The impact of performance incentives on child health outcomes: results from a cluster randomized controlled trial in the Philippines

John W Peabody,[1,2]* Riti Shimkhada,[1,2] Stella Quimbo,[3] Orville Solon,[3] Xylee Javier[3] and Charles McCulloch[4]

[1]Institute for Global Health, Global Health Sciences, University of California, San Francisco, CA, USA, [2]QURE Healthcare, San Rafael, CA 94901, USA, [3]School of Economics, University of the Philippines, Diliman and [4]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, USA

*Corresponding author. QURE Healthcare, 1000 Fourth St. Suite 300, San Rafael, CA 94901, USA. Tel: 415 269 5421. Fax: 415 459 4932. E-mail: peabody@psg.ucsf.edu

Improving clinical performance using measurement and payment incentives, including pay for performance (or P4P), has, so far, shown modest to no benefit on patient outcomes. Our objective was to assess the impact of a P4P programme on paediatric health outcomes in the Philippines. We used data from the Quality Improvement Demonstration Study. In this study, the P4P intervention, introduced in 2004, was randomly assigned to 10 community district hospitals, which were matched to 10 control sites. At all sites, physician quality was measured using Clinical Performance Vignettes (CPVs) among randomly selected physicians every 6 months over a 36-month period. In the hospitals randomized to the P4P intervention, physicians received bonus payments if they met qualifying scores on the CPV. We measured health outcomes 4–10 weeks after hospital discharge among children 5 years of age and under who had been hospitalized for diarrhoea and pneumonia (the two most common illnesses affecting this age cohort) and had been under the care of physicians participating in the study. Health outcomes data collection was done at baseline/pre-intervention and 2 years post-intervention on the following post-discharge outcomes: (1) age-adjusted wasting, (2) C-reactive protein in blood, (3) haemoglobin level and (4) parental assessment of child's health using general self-reported health (GSRH) measure. To evaluate changes in health outcomes in the control vs intervention sites over time (baseline vs post-intervention), we used a difference-in-difference logistic regression analysis, controlling for potential confounders. We found an improvement of 7 and 9 percentage points in GSRH and wasting over time (post-intervention vs baseline) in the intervention sites relative to the control sites ($P \leq 0.001$). The results from this randomized social experiment indicate that the introduction of a performance-based incentive programme, which included measurement and feedback, led to improvements in two important child health outcomes.

Keywords   Pay for performance, quality of care, Philippines, child health, health policy

KEY MESSAGES

- This randomized controlled trial of a pay for performance programme in a low- to middle-income country showed that performance incentives improve quality of care.

- Performance incentives can go beyond improvements in quality of physician practice and lead to actual improvements in child health outcomes.

Introduction

In the last decade, there has been widespread interest in improving clinical performance (Institute of Medicine 2002). Measurement and pay for performance (P4P) are two means used to incentivize clinical practice behaviour. Though programmes vary, these approaches generally try to link provider compensation to demonstrable improvements in measures of clinical practice quality (Basinga *et al.* 2011; Epstein 2007). Improving clinical performance and the systems of care in health requires (1) knowing which clinical practices lead to better health, (2) having the ability to put improvements into a clinical practice framework and (3) the ability to isolate and measure changes in practice that have led to better health (Derose and Petitti 2003). Thus, one of the most important technical obstacles facing clinical performance improvement initiatives is the ability to measure performance of the providers after being incentivized and determine if there is a causal link to actual improvements in health status (Werner and Asch 2007).

To date, research shows that P4P might be linked to, at best, modest improvements in quality of care (Witter *et al.* 2012; Lindenauer *et al.* 2007; Petersen *et al.* 2006; Rosenthal *et al.* 2005; Epstein *et al.* 2004). Recent work further suggests that while performance incentives led to improvements in quality of chronic conditions in the short term, the improvements levelled off once performance targets were achieved (Dixon and Khachatryan 2010). The body of literature on P4P is challenged, however, particularly by its paucity in low- and middle-income countries (Witter *et al.* 2012). While most studies are not experimentally designed and participation in P4P programmes is voluntary, there is interest in identifying the impacts of performance incentives in a rigorous randomized study setting (Basinga *et al.* 2011; Garner *et al.* 2012). From a policy perspective, there is yet another issue to consider: findings have primarily been reported from hospitals or hospital-systems and have not looked at attempts to improve health in the wider context of a local population.

In the Philippine Child Health Experiment, known in-country as the Quality Improvement Demonstration Study (QIDS), we had the opportunity to examine the effectiveness of a P4P intervention by introducing a P4P programme in a randomized control manner to 30 communities with district hospitals, and following child health outcomes over time as children were discharged from the hospital and returned home. The policy targeted physicians working in the district community hospitals; quality measurements were collected using Clinical Performance Vignettes (CPVs) for all physicians in the control and intervention arms but the intervention physicians received bonus payments based on their quality scores. We measured the health status 4–10 weeks post-hospital discharge of children 5 years of age and under who had been admitted to the hospitals for diarrhoea or pneumonia and under the care of participating physicians. We used anthropometric, blood and self-reports to assess four different health outcomes of interest.

Methods

QIDS study design

The QIDS was funded by the US National Institutes of Health, as the Philippine Child Health Experiment (NIH: R01 HD042117, ClinicalTrials.gov Identifier: NCT00678197). The study was a large policy experiment that followed the impact of two interventions on physician practices, health behaviours, and health status of children 5 years and under in the Philippines. Further details regarding the study, including the participant flow into the sample, are provided in an earlier publication (Shimkhada *et al.* 2008). QIDS took place at 30 community district-level hospitals financed by the Philippine Health Insurance Corporation (PhilHealth).

Thirty district hospitals, participating on a voluntary basis in the study, are located in 11 provinces in the central region of the Philippines, called the Visayas. These hospitals serve largely poor populations from urban and rural areas. The 30 hospitals were organized into groups of three on the basis of similar demand and supply characteristics, such as average household income, number of beds, average case load, and PhilHealth accreditation status. After the groups were made, randomization was done within each group such that one district was randomized to a control group, one was randomized to a P4P scheme rewarding high quality care with financial incentives to providers at the district/community hospitals, and the other was randomized to another intervention, expanded insurance coverage to households with children under 5 years of age (not examined here, but discussed in previous publications: Solon *et al.* 2009; Quimbo *et al.* 2011).

In the hospitals, physicians were eligible to enrol in the study if they were in good standing with licensure and review boards, accredited by the national insurance corporation, worked at an accredited facility, and cared for patients 5 years of age and under. All physicians participated voluntarily and were consented to participate. The eligible physicians were rostered at each hospital (an average of 11 physicians per hospital at baseline) and three were randomly selected to complete CPVs to measure clinical practice.

The CPV is a validated quality measure of a physician's ability to evaluate, diagnose and treat specific diseases and conditions (Dresselhaus *et al.* 2000). A CPV is a typical case given to a group of doctors simultaneously. The CPVs are open ended and comprehensively assess a provider's clinical practice in five domains: (1) taking a medical history, (2) performing a physical exam, (3) ordering tests, (4) making a diagnosis, and (5) prescribing a

treatment plan. We used CPVs to measure the details of the clinical care activities for diarrhoea and pneumonia (Peabody *et al.* 2000; 2004). Every 6 months, when the CPVs were re-administered, three physicians were randomly sampled so that for each round of data collection, physicians had a recurring probability of being sampled. The randomly selected physicians at each hospital took a total of three vignettes, one for three conditions, pneumonia, diarrhoea and dermatitis (we examine the pneumonia and diarrhoea CPV scores in this paper). Care was exercised so that the physicians did not take the same vignette more than once. Two trained physician abstractors blinded to the CPV-taker's identity scored each vignette. The vignettes were accompanied by a physician survey, which collected data on demographics, education and training, practice characteristics, clinic characteristics and income. CPVs and physician surveys were administered at both the intervention and the control sites. Physicians received individual feedback on their CPV scores. The scores of all the doctors at each hospital were also aggregated and given to the hospital director and the provincial public health officials, including the governors to introduce an element of transparency.

The P4P intervention

In the group of hospitals randomized into the intervention P4P scheme, doctors who met pre-determined quality standards were eligible for bonus payments. In these sites, physicians were told that they had been randomly assigned to the P4P scheme and that they could earn a bonus based on their CPV quality score. A passing cut-off score on the CPV determined eligibility for the bonus (see Solon *et al.* 2009). The bonus amount for each qualifying doctor was calculated by multiplying the number of inpatients during the quarter by a bonus of 100 Philippine pesos (PhP) [2.4 U.S. dollars (USD)] per patient. Bonuses were paid quarterly to qualifying intervention sites. The bonus payments represented 5% of total physician salaries.

Data collection

In the Philippines, diarrhoea and pneumonia are responsible for ~25% of all deaths in children 5 years and under (World Health Organization 2006), posing challenges to the health system as it struggles to implement preventive and proven clinical measures (Black *et al.* 2003; Lewin *et al.* 2008). Diarrhoea and pneumonia care were chosen as target conditions because the overarching goal of the QIDS study was to improve the health status of hospitalized children 5 years old and under and focus on diseases causing the greatest burden of child morbidity and mortality in developing countries.

To measure health status after the hospital stay, we administered surveys to the parents of all children 5 years old and under who were admitted to the hospital for diarrhoea or pneumonia, 4–10 weeks after they were discharged from the study hospitals in each of the two survey rounds. Data collection was done by trained interviewers who went to both intervention and control sites. The interviewers were independent teams who were not part of the analytic team or from the study hospital staff.

After discharge, all children with the diagnosis of pneumonia or diarrhoea (roughly one-half to two-thirds of all the discharges) were eligible. We asked parents of those children to voluntarily provide information on the child's symptoms, their health seeking behaviour and utilization for the reference illness, services provided, drugs prescribed and purchased, payments and financing arrangements (including PhilHealth membership status). We selected four health status outcome measures that would potentially be affected by better quality care for diarrhoea and pneumonia.

We directly measured weight and height measurements (age adjusted) to assess for wasting (as defined by the Waterlow system for wasting in the 2006 World Health Organization Standard Reference) (World Health Organization 2006). We drew blood samples from the children to test for the qualitative presence of C-reactive protein (CRP), a possible measure of acute ongoing infection, and checked for anaemia, as measured by blood haemoglobin levels (anaemia defined as <10 g/dl blood haemoglobin).

Last, we asked parents to rate their child's health status as excellent, very good, good, fair or poor to obtain a parental assessment of improvement in health using the general self-reported health (GSRH) measure, which is commonly dichotomized into greater than or equal to good vs less than good (DeSalvo *et al.* 2006).

Baseline data were collected in 2003 (Round 1), the P4P policy intervention was introduced in 2004, and then a post-intervention assessment was done in 2006–07 (Round 2). For the child assessment, there was an average of 15–20 children from each facility: at baseline, there was a sample of 501 children in the control arm and 496 in the intervention arm; at Round 2, there were 560 and 596 children, respectively. Of those who participated in the baseline, the refusal rate at Round 2 was 1.5% and there was another 1% lost to follow-up (i.e. could not be located). To collect data from the physicians, we created a roster of all of the physicians caring for paediatric patients in the QIDS facility. All physicians who were deemed eligible, defined as caring for children at least 20% of the time, were asked to participate and all agreed to be in the study (i.e. no refusals). For each study site, an average of three physicians was randomly drawn from the hospital roster; a total of 119 out of 119 eligible physicians participated in the study (61 at baseline and 58 in Round 2).

Model

We used logistic regression models to estimate the effects of the P4P intervention on the four health outcomes. We employed a basic 'difference-in-difference' approach to compare the change in health outcomes (before and after) between the intervention sites and the control sites. By looking at the change over time, we adjusted for (1) characteristics that do not change over time within control and treatment localities and (2) characteristics that change over time but are common to both control and treatment areas.

The difference-in-difference model can be specified in regression form as

$$P\{Y_{ijt} = 1\} \equiv p_{ijkt} = \text{probability of an outcome at facility } i,$$
$$\text{individual } j, \text{ intervention } k, \text{ at time } t.$$
$$\text{logit}(p_{ijkt}) = \mu + a_i + \beta_k + \gamma_t + (\beta\gamma)_{kt} + \sum_p \varphi x_{ijktp}.$$

In each facility (i) there are individuals (j) that are assigned to a treatment group (k), and we look at effects over time (t). Thus, in our model the right-hand-side variables include a

random effect for each facility, α_i, an effect for the intervention, β_k, each time period (γ), and an interaction effect $(\beta\gamma)$ indicating whether the individual is in an intervention locality and an indicator of whether it is the post-intervention period. There is also a series of control variables, indexed by p, $(xijktp)$ that may vary over time and across individuals. These include (1) PhilHealth (insurance) membership, (2) age of child (months), (3) mother's education (years of schooling), and (4) household income (PhP), (5) initially visited a lower-level facility prior to hospitalization and (6) length of stay in hospital. The individual effects control for individual, household and area specific factors that are fixed over time. Time controls for factors that vary over time but are common across all areas—both treatment and control. The coefficient $\beta\gamma$ is the difference-in-difference estimate of the impact of P4P on health. Our threshold for statistical significance was set at a P value < 0.05 and we used the Bonferroni correction to adjust this P value for our four outcomes. And, to account for facility level clustering, we adjusted standard errors using the xtlogit, random effects command in STATA 10.

Results

At baseline assessment we found that the intervention and control groups were equally divided between cases of diarrhoea and pneumonia, both groups had mothers with nearly 9 years of schooling, hospital stays of ~3 days, and ~30% were insured. The baseline health status of children discharged from the hospital revealed that ~40% were wasted, 22% had elevated CRP, 16% were anaemic, and ~20% of parents reported their children as having fair or poor health status (see Table 1).

To determine the effects of the P4P intervention on health status we modelled whether health status was different between intervention and control groups 4–10 weeks after they were discharged, controlling for potential confounding socioeconomic and clinical variables (see Table 2). We found that the number of children who were wasted increased by 9 percentage points from baseline for the patients cared in the control group compared with the incentivized doctors who received the quality bonus payments where there was no change ($P < 0.001$). We also found that parents reported an improvement in GSRH of 7 percentage points in the P4P sites compared with controls ($P < 0.001$) (see Table 3 for a summary of difference-in-difference estimates). CRP and haemoglobin did not change and were statistically similar between the two groups.

Discussion

We present herein the key finding from the QIDS, a randomized trial conducted at the community level in the Philippines examining the effects of a measurement and incentive based (P4P) policy programme rewarding physicians

Table 1 Study population characteristics at baseline for control ($n = 501$) and intervention ($n = 496$) sites, mean and standard deviation (in parentheses)

	Control	Intervention	Average
Age (months)	20.22 (12.62)	19.50 (12.16)	19.86 (12.39)
Household income (annual, PhP)	52 367 (81 585)	58 266 (64 175)	55 305 (73 455)
Mother's education (years of schooling)	8.54 (3.33)	8.89 (3.33)	8.72 (3.33)
Length of stay (days)	3.22 (1.49)	3.32 (1.46)	3.27 (1.47)
Visited a lower-level facility (%)	27.54 (44.72)	35.48 (47.89)	31.49 (46.47)
Insured—PHIC member (%)	29.14 (45.49)	31.25 (46.40)	30.19 (45.93)
CRP (% positive)	6.0 (23.7)	4.0 (19.7)	5.0 (21.8)
Anaemic [Hgb < 10 µg/ml (%)]	14.7 (35.4)	9.5 (29.7)	12.1 (32.6)
GSRH (% rating ≥ good)	83.4 (37.2)	75.0 (43.3)	79.2 (40.6)
Wasted (%)	26.4 (26.4)	31.2 (46.4)	28.9 (45.3)

Table 2 Marginal effects of the bonus intervention in Round 2 from logistic regression models for four health outcomes of interest

	CRP negative		Not wasted		GSRH ≥ good		Not anaemic	
	Coeff	P-value	Coeff	P-value	Coeff	P-value	Coeff	P-value
Bonus intervention in Round 2	0.008	0.497	0.093	0.000	0.074	0.001	0.049	0.253
Bonus sites (vs control sites)	0.016	0.312	−0.049	0.219	−0.083	0.222	−0.042	0.226
Round 2 (vs baseline)	−0.005	0.664	−0.098	0.001	−0.008	0.681	−0.030	0.095
PHIC member	−0.000	0.977	−0.021	0.321	0.023	0.130	0.010	0.558
Length of stay (No. days in hospital)	0.001	0.678	−0.011	0.074	−0.011	0.035	0.006	0.228
Initially visited another lower-level facility	−0.014	0.205	0.007	0.704	−0.016	0.325	−0.010	0.502
Age of child (months)	−0.0001	0.308	−0.002	0.026	0.000	0.414	−0.004	0.001
Mother's education (years of schooling)	0.003	0.056	0.014	0.000	0.010	0.003	−0.008	0.011
Household income (PhP)	0.0001	0.897	0.0001	0.009	1.39E-07	0.228	0.0001	0.024

Table 3 Summary of difference-in-difference estimates of health outcomes comparing intervention and control

Health outcome	Pre-QIDS bonus intervention prevalence (%)	Post-QIDS bonus intervention prevalence (%)	Improvement (Post–Pre), %	P-value
CRP negative				
Intervention	97.69	98.07	0.38	
Control	96.06	95.60	−0.46	
Difference	1.63	2.47	0.84	0.497
Not anaemic				
Intervention	93.80	91.95	−1.85	
Control	89.59	92.61	3.02	
Difference	4.21	−0.66	−4.97	0.253
Not wasted				
Intervention	70.09	69.57	−0.51	
Control	75.02	65.25	−9.77	
Difference	−4.93	4.32	9.25	<0.0001
GSRH at least good				
Intervention	78.50	85.02	6.53	
Control	86.79	85.94	−0.85	
Difference	−8.29	−0.92	7.37	0.001

financially when they provided higher quality care to children. We compared health outcomes in intervention vs control groups to evaluate the health of children 4–10 weeks after a hospital stay for diarrhoea or pneumonia. While over time (post-intervention vs baseline) we found an overall increase in children who were wasted (underweight for height), the increase was significantly greater in the control group than in the intervention group such that the difference-in-difference estimate suggests a statistically significant improvement of 9 percentage points in being not wasted in the intervention group vs the control group over time. We also found a statistically significant increase of 7 percentage points of parents reporting at least good health for their children in the intervention sites compared with controls over time. Other reports show that subjective reports are good predictors of future health status and future costs of care (Peabody *et al.* 2000; DeSalvo *et al.* 2009; Idler and Benyamini 1997). We note that the GSRH and wasting were more responsive to changes in quality than our two blood drawn measures. Other studies reveal that even small changes in GSRH are associated with increased future health expenditures and that childhood wasting is associated with chronic disease later in life, less educational performance, lower labour productivity and income attainment (Victora *et al.* 2008). Also, under-nutrition and the more severe cases of wasting have been attributed to 61% of diarrhoea deaths and 52% of pneumonia deaths (Caulfield *et al.* 2004).

Our findings underscore two clear reasons to conduct carefully designed, randomized controlled evaluations. First, the decrease in wasting was relative and only seen against a backdrop where the controls actually worsened over time. From contextual evidence noted at the time of the study, the increase in the prevalence of wasting was due to severe weather disturbances (hurricanes) in 2006 that affected crops (food

supply), shelter, and infrastructure and led to outbreaks of water borne diseases. Without the control, the policy would have been assumed to be ineffective. Second, there exists a dearth of P4P studies that have been conducted in a randomized controlled manner, especially on paediatric outcomes (Chien 2012). In most studies on incentivizing doctors, the interventions are not randomly assigned and this introduces the possibility of selection bias wherein providers who elect to adopt the incentives may well be the ones most likely to respond and improve their clinical practice (Petersen *et al.* 2006; Fairbrother *et al.* 1999; 2001; Kouides *et al.* 1998). Three US hospital-based studies, examining inpatient-based P4P programmes as we did in this study, have included control hospitals but not random assignment (Lindenauer *et al.* 2007; Glickman *et al.* 2007; Grossbart 2006). These studies found a more modest benefit of 2- to 4-percentage point improvement in outcomes beyond the improvement seen in controls (Lindenauer *et al.* 2007; Grossbart 2006). Just recently, one of the first studies on P4P done in a randomized controlled setting found that P4P improved adolescent substance-use disorder treatment implementation by clinicians (Garner *et al.* 2012). Providers in the P4P group were significantly more likely to demonstrate clinical competence than controls (24.0% vs 8.9% respectively, $P = 0.02$), and patients in the P4P group were significantly more likely to receive target treatment than controls (17.3% vs 2.5% respectively, $P = 0.01$). Recent discussion of P4P includes the recognition that evidence of effectiveness from randomized controlled studies on P4P must be one of the foremost criteria prior to implementing such a programme (Glasziou *et al.* 2012).

Our study is also one of the first studies to examine effects of physician supply-side incentives in paediatric health outcomes in a developing county. Recently, Basinga *et al.* (2011) reported on findings from Rwanda where a P4P programme for maternal and child health-care providers led to improvements in the number of institutional deliveries and increases in the number of preventive care visits by children. Interest is clearly growing in identifying real world examples coupled with rigorous evaluation on policies and practices that lead to better health. Performance-based incentives are thought to be one of the best ways to improve health, particularly in the developing world, when physicians are not adequately incentivized to provide quality care (Eichler and Levine 2009). Other recent large studies have looked primarily at demand side interventions, such as Oportunidades in Mexico, which was effective in reducing the prevalence of stunting in preschool children (Gertler 2004; Rivera *et al.* 2004).

The issues of how to measure performance in P4P schemes and what health outcome measures to use in community-based studies are the subjects of important debate. Measuring health outcomes as a determinant of better quality is expensive, expansive and requires a large sample to capture infrequent health events. We found that CPVs might be a valuable alternative to as they provide a detailed measure of the clinical encounter that capture, on a case-mix adjusted basis, the evidence-based elements of care quality that lead to better health outcomes (Peabody and Liu 2007). Furthermore, in a prior publication we showed that CPV measurement was sustainable, minimally disruptive and readily grafted into existing routines (Solon *et al.* 2009). CPVs are also better

designed to overcome gaming, a vexing problem for P4P programmes, when physicians are able to make small changes in their practices or differentially select patients based on the likelihood of good clinical outcomes (Petersen *et al.* 2006). Perhaps as important, from an operational point of view, the CPV was more responsive to the immediate impacts of policy change than standard structural measures and at a cost of only US$305 per site assessment it was considered affordable to the local administrating body.

We also found the subjective parental report of health outcomes (the GSRH in this study) to be a responsive measure of overall heath status. Other reports show that subjective reports are good predictors of future health status and future costs of care (Peabody *et al.* 2000; DeSalvo *et al.* 2009; Idler and Benyamini 1997). We note that the GSRH and wasting were more responsive to changes in quality than our two blood drawn measures.

One challenge is parsing out exactly how much of the improvements in outcomes can be attributed to the bonus payment or from just the measurement, and feedback, of quality scores. In a recent publication, we report on findings that suggest indeed measurement and feedback itself may be important in motivating clinical practice change (Peabody *et al*. 2013). At 36 months after the intervention, the intervention sites receiving the performance-based bonus had CPV scores ~10 percentage points higher than baseline. Quality scores also improved in control sites but by a smaller degree (5 percentage points), suggesting that measurement and feedback alone without the performance bonus contributed significantly to the quality effect. We hypothesize that measurement and feedback of quality scores likely play a primary role in driving improvements in outcomes, with the bonus emolument acting as an accelerator.

Another limitation is that we had only a selected set of outcomes measures; clearly, there are others that could have been used, and not all of our measures were impacted by the intervention. While we accounted for multiple outcomes by using the Bonferroni correction, it is also possible that if we used more or different measures, we would have found other outcomes that were impacted by higher quality. Other limitations include a lack of generalizability of our findings to areas and hospitals outside of the Philippines. We also believe with longer-term follow-up of children we could better elucidate health outcome changes in children who were under the care of study physicians as well as trickle-down effects to other members of the household and community. Finally, operationalizing the P4P scheme requires attention to administrative detail that began with a high degree of co-operation with our host partner, the Philippine National Heath Insurance Corporation and other government leaders on down to the minutiae of processing claims correctly, providing feedback to the doctors and ensuring the distribution of the cheques. While critical to success, it may, in an output-based reimbursement system, re-focus efforts on performance, thereby obviating the need for detailed accounting and input monitoring, such as medical claims review. Instead, output-based reimbursement allows administrators to focus efforts on improving quality.

In summary, the findings in this paper add to the notion that P4P policy impacts can go beyond improvements in the measured quality of physician practice and lead to actual improvements in health outcomes. Based on the model results presented here, we estimate that the impacts of higher quality are potentially quite large: P4P resulted in 294 averted cases of wasting and 229 more children reporting at least good health. Were this to become national policy, it would translate into 15 000–20 000 fewer annual cases of wasting in children 5 years and under. Finally, our work further suggests that a randomized controlled study on P4P can be done in a relatively short time frame (performance measurement and outcome assessment done over a 2-year period), but only when evaluation is done in concert with original policy-making and anticipated prior to implementation. This adds to the notion that policy effectiveness is possible and that ineffective polices can be unmasked early on (Shimkhada *et al.* 2008).

Ethics committee approval

Study conducted in accordance with the ethical standards of the applicable national and institutional review boards (IRBs) of the University of the Philippines and the University of California, San Francisco (CHR Approval Number: H10609-19947-05).

Funding

Conflict of interest

There are no actual or potential Conflicts of interest by any of the authors. This work was conducted through the University of California, San Francisco under an NIH grant and with a subcontract to the University of the Philippines. After the conclusion of this study and writing of this report, authors J.P. and R.S. have joined QURE Healthcare. QURE Healthcare is a privately owned company (President, J.P.), which uses the performance measure described in this report, the CPV vignettes, in different clinical settings, and in a proprietary manner.

References

Basinga P, Gertler PJ, Binagwaho A, Soucat AL, Sturdy J, Vermeersch CM. 2011. Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation. *Lancet* **377**: 1421–8.

Black RE, Morris SS, Bryce J. 2003. Where and why are 10 million children dying every year? *Lancet* **361**: 2226–34.

Caulfield LE, de Onis M, Blössner M, Black RE. 2004. Undernutrition as an underlying cause of child deaths associated with diarrhea, pneumonia, malaria, and measles. *The American Journal of Clinical Nutrition* **80**: 193–8.

Chien AT. 2012. Can pay for performance improve the quality of adolescent substance abuse treatment? *Archives of Pediatrics and Adolescent Medicine* **166**: 964–5.

Derose SF, Petitti DB. 2003. Measuring quality of care and performance from a population health care perspective. *Annual Review of Public Health* **24**: 363–84.

DeSalvo KB, Bloser N, Reynolds K, He J, Muntner P. 2006. Mortality prediction with a single general self-rated health question. A meta-analysis. *Journal of General Internal Medicine* **21**: 267–75.

DeSalvo KB, Jones T, Peabody JW et al. 2009. Health care expenditure prediction with a single item, self-rated health measure. *Medical Care* **47**: 440–7.

Dixon A, Khachatryan A. 2010. A review of the public health impact of the Quality and Outcomes Framework. *Quality in Primary Care* **18**: 133–8.

Dresselhaus TR, Peabody JW, Lee M, Wang MM, Luck J. 2000. Measuring compliance with preventive care guidelines: standardized patients clinical vignettes and the medical record. *Journal of General Internal Medicine* **15**: 782–8.

Eichler R, Levine R. 2009. *Performance-Based Incentives Working Group. Performance Incentives for Global Health: Potential and Pitfalls*. Washington, DC: Center for Global Development.

Epstein AM. 2007. Pay for performance at the tipping point. *The New England Journal of Medicine* **356**: 515–7.

Epstein AM, Lee TH, Hamel MB. 2004. Paying physicians for high-quality care. *The New England Journal of Medicine* **350**: 406–10.

Fairbrother G, Hanson KL, Friedman S, Butts GC. 1999. The impact of physician bonuses, enhanced fees, and feedback on childhood immunization coverage rates. *American Journal of Public Health* **89**: 171–5.

Fairbrother G, Siegel MJ, Friedman S, Kory PD, Butts GC. 2001. Impact of financial incentives on documented immunization rates in the inner city: results of a randomized controlled trial. *Ambulatory Pediatrics* **1**: 206–12.

Garner BR, Godley SH, Dennis ML, Hunter BD, Bair CM, Godley MD. 2012. Using pay for performance to improve treatment implementation for adolescent substance use disorders: results from a cluster randomized trial. *Archives of Pediatrics and Adolescent Medicine* **166**: 938–44.

Gertler PJ. 2004. Do conditional cash transfers improve child health? Evidence from PROGRESA's control randomized experiment. *The American Economic Review* **94**: 336–41.

Glasziou PP, Buchan H, Del Mar C et al. 2012. When financial incentives do more good than harm: a checklist. *BMJ* **345**: e5047.

Glickman SW, Ou FS, DeLong ER. 2007. Pay for performance, quality of care, and outcomes in acute myocardial infarction. *JAMA* **297**: 2373–80.

Grossbart SR. 2006. What's the return? Assessing the effect of 'pay-for-performance' initiatives on the quality of care delivery. *Medical Care Research and Review* **63**: 29S–48S.

Idler EL, Benyamini Y. 1997. Self-rated health and mortality: a review of twenty-seven community studies. *Journal of Health and Social Behavior* **38**: 21–37.

Institute of Medicine. 2002. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: National Academy Press.

Kouides RW, Bennett NM, Lewis B, Cappuccio JD, Barker WH, LaForce FM. 1998. Performance-based physician reimbursement and influenza immunization rates in the elderly. *American Journal of Preventive Medicine* **14**: 89–95.

Lewin S, Lavis JN, Oxman AD, Bastías G, Chopra M, Ciapponi A. 2008. Supporting the delivery of cost-effective interventions in primary health-care systems in low-income and middle-income countries: an overview of systematic reviews. *Lancet* **372**: 928–39.

Lindenauer PK, Remus D, Roman S et al. 2007. Public reporting and pay for performance in hospital quality improvement. *The New England Journal of Medicine* **356**: 486–96.

Peabody JW, Liu A. 2007. A cross-national comparison of the quality of clinical care using vignettes. *Health Policy and Planning* **22**: 294–302.

Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. 2000. Comparison of vignettes standardized patients and chart abstraction: a prospective validation study of 3 methods for measuring quality. *JAMA* **283**: 1715–22.

Peabody JW, Luck J, Glassman P et al. 2004. Measuring the quality of physician practice by using clinical vignettes: a prospective validation study. *Annals of Internal Medicine* **141**: 771–80.

Peabody JW, Shimkhada R, Quimbo S et al. 2011. Financial incentives and measurement improved physicians' quality of care in the Philippines. *Health Aff (Millwood)* **30**: 773–81.

Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S. 2006. Does pay-for-performance improve the quality of health care? *Annals of Internal Medicine* **145**: 265–72.

Quimbo SA, Peabody JW, Shimkhada R, Florentino J, Solon O. 2011. Evidence of a causal link between health outcomes, insurance coverage, and a policy to expand access: experimental data from children in the Philippines. *Health Economics* **20**: 620–30.

Rivera JA, Sotres-Alvarez D, Habicht JP, Shamah T, Villalpando S. 2004. Impact of the Mexican program for education, health, and nutrition (Progresa) on rates of growth and anemia in infants and young children: a randomized effectiveness study. *JAMA* **291**: 2563–70.

Rosenthal MB, Frank RG, Li Z, Epstein AM. 2005. Early experience with pay-for-performance: from concept to practice. *JAMA* **294**: 1788–93.

Shimkhada R, Peabody JW, Quimbo SA, Solon O. 2008. The Quality Improvement Demonstration Study: an example of evidence-based policy-making in practice. *Health Research Policy and Systems* **25**: 5.

Solon O, Woo K, Quimbo SA, Peabody JW, Florentino J, Shimkhada R. 2009. A novel method for measuring health care system performance: experience from QIDS in the Philippines. *Health Policy and Planning* **24**: 167–74.

Victora CG, Adair L, Fall C et al. 2008. Maternal and child under-nutrition: consequences for adult health and human capital. *Lancet* **371**: 340–57.

Werner RM, Asch D. 2007. Clinical concerns about clinical performance measurement. *Annals of Family Medicine* **5**: 159–63.

Witter S, Fretheim A, Kessy FL, Lindahl AK. 2012. Paying for performance to improve the delivery of health interventions in low- and middle-income countries. *Cochrane Database of Systematic Reviews* **2012**: CD007899.

World Health Organization. 2006. *WHO Global Database on Child Growth and Malnutrition*. http://www.who.int/nutgrowthdb/en/, accessed 5 November 2012.