). The individual hospitals have then a large accountability on how the allocate the incentive payments. Typically, provider institutions allocate the financial resources to make general improvements in the service delivered, and in particular related to the performance measures. In the case of the Lombardy region, it is also possible that the physicians and/or nurses working in the treated wards received a direct bonus as a drive to performance improvement. This is, however, bound to vary across hospitals, so we do not expect to see the impact of this in our policy evaluation.

3 The econometric approach

We test the effect of the policy using a difference-in-differences (DID) approach (Abadie 2005; Blundell et al. 2004) on data between 2010 and 2013 (two year pre- and two year post-policy). To justify the suitability of this approach, the following considerations are needed:

1. The wards are split into a *treatment* group—the 9 wards that are used for the hospital evaluation—and a *control* group—the remaining wards. The allocation of the wards in one of these groups was made exogenously prior to the introduction of the policy (ACT 2633 2011). There is an underlying assumption here that, although the incentive is provided to the hospital as a whole, the incentive is dictated only by the performance of the wards *treated*. Combined with the fact that the individual wards operate autonomously, the *untreated* wards can be considered as an independent group. A similar analysis was conducted by Sutton et al. (2012), where the treatment and control groups are defined within each hospital on the basis of selected diagnoses.
2. Units do not switch between the control and the treatment group: improvements in performance of the control group do not affect the financial incentives gained

by the hospital. We will, however, test whether there is evidence of a distortion of the hospital behavior aimed at inflating the performance evaluation, such as the lift of resources in favor of the treated wards.

3. Any macro-changes affect both groups equally and differences between the treatment and the control group remain constant in the absence of treatment, i.e., a parallel trend prior to treatment. The check of this assumption is going to be discussed later in the results section. Of notice is also the fact that the regional resolution was formally announced in December 2011 (ACT 2633 2011) and applied from early January 2012 (ACT 349 2012). Thus, hospitals had no possibility to anticipate changes.

As discussed in Sect. 2, the policy evaluation is based on five health outcomes. Given the mix of patients in the different wards, the outcomes are first adjusted by patients characteristics via the use of a multilevel logistic mixed effect model (Snijders 2011; Goldstein 2011). This model allows to account for the hierarchical structure of the data whereby patients are clustered into wards and wards are nested into hospitals. In addition, the longitudinal structure of the data means that a time effect is also to be expected. In detail, let Y_{pwht} represent a binary health outcome for patient p (with $p = 1, \ldots, P_{wht}$) in the ward w (with $w = 1, \ldots, W_{ht}$), belonging to the hospital h (with $h = 1, \ldots, H_t$), hospitalized at time t (in years, $t = 2010, \ldots, 2013$). Let π_{pwht} be the conditional probability of Y_{pwht} being equal to 1. We consider the model

$$\log\left(\frac{\pi_{pwht}}{1-\pi_{pwht}}\right) = \alpha + \eta X_{pwht} + \mu_{wht} + \nu_{ht}, \tag{1}$$

where η is a vector of coefficients for the X_{pwht} patient-level covariates, μ_{wht} is a random effect of the ward w nested within hospital h at time t, capturing the latent heterogeneity of the wards, whereas ν_{ht} captures the latent heterogeneity of the hospital h at time t. μ_{wht} and ν_{ht} are independent and identically distributed, $N(0, \sigma_\mu^2)$ and $N(0, \sigma_\nu^2)$, respectively, and are assumed to be uncorrelated with the regressors.

The model in Eq. (1) returns the patients' predicted probabilities

$$\hat{\pi}_{pwht} = \frac{\exp\left(\hat{\alpha} + \hat{\eta}\, X_{pwht} + \hat{\mu}_{wht} + \hat{\nu}_{ht}\right)}{1 + \exp\left(\hat{\alpha} + \hat{\eta}\, X_{pwht} + \hat{\mu}_{wht} + \hat{\nu}_{ht}\right)}, \tag{2}$$

which we collapse at the ward level over time in order to obtain the average predicted health outcome

$$\mathrm{HO}_{wht_m} = \frac{\sum_{p \in P_{wht_m}} \hat{\pi}_{pwht}}{|P_{wht_m}|}, \tag{3}$$

where P_{wht_m} is the set of patients admitted in the ward w of the hospital h in the month m ($m = 1, \ldots, 12$) of the year t and $|P_{wht_m}|$ is the cardinality of this set.

The aim is now to quantify the policy effect on the basis of the five (adjusted) health outcomes. As we anticipate a correlation between the five health outcomes, we consider a multivariate DID model, rather than a separate model for each outcome. In this way, we are able to quantify the overall effect of the policy across all health outcomes, as well as at the individual level. Let then $\mathrm{HO}^{(\theta)}_{wht_m}$ denote the health outcome θ, namely

readmissions ($\theta = 1$), mortality ($\theta = 2$), return to the surgical room ($\theta = 3$), transfers ($\theta = 4$) and voluntary discharges ($\theta = 5$), at month m of year t ($t = 2010, \ldots, 2013$) of ward w ($w = 1, \ldots, W_h$) belonging to hospital h (with $h = 1, \ldots, H$). We consider the following multivariate mixed model:

$$\begin{aligned} \text{HO}^{(\theta)}_{wht_m} = \alpha^{(\theta)}_h + \beta^{(\theta)} \text{TREATED}_{wh} + \sum\nolimits_{j=2011}^{2013} \gamma^{(\theta)}_j I(j = t) \\ + \sum\nolimits_{j=2011}^{2013} \delta^{(\theta)}_j \left(I(j = t) \cdot \text{TREATED}_{wh}\right) + \upsilon^{(\theta)} \text{MONTH}_{t_m} + \epsilon^{(\theta)}_{wht_m}, \end{aligned} \tag{4}$$

where the dummy variable TREATED_{wh} indicates whether the ward w is in the treatment group or not, the indicator variable $I(j = t)$ indexes the four years of the study (two pre- and two post- policy), with 2010 set as reference category, MONTH is a continuous variable, taking values 1–48 and added to correct for a possible seasonality effect, $\alpha^{(\theta)}_h$ is the random hospital effect for outcome θ, and the error $\epsilon^{(\theta)}_{wht_m} = (\epsilon^{(1)}_{wht_m}, \ldots, \epsilon^{(5)}_{wht_m})$ has a multivariate distribution $\epsilon_{wht_m} \sim N(0, \Sigma)$, with the covariance Σ accounting for possible dependencies between the different outcomes. The parameter $\delta^{(\theta)}_j$ is of interest in this model. Under the assumption of a parallel trend pre-policy, we expect $\delta^{(\theta)}_{2011} = 0$ for all outcomes, whereas the parameters $\delta^{(\theta)}_{2012}$ and $\delta^{(\theta)}_{2013}$ represent the DID of average outcomes between the treated and control wards from the pre- to the post-policy years. The two different parameters for the post-policy period let us detect whether the impact of the policy was immediate in the first year of its introduction or whether it was delayed in the second year (Ayyagari and Shane 2015). This model allows us to detect the effect of the policy across all wards.

A second objective of the study is to detect whether the reaction to the P4P adoption is different depending on the ward's type. In particular, we group all wards into two types: surgical and medical, and extend the model in Eq. (4) to:

$$\begin{aligned} \text{HO}^{(\theta)}_{wht_m} = \alpha^{(\theta)}_h + \beta^{(\theta)} \text{TREATED}_{wh} + \sum\nolimits_{j=2011}^{2013} \gamma^{(\theta)}_j I(j = t) \\ + \sum\nolimits_{k=1}^{2} \lambda^{(\theta)}_k I(k = \text{SURGICAL}_{wh}) \\ + \sum\nolimits_{j=2011}^{2013} \left(\delta^{(\theta)}_j I(j = t) \cdot \text{TREATED}_{wh}\right) \\ + \sum\nolimits_{j=2011}^{2013} \sum\nolimits_{k=1}^{2} \left(\mu^{(\theta)}_{jk} I(j = t) \cdot I(k = \text{SURGICAL}_{wh})\right) \\ + \sum\nolimits_{k=1}^{2} \left(\nu^{(\theta)}_k I(k = \text{SURGICAL}_{wh}) \cdot \text{TREATED}_{wh}\right) \\ + \sum\nolimits_{j=2011}^{2013} \sum\nolimits_{k=1}^{2} \left(\tau^{(\theta)}_{jk} I(j = t) \cdot I(k = \text{SURGICAL}_{wh}) \cdot \text{TREATED}_{wh}\right) \\ + \upsilon^{(\theta)} \text{MONTH}_{t_m} + \epsilon^{(\theta)}_{wht_m}, \end{aligned} \tag{5}$$

with the variable SURGICAL defined as 1 if the prevalent activity of the ward is surgical and 0 otherwise. In this model, the DID parameters $\tau^{(\theta)}_{jk}$, $j = (2012, 2013)$, are of interest as they represent the differences in average outcomes between the

surgical treated wards and the surgical control wards, from the pre- to the post-policy period and with respect to the medical wards which are taken as the reference category. For this model, we do not consider the health outcome returns to the surgery room as this is observed only for the surgical wards.

Finally, in the results section, we also consider a similar model for the detection of possible differences in the reaction to the P4P adoption depending on the type of hospital ownership. In particular, we compare private for-profit, private not-for-profit and public hospitals. Due to the more strict budget constrains for private hospitals, these hospitals may react more actively to the policy than public ones. Furthermore, private for-profit hospitals are more oriented toward profit than the other hospitals and may therefore be more driven to increase their outcome measures in order to obtain a financial reward.

4 Data and descriptive statistics

The database was gathered from the Lombardy healthcare information system. Data were collected on patients admitted to 142 hospitals during the four years 2010–2013 (two before and two in the policy-on period). In this period, the hospitals provided 3,581,389 hospitalizations, coded in the available hospital discharge chart. In our analysis, we included patients admitted for acute care and we excluded patients living outside the region, patients younger than two years old or patients hospitalized in day-hospital, rehabilitation or palliative treatments.

Table 1 provides details for the variables considered in the study and the five outcomes. We used variables at both the patient and ward/hospital level. At the patient level, there is information on their gender, age, number of transit to the intensive care unit during hospitalization, the weight of the financial reimbursement corresponding to the patient's disease, and the comorbidity index. The latter is measured as in Elixhauser et al. (1998) and indicates the presence of one or more additional diseases or disorders co-occurring with a primary disease or disorder. At the hospital level, we know whether the hospital is affiliated to a medical school in which medical students receive practical training, whether the hospital is mono-specialistic or general, and whether there is presence of high-technology instrumentation in the ward. Finally, we include the hospitals' ownership, which categorizes the hospital as private for profit, private not-for-profit or public, and we distinguish wards whose prevalent activity is surgical from the medical ones. The effectiveness of the policy is evaluated over the five health outcomes described in the previous section, namely mortality, readmissions, transfers, returns and voluntary discharges. We should clarify that the outcome return to the surgery room can be evaluated only for the surgical wards.

Table 1 reports the average (and the standard deviations in brackets) of the variables in the dataset by treatment and across the four years of the study (two pre- and two post- policy). It appears that the mix of patients within the treated and untreated wards is relatively stable over time, but that there are differences between the two groups. In particular, patients that are admitted to the treated wards are on average older than those admitted to the untreated ward. In addition, the treated wards consider higher-risk patients than the untreated wards in terms of DRGs weight, number of

Table 1 Sample means and standard deviations in brackets for the covariates in the study from the Lombardy hospital inpatient stays for each year before and after the policy introduction

	Untreated				Treated			
	Pre-policy		Post-policy		Pre-policy		Post-policy	
	2010	2011	2012	2013	2010	2011	2012	2013
Patient								
MALE	0.2589	0.2613	0.2646	0.2673	0.5399	0.5413	0.5397	0.5383
	(0.43)	(0.43)	(0.44)	(0.44)	(0.49)	(0.49)	(0.49)	(0.49)
AGE	46.076	46.585	46.973	47.212	64.526	64.877	65.054	65.384
	(21.1)	(21.1)	(21.2)	(21.3)	(18.7)	(18.5)	(18.6)	(18.5)
DRGWEIGHT	0.892	0.9127	0.9139	0.919	1.2974	1.3252	1.3167	1.3277
	(0.81)	(0.84)	(0.83)	(0.85)	(1.12)	(1.15)	(1.12)	(1.13)
COMORBIDITY	0.2379	0.2128	0.2156	0.2099	0.4082	0.3303	0.325	0.3121
	(0.58)	(0.55)	(0.56)	(0.55)	(0.72)	(0.66)	(0.65)	(0.64)
INTCARE	0.015	0.0164	0.017	0.0174	0.0644	0.0676	0.0677	0.0687
	(0.12)	(0.12)	(0.12)	(0.13)	(0.24)	(0.25)	(0.25)	(0.25)
Ward/Hospital								
TECHNOLOGY	0.8585	0.8588	0.8614	0.8683	0.8079	0.807	0.8111	0.8119
	(0.34)	(0.34)	(0.34)	(0.33)	(0.39)	(0.39)	(0.39)	(0.39)
TEACHING	0.2684	0.2708	0.2754	0.2734	0.2455	0.2456	0.2471	0.2456
	(0.44)	(0.44)	(0.44)	(0.44)	(0.43)	(0.43)	(0.43)	(0.43)
SPECIALISED	0.052	0.0474	0.0482	0.049	0.0387	0.0386	0.0406	0.0393
	(0.22)	(0.21)	(0.21)	(0.21)	(0.19)	(0.19)	(0.19)	(0.19)
SURGICAL	0.5637	0.5535	0.5646	0.562	0.5088	0.4884	0.4942	0.487
	(0.49)	(0.49)	(0.49)	(0.49)	(0.49)	(0.49)	(0.5))	(0.49)
OWN:NOPROFIT	0.0758	0.0765	0.077	0.0793	0.0947	0.0948	0.0975	0.096
	(0.26)	(0.26)	(0.26)	(0.27)	(0.29)	(0.29)	(0.29)	(0.29)
OWN:PROFIT	0.1376	0.1373	0.1346	0.1264	0.2314	0.2354	0.2308	0.2327
	(0.34)	(0.34)	(0.34)	(0.33)	(0.42)	(0.42)	(0.42)	(0.42)
OWN:PUBLIC	0.7866	0.7862	0.7884	0.7943	0.6739	0.6698	0.6717	0.6713
	(0.49)	(0.49)	(0.49)	(0.49)	(0.49)	(0.49)	(0.5))	(0.49)
Outcomes								
TRANSFERS	0.0056	0.0052	0.0036	0.0035	0.0127	0.0127	0.0053	0.0051
	(0.07)	(0.07)	(0.06)	(0.05)	(0.11)	(0.11)	(0.07)	(0.07)
RETURN	0.0592	0.0632	0.0099	0.0108	0.0431	0.0443	0.0154	0.0161
	(0.23)	(0.24)	(0.09)	(0.10)	(0.20)	(0.20)	(0.12)	(0.12)
MORTALITY	0.0268	0.0276	0.029	0.0273	0.0593	0.0608	0.0611	0.0601
	(0.16)	(0.16)	(0.16)	(0.16)	(0.23)	(0.23)	(0.23)	(0.23)
READMISSIONS	0.1216	0.1149	0.1117	0.1091	0.1335	0.1277	0.1211	0.1111
	(0.32)	(0.31)	(0.31)	(0.31)	(0.34)	(0.33)	(0.32)	(0.31)
VOLDISCH	0.0084	0.0085	0.0082	0.0084	0.0088	0.0081	0.0076	0.007
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.08)	(0.08)	(0.08)

comorbidities and intensive treatment. The percentage of comorbidities (roughly 30%) is, however, still relatively small compared to other countries, e.g., 0.69% in Northern Ireland in 2011/2012 (Reilly et al. 2015). This is justified by the coding rules that affect the healthcare system in Lombardy, whereby only the comorbidities directly connected with the treated DRGs are registered. Considering the variables related to the hospitals and the wards, we observe that the overall composition of the hospitals has not changed during the policy period, with surgical wards covering around 51% of the overall admissions. Moreover, 71% of the hospitalizations are provided by the public hospitals, whereas 30% of the patients are admitted to a private provider (20% in the for-profit hospitals and 9% in the not-for-profit). With regard to the health outcome measures, three out of the five outcomes, namely transfers, return to the surgery room and readmissions, show a reduction after the introduction of the P4P program. The aim is to assess the significance of this finding after adjusting for the patient-level covariates identified in Table 1 using Eq. (1).

5 Policy evaluation

In order to assess whether there has been an improvement in the healthcare quality following the introduction of the P4P policy, we use a multivariate DID approach as discussed in Sect. 3. Table 2 reports the fixed effects estimates of the model in Eq. (4). As all outcomes are constrained to be between 0 and 1, the parameter estimates and the p values are computed by a nonparametric bootstrap approach. For this, we use a method specifically developed for multilevel modeling (Wang et al. 2011; Carpenter et al. 2003).

5.1 Testing the assumptions of a DID approach for policy evaluation

Table 2 shows how the parameters δ^{θ}_{2011} of the interaction between TREATED and YEAR_{2011} are not significantly different from zero. This provides evidence in favor of the parallel trend assumption for each individual health outcome, i.e., the differences between the average outcome of the treatment and control group are constant prior to the introduction of the policy. This assumption is needed in order to evaluate the impact of the policy using a DID approach. As we require a parallel trend to be satisfied for all health outcomes simultaneously, we use a multivariate analysis of variance test (MANOVA) to test the null hypothesis $H_0 : \delta^{(1)}_{2011} = \ldots \delta^{(5)}_{2011} = 0$ under the model in Eq. (4). The Wilks' lambda statistics returns a p value of 0.2676, which provides further evidence in support of the parallel trend assumption across all health outcomes.

Given that the incentive is provided to the hospital as a whole, it is also necessary to test whether the introduction of the P4P may have had a negative spillover effect between the treated and the untreated wards. This would violate the assumption of independence between the two groups and thus bias the policy evaluation. Although within each ward the physicians and nurses detain managerial freedom on whether and how to treat the patients, spillover effects could take the form of hospitals lifting resources in favor of the treated wards to the expense of the untreated wards. To

Table 2 Estimates for the fixed effects for the model in Eq. (4)

	MORTALITY	READMISSIONS	RETURN	TRANSFERS	VOL. DISCH.
MONTHS	0.001	−0.001	0.001	−0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
TREATED	0.02***	0.004***	−0.037***	0.006***	0.001
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
$YEAR_{2010}$	0.044***	0.13***	0.084***	0.009***	0.009***
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)
$YEAR_{2011}$	0.044***	0.125***	0.082***	0.008***	0.008***
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)
$YEAR_{2012}$	0.045***	0.122***	0.021***	0.006*	0.008**
	(0.003)	(0.003)	(0.005)	(0.003)	(0.003)
$YEAR_{2013}$	0.041***	0.118***	0.022***	0.005	0.008**
	(0.004)	(0.004)	(0.006)	(0.004)	(0.004)
TREATED·$YEAR_{2011}$ (δ_{2011})	0.002	0.001	0.002	0.001	−0.001
	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)
TREATED·$YEAR_{2012}$ (δ_{2012})	0.001	−0.005***	0.026***	−0.005***	−0.001
	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)
TREATED·$YEAR_{2013}$ (δ_{2013})	0.005***	−0.011***	0.025***	−0.005***	−0.001
	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)

The coefficients and standard errors (in brackets) are reported
***Significance at the 1% level
**Significance at the 5% level
*Significance at the 10% level

this aim, we assess whether there has been a difference in the total number of hours worked by physicians and nurses within each hospital between the treated and the untreated wards from the year 2011 (pre-policy) to 2012 (post-policy). We consider 58 hospitals which have a balanced proportion of treated/untreated wards. Figure 1 shows the box plot of the number of hours worked by hospital and year. The figure shows how, within each hospital, the number of hours worked is stable across the two groups and between the pre- and post-policy period, suggesting that no shift of resources occurred, at least at the level of labor. This is supported by a nonsignificant p value for the year–treatment interaction term (0.812) from a negative binomial generalized linear model which includes also fixed effects for hospitals. In addition to the allocation of resources, another possible spillover effect could result from the sharing of technological resources between the different wards. This may have an impact on surgical outcomes, such as the return to the surgery room in our case. We have no data to evaluate this, but we will take this into consideration when interpreting the results of the policy evaluation analysis.

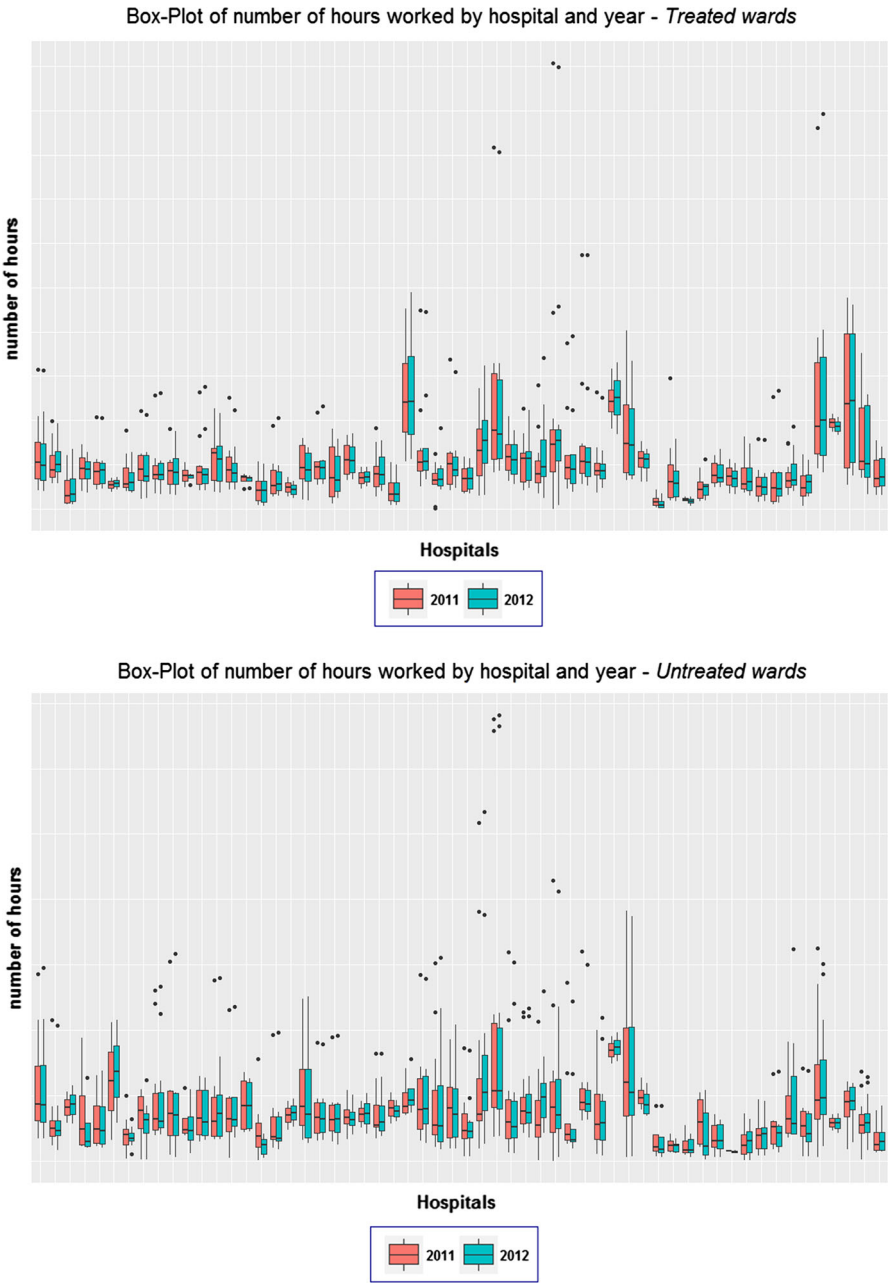

Fig. 1 Box plot of the number of hours worked by hospital and year for the *Treated* (top) and *Untreated* (bottom) wards

Together with the spillover effects mentioned above between wards within the same hospital, the different providers may have also reacted to the policy by avoiding to treat high-risk patients (Levaggi and Montefiori 2013). In order to check for this potential distortion, we have analyzed whether the cream skimming index, calculated as in Berta et al. (2010), changed significantly between the pre- and the post-policy period. As above, we restrict the analysis to the hospitals which have a balanced proportion of treated/untreated wards and we perform the pre–post analysis separately for the treated and untreated groups. Using a multiple regression model, we find only four hospitals (out of 58) with a significant negative interaction with the post-policy term, two for the treated wards (p values: $4.54E-08$, 0.0025) and two for the untreated ones (p values: 0.02, 0.0314). Thus, we conclude that overall the hospitals show no evidence of a gaming behavior in selecting the mix of patients in the post-policy period.

5.2 Do the hospitals react positively to the policy?

We are now in a position to evaluate the impact of the P4P policy by considering the estimates of the coefficients of the interaction between the treatment variable and the post-policy years in Table 2, i.e., δ^{θ}_{2012} and δ^{θ}_{2013}. As all health outcomes are improved if they are reduced, a significant and negative coefficient for these interactions would mean that the P4P introduction had a positive effect on quality. This result is confirmed for readmissions ($\delta_{2012} = -0.0051, \delta_{2013} = -0.0112$) and transfers ($\delta_{2012} = -0.0046, \delta_{2013} = -0.0047$). This is a clear signal that the hospital activity was modified as a result of the P4P introduction, as both readmissions and transfers are directly affected by the hospital organization. In particular, the results show that the P4P program may have reduced the hospital attitude of readmitting patients in order to increase the number of the DRGs provided (Berta et al. 2010). The reduction in the transfers of the patients between hospitals in the treated wards is also particularly encouraging, considering that transfers are directly linked to the patient safety and continuity of care.

In order to further quantify the impact of the policy and to confirm the significance of the results on the health outcomes in absolute terms, Fig. 2 plots the marginal effects of each health outcome in Eq. (4) for treated and untreated wards and over the observation period (Karaca-Mandic et al. 2012; Ai and Norton 2003). As well as verifying the parallel trend in the pre-policy period, the plots show a clear improvement for readmissions and transfers. In particular, there is an absolute difference of 0.91 and 1.52% in the average number of readmissions between the treated and untreated wards in the year 2012 and 2013, respectively, and of 0.31% in the year 2011, whereas there is a difference of 0.19 and 0.18% in the average number of transfers between the treated and untreated wards in the year 2012 and 2013, respectively, and of 0.72% in the year 2011. This leads to DID reductions of 0.60% (readmissions) and 0.53% (transfers) in 2012 compared to 2011 and a further reduction of 0.61% (readmissions) and 0.01% (transfers) in 2013. The predicted percentages of reduction correspond to a P4P-related saving of 4324 readmissions and 4295 transfers in the treated wards in 2012 and a further reduction of 4871 readmissions and 157 transfers in 2013.

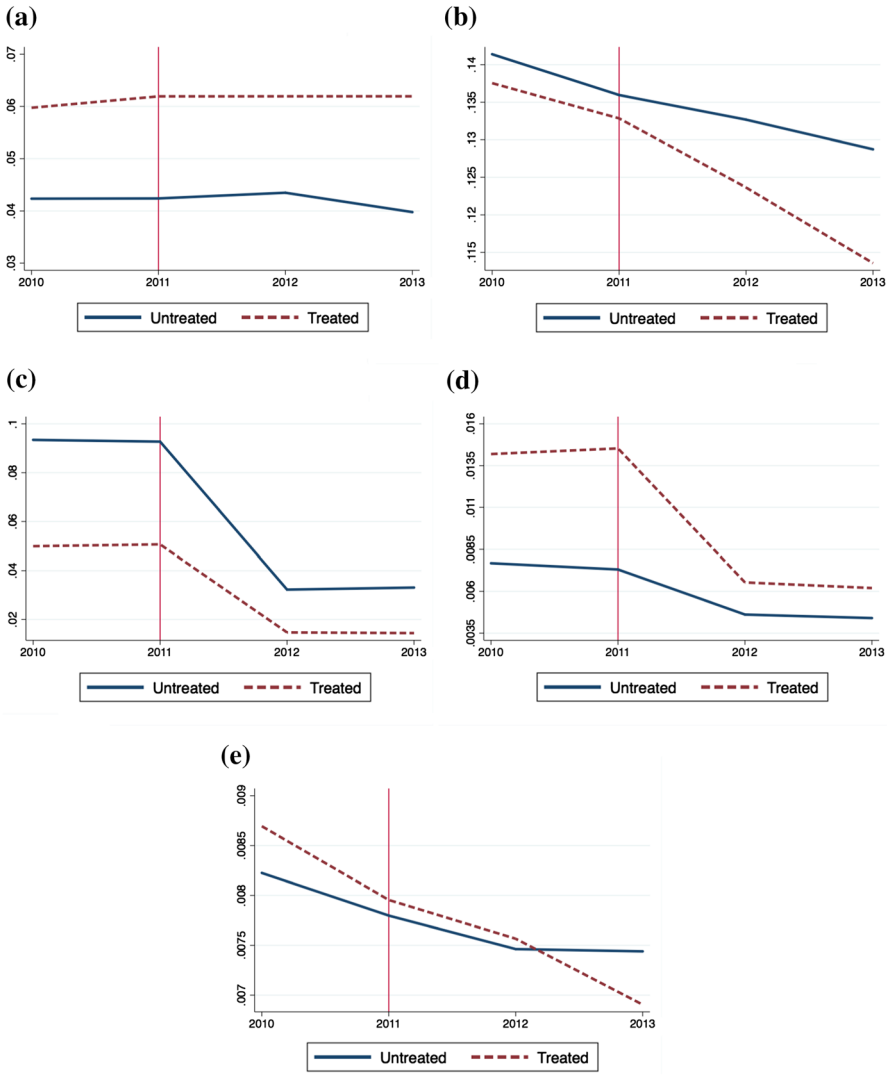

Fig. 2 Marginal effects of all health outcomes per year and treatment for the model in Eq. (4). (**a**) Expected Mortality, (**b**) Expected Readmissions, (**c**) Expected Returns to OR, (**d**) Expected Transfers, (**e**) Expected Voluntary discharges

The picture for the other three health outcomes is more complex than for transfers and readmissions. The average number of returns to the surgery room seems to increase in the treated wards more than in the untreated after the introduction of the policy, as δ_{2012} and δ_{2013} are positive and significant. This is shown in Fig. 2, which, on the other hand, shows also how the P4P incentives improve the performance for both the treated and untreated wards. This is an interesting result, suggesting that the managerial impact in the hospital organization caused by the adoption of the P4P program has changed the overall hospital performance with regard to the surgical activity. A possible explanation

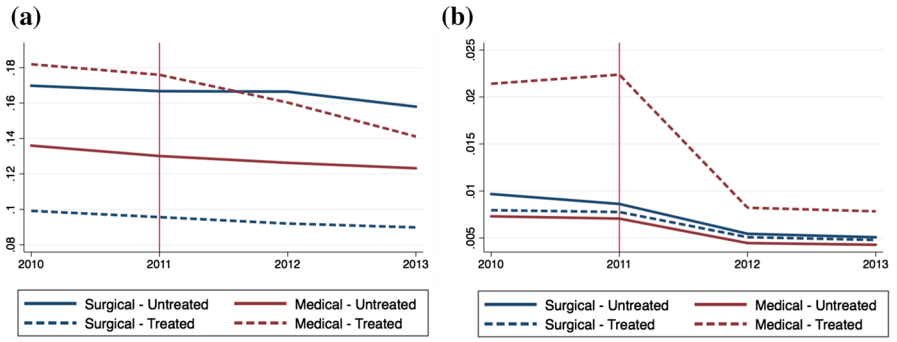

Fig. 3 Marginal effects of readmissions and transfers per type of ward, year and treatment for the model in Eq. (5). **a** Expected readmissions, **b** expected transfers

to this could be given by a spillover effect between the treated and the untreated wards, as all wards may be benefitting from potentially improved technology in the surgery room.

For the other two health outcomes, voluntary discharges and mortality, the coefficients of δ_{2012} and δ_{2013} are not significantly different from zero. Figure 2 shows how the number of voluntary discharges decreases already before the P4P introduction. With regard to mortality, it is reasonable to believe that, when hospitals are checked for effectiveness on more than one output, they will focus on those outcomes that are easily measurable. This is observed by Propper et al. (2008) in the context of a competition analysis. From this point of view, readmissions, transfers and return to the surgery room represent well-measured outcomes. Hence, it is possible that hospitals have focused their efforts on those easily measured and better observable activities in order to increase their performance and then gain financial rewards.

5.3 Do surgical and medical wards react differently to the policy?

We fit the model in Eq. (5) to the data in order to answer this question. The results, omitted in full for brevity, show evidence of a differential impact of the P4P introduction for the two health outcomes that were significant in the global analysis above. In particular, there is evidence that the P4P program impacted more on the medical wards than on the surgical ones in terms of number of readmissions ($\tau_{2012} = 0.008$, p value $= 0.0102$; $\tau_{2013} = 0.0307$, p value $=< 0.0001$) and number of transfers ($\tau_{2012} = 0.0117$, p value $= 0.0002$, $\tau_{2013} = 0.012$, p value $= 0.0001$). This is shown visually also by the marginal effects in Fig. 3. This finding can be explained by the fact that the surgical healthcare pathways are more rigorous and more linked to fixed guidelines than those on medical hospitalizations, which instead tend to be more flexible and more dependent on managerial actions and hospital organization.

5.4 Do private and public hospitals react differently to the policy?

Previous studies have found no dependency between hospital ownership and efficiency (Barbetta et al. 2007) or hospital ownership and competition (Berta et al. 2016),

suggesting that the long-term adoption of a quasi-market system in Lombardy has reduced the expected differences between the hospital types.

In this paper, we test whether the hospitals reacted differently to the introduction of the P4P policy, depending on their ownership. In order to answer this question, we use a model like Eq. (5), but with SURGICAL replaced by a variable representing the ownership type (OWN), where public is taken as the reference category. Once again, the interactions $\tau_{jk}^{(\theta)}$ are of interest in this model. In line with the existing literature, the results show only limited evidence in support to a hypothesis of a different reaction: apart from readmissions in 2012 ($\tau_{2012,\text{not-for-profit}} = -0.01964$, p value $= 0.0004$; $\tau_{2012,\text{private}} = -0.0096$, p value $= 0.0062$), the interaction for readmissions in 2013 and all interactions for transfers, for both the private for-profit and not-for-profit categories, are not statistically significant. This is an interesting result meaning that the monetary incentive is an interesting motivation to improve the quality of care for all types of ownership and not only for the profit-maximizer providers (profit hospitals).

6 Conclusions

The P4P approach has been adopted in many countries in order to encourage improvements in the quality of health care by supplying financial incentives to healthcare providers. In this study, we evaluate the impact of a specific P4P program adopted in the Lombardy region (Italy) in 2012. Differently to previous studies, we perform the analysis considering the whole healthcare system, evaluating multiple health outcomes over a number of clinical areas. We analyze data over four years, two before (2010/2011) and two after (2012/2013) the implementation of the program. The policy was applied to all hospitals in the Lombardy region, but the incentive was calculated only on the basis of the performance of 9 wards. The fact that the selection of these wards was made exogenously, combined with the fact that we observe a parallel trend pre-introduction of the policy and that we have found no evidence of spillover effects between the treated and untreated wards in terms of allocation of resources, have led us to use a multivariate DID approach for the evaluation of the impact of the policy.

Our study shows that two out of the five health outcomes considered, i.e., readmissions and transfers, support the hypothesis that the P4P introduction had a positive effect on quality. The picture for the other three health outcomes is more complex than for transfers and readmissions. Considering the returns to the surgery room, our results show that the P4P incentives improve the performance for both the treated and untreated wards. We speculate that this may be the result of improved technology in the surgery room which all the wards have benefitted from. The last two health outcomes, voluntary discharges and mortality, did not show changes that can be attributed to the P4P adoption. This can be explained by considering the fact that when hospitals are checked for effectiveness on more than one output, they will focus on those outcomes which are more easily driven by a managerial intervention in order to improve their performance and to obtain the financial incentives.

Moreover, our study shows that the medical wards have reacted to the P4P program more strongly than the surgical wards, whereas only limited evidence is found to suggest that the policy reaction was different across different types of hospital owner-

ship. Overall, the results show that the healthcare system in Lombardy was positively impacted by the P4P implementation, as anticipated by Castaldi et al. (2011): there is evidence of a reduction in some adverse health outcomes and of a general change in the hospital organization in order to improve the healthcare services provided to the citizens. Lastly, the evaluation study found no evidence of a distortion of the hospital behavior aimed at inflating the performance evaluation, such as cream skimming behavior.

This study has some implications. Firstly, Lombardy should extend the adoption of the P4P program across the whole regional healthcare system in order to improve the overall hospital activity. Secondly, given the positive impact of the P4P program in Lombardy, the adoption of a similar strategy is suggested to the other regional healthcare systems in Italy. This would stimulate improvements in quality for the regions that already perform relatively well, but, in particular, this would be an important incentive for these regions with a lower qualified healthcare system.

Future work on the evaluation of P4P programs could explore additional aspects, for which data were currently not available. Firstly, it would be interesting to test the impact of the P4P program in terms of the number of intra-hospital infections and complications, or other outcomes directly related to the performance of the hospitals' physicians and the improvement of technology. Secondly, it would be useful to conduct a comparative analysis between the Lombardy region and neighboring regions which are not subjected to P4P programmes. This would help also in controlling for spillover effects between the treated and the untreated wards within the same hospital, such as those resulting from the sharing of common technology and resources. Thirdly, our analysis has focussed solely on the impact of the P4P programs on the hospital effectiveness. It would be interesting to extend the current analysis to understand whether the monetary incentive had an impact also on the hospital efficiency. Finally, we believe that further research is needed to assess the impact of P4P programs over a long time frame, as encouraged by Werner et al. (2011).

References

Abadie A (2005) Semiparametric difference-in-differences estimators. Rev Econ Stud 72(1):1–19

ACT 2633 Lr (2011) Determinazioni in ordine alla gestione del servizio socio sanitario regionale per l'esercizio 2012

ACT 349 Lr (2012) Approvazione del metodo per l'individuazione dell'indice sintetico di performance per le strutture di ricovero

Ai C, Norton EC (2003) Interaction terms in logit and probit models. Econ Lett 80(1):123–129

Alshamsan R, Majeed A, Ashworth M, Car J, Millett C (2010) Impact of pay for performance on inequalities in health care: systematic review. J Health Serv Res Policy 15(3):178–184

Ayyagari P, Shane DM (2015) Does prescription drug coverage improve mental health? Evidence from medicare part d. J Health Econ 41:46–58

Barbetta GP, Turati G, Zago AM (2007) Behavioral differences between public and private not-for-profit hospitals in the Italian national health service. Health Econ 16(1):75–96

Berta P, Callea G, Martini G, Vittadini G (2010) The effetcs of upcoding, creamskimming and readmissions on the italian hospitals efficiency modelling: a populaiton-based investigtion. Econ Model 27(4):789–890

Berta P, Seghieri C, Vittadini G (2013) Comparing health outcomes among hospitals: the experience of the Lombardy region. Health Care Manag Sci 16(3):245–257

Berta P, Martini G, Moscone F, Vittadini G (2016) The association between asymmetric information, hospital competition, and the quality of health care: evidence from italy. J R Stat Soc Ser A 179:907–926

Blundell R, Meghir C, Dias MC, Reenen JV (2004) Evaluating the employment impact of a mandatory job search program. J Eur Econ Assoc 2(4):569–606

Brenna E (2011) Quasi-market and cost-containment in Beveridge systems: the Lombardy model of Italy. Health Policy 103(2):209–218

Carpenter JR, Goldstein H, Rasbash J (2003) A novel bootstrap procedure for assessing the relationship between class size and achievement. J R Stat Soc Ser C 52(4):431–443

Cashin C, Chi YL, Smith P et al (2014) Paying for performance in health care: implications for health system performance and accountability. Open University Press, Maidenhead

Castaldi S, Bodina A, Bevilacqua L, Parravicini E, Bertuzzi M, Auxilia F (2011) Payment for performance (p4p): Any future in italy? BMC Public Health 11(1):1

Eijkenaar F (2012) Pay for performance in health care: an international overview of initiatives. Med Care Res Rev 69(3):251–276

Eijkenaar F, Emmert M, Scheppach M, Schöffski O (2013) Effects of pay for performance in health care: a systematic review of systematic reviews. Health Policy 110(2):115–130

Elixhauser A, Steiner C, Harris DR, Coffey RM (1998) Comorbidity measures for use with administrative data. Med Care 36(1):8–27

Emmert M, Eijkenaar F, Kemter H, Esslinger AS, Schöffski O (2012) Economic evaluation of pay-for-performance in health care: a systematic review. Eur J Health Econ 13(6):755–767

Glickman SW, Ou FS, DeLong ER, Roe MT, Lytle BL, Mulgund J, Rumsfeld JS, Gibler WB, Ohman EM, Schulman KA et al (2007) Pay for performance, quality of care, and outcomes in acute myocardial infarction. JAMA 297(21):2373–2380

Goldstein H (2011) Multilevel statistical models, vol 922. Wiley, New York

Jha AK, Joynt KE, Orav EJ, Epstein AM (2012) The long-term effect of premier pay for performance on patient outcomes. N Engl J Med 366(17):1606–1615

Karaca-Mandic P, Norton EC, Dowd B (2012) Interaction terms in nonlinear models. Health Serv Res 47(1pt1):255–274

Levaggi R, Montefiori M (2013) Patient selection in a mixed oligopoly market for health care: the role of the soft budget constraint. In Rev Econ. https://doi.org/10.1007/s12232-013-0175-3

Levin-Scherz J, DeVita N, Timbie J (2006) Impact of pay-for-performance contracts and network registry on diabetes and asthma hedis® measures in an integrated delivery network. Med Care Res Rev 63(1 suppl):14S–28S

Lindenauer PK, Remus D, Roman S, Rothberg MB, Benjamin EM, Ma A, Bratzler DW (2007) Public reporting and pay for performance in hospital quality improvement. N Eng J Med 356(5):486–496

Martini G, Berta P, Mullahy J, Vittadini G (2014) The effectiveness-efficiency trade-off in health care: the case of hospitals in Lombardy, Italy. Reg Sci Urban Econ 49:217–231

Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S (2006) Does pay-for-performance improve the quality of health care? Ann Intern Med 145(4):265–272

Propper C, Burgess S, Gossage D (2008) Competition and quality: evidence from the NHS internal market 1991–9. Econ Jl 118(525):138–170

Reilly S, Olier I, Planner C, Doran T, Reeves D, Ashcroft DM, Gask L, Kontopantelis E (2015) Inequalities in physical comorbidity: a longitudinal comparative cohort study of people with severe mental illness in the UK. BMJ Open. https://doi.org/10.1136/bmjopen-2015-009010

Rosenthal MB, Frank RG, Li Z, Epstein AM (2005) Early experience with pay-for-performance: from concept to practice. JAMA 294(14):1788–1793

Shih T, Nicholas LH, Thumma JR, Birkmeyer JD, Dimick JB (2014) Does pay-for-performance improve surgical outcomes? An evaluation of phase 2 of the premier hospital quality incentive demonstration. Ann Surg 259(4):677

Snijders TA (2011) Multilevel analysis. Springer, Berlin

Sutton M, Nikolova S, Boaden R, Lester H, McDonald R, Roland M (2012) Reduced mortality with hospital pay for performance in England. N Eng J Med 367(19):1821–1828

Van Herck P, De Smedt D, Annemans L, Remmen R, Rosenthal MB, Sermeus W (2010) Systematic review: effects, design choices, and context of pay-for-performance in health care. BMC Health Serv Res 10(1):247

Wang J, Xie H, Fisher JF (2011) Multilevel models: applications using SAS®. Walter de Gruyter, Berlin

Werner RM, Kolstad JT, Stuart EA, Polsky D (2011) The effect of pay-for-performance in hospitals: lessons for quality improvement. Health Aff 30(4):690–698