RAND Journal of Economics
Vol. 41, No. 1, Spring 2010
pp. 64–91

Can you get what you pay for? Pay-for-performance and the quality of healthcare providers

Kathleen J. Mullen*

Richard G. Frank**

and

Meredith B. Rosenthal***

Despite the popularity of pay-for-performance (P4P) among health policy makers and private insurers as a tool for improving quality of care, there is little empirical basis for its effectiveness. We use data from published performance reports of physician medical groups contracting with a large network HMO to compare clinical quality before and after the implementation of P4P, relative to a control group. We consider the effect of P4P on both rewarded and unrewarded dimensions of quality. In the end, we fail to find evidence that a large P4P initiative either resulted in major improvement in quality or notable disruption in care.

1. Introduction

■ In 1999, the Institute of Medicine (IOM) issued a startling report estimating that, every year, between 44,000 and 98,000 people admitted to U.S. hospitals die as a result of preventable medical errors (IOM, 1999). On average, U.S. patients receive only 55% of recommended care, including regular screenings, follow-ups, and appropriate management of chronic diseases such as asthma and diabetes (McGlynn et al., 2003). In response to widespread concerns over high rates of medical errors and inconsistent healthcare quality that have persisted in the face of public reporting of quality, health policy makers and private insurers are turning to pay-for-performance (P4P) as a more direct line of attack. More recently, the IOM cited over 100 P4P programs in

*RAND Corporation; kmullen@rand.org.
**Harvard Medical School and NBER; frank@hcp.med.harvard.edu.
***Harvard School of Public Health; mrosenth@hsph.harvard.edu.

We are grateful to Zhonghe Li for assistance with the PacifiCare data. We also thank Cheryl Damberg, Steve Levitt, Dayanand Manoli, Tom McGuire, and Eric Sun, seminar participants at Harvard University, the Public Policy Institute of California, RAND, and Yale University, and participants of the June 2006 Robert Wood Johnson Scholars in Health Policy Research Annual Meeting (Aspen, Colorado) for helpful comments and advice. We are particularly grateful to Philip Haile and two anonymous referees for extremely helpful suggestions. Financial support from the Commonwealth Fund, the Harvard Interfaculty Initiative for Health Systems Improvement, and the Robert Wood Johnson Foundation is gratefully acknowledged.

place in private healthcare markets, and recommended that Medicare incorporate P4P into its reimbursement structure (IOM, 2006). As Mark McClellan, former administrator of the Centers for Medicare and Medicaid Services (CMS), put it, "You get what you pay for. And we ought to be paying for better quality" (quoted in the *New York Times*, Leonhardt, 2006).

In contrast to public reporting campaigns, which rely on consumer response to information, P4P programs focus their efforts on the price margin directly to motivate quality improvement. A typical P4P program rewards healthcare providers (e.g., physician medical groups) with bonuses for high marks on one or more quality measures, such as rates of preventative screenings or adherence to guidelines for chronic disease management (e.g., regular blood sugar testing for diabetics). These measures are based on clinical studies showing that better outcomes result when these processes are followed for patients meeting certain criteria. The rationale for pay-for-performance is simple. If quality of care becomes a direct component of their financial success, providers will shift more resources toward quality improvement. Economic theory, however, suggests the story may not be this simple. In particular, providers may shift resources toward rewarded dimensions of quality at the expense of unrewarded dimensions, which may result in a decline in the overall quality of patient care.

In this article, we use data from the performance reports of medical groups contracting on a capitated basis with a large network HMO, PacifiCare Health Systems, before and after implementation of two P4P programs in California. We compare the performance of these groups to medical groups in the Pacific Northwest that were not affected by either program. In early 2002, PacifiCare announced the creation of a new quality incentive program (QIP), which paid quarterly bonuses to medical groups performing at or above the 75th percentile from the preceding year on one or more of five clinical quality measures. On average, PacifiCare accounts for 15% of total capitated revenues among medical groups in our sample. One year after the QIP went into effect, PacifiCare joined forces with five other health plans in a coordinated P4P program sponsored by California's Integrated Healthcare Association (IHA), a nonprofit coalition of health plans, physician groups, hospitals, and purchasers. Together, the plans participating in the IHA program account for 60% of revenues for the medical groups in our data. Five of the six measures selected by the IHA were also targets of the original PacifiCare program.

We address two main questions. First, were either of these P4P programs effective at inducing changes in quality of care? Second, if so, did the programs encourage healthcare providers to divert effort away from unrewarded toward rewarded dimensions of quality? We find that pay-for-performance did have a positive impact on some of the clinical measures rewarded by the programs, and the impact increased with the size of the average expected reward. However, we fail to find evidence that the programs either resulted in major improvement or notable disruption in care.

Our data have several unique features which make it possible for us to investigate these questions. First, although PacifiCare announced its P4P program early in 2002, it has been collecting quality information on its providers since 1993 and making that information public since 1998. This allows us to estimate and control for preperiod trends in quality improvement irrespective of the QIP. We can also attribute any postperiod trend breaks to the QIP without confounding our results with the effects of the public reporting. To control for macro shocks to quality trends, we have data on a control group of PacifiCare providers in the Pacific Northwest where there is also public reporting of quality of care but no P4P scheme. In addition, we have data on performance measures not explicitly rewarded, or differentially rewarded, by the incentive programs, which allows us to investigate spillover effects to other measures along rewarded and unrewarded dimensions of quality.

Despite the rising popularity of P4P, little is known about how providers actually respond to such schemes. Randomized controlled trials of P4P are rare and tend to be small in scale. Additionally, P4P programs are often introduced at the same time as other quality improvement strategies such as public reporting, making it difficult to isolate the effects of P4P. In a review of the empirical evidence on P4P, Rosenthal and Frank (2006) identified only seven published,

peer-reviewed studies of the impact of P4P in health care, with mixed results (zero or small positive effects on rewarded quality measures). These studies focused on outcomes such as flu vaccinations, childhood immunizations, and dispensation of smoking cessation advice, and they tended to be small in terms of both sample size (15–60 medical groups or physicians) and financial impact (with potential bonuses ranging from $500–$5000 annually). In 2004, Britain's National Health Service rolled out a new P4P program for general practitioners. This program was much larger than most P4P programs in the United States, with practices earning average bonuses of $133,200 (Doran et al., 2006). Campbell et al. (2007) estimated that quality indicators for asthma and diabetes (but not coronary heart disease) improved in 2005 after P4P was implemented in the United Kingdom, relative to projected performance based on trends from 1998 to 2003. They found that rewarded and unrewarded measures improved about the same.

We build on an earlier study by Rosenthal et al. (2005) which examined the effects of the PacifiCare intervention on three clinical service measures rewarded by that program: cervical cancer screening, breast cancer screening, and hemoglobin A1c testing for diabetics. Using a difference-in-differences approach, they found that cervical cancer screening was the only measure with a statistically significant response to the program, on the order of 3 percentage points (10%). Our article extends the time period of that study in order to separate the estimated effect of the PacifiCare intervention from that of the larger-scale, coordinated P4P program introduced roughly 6 months into the postperiod. In addition, we examine both measures that were explicitly rewarded by P4P and measures that were differentially rewarded, or not rewarded at all, by either P4P policy.

In addition to contributing to the literature on quality improvement in health care, our article contributes to the growing empirical literature on Holmstrom and Milgrom's (1991) theory of multitasking (see, e.g., Jacob [2005] for an analysis of teachers' responses to test-based accountability, and Lu [2009] for an application of multitasking theory to public reporting in the nursing home industry). We consider two ways in which medical groups can respond to P4P: (i) they can divert resources away from unrewarded measures to focus on the targeted measures; or (ii) they can make more general quality improvements, boosting both rewarded and unrewarded measures of performance. Which response dominates will depend on the technology of quality improvement in medical practices, about which little is known. For example, screening and follow-up measures, such as mammography and hemoglobin A1c (blood sugar) testing for diabetics, may both be increased by a general improvement in information technology, such as a computerized reminder program, despite differences in administration technique and patient populations. The degree of commonality in the production of quality measures is crucial to whether we expect to see positive or negative spillovers.

The remainder of the article is organized as follows. In the next section, we develop a model of provider response to P4P. In Section 3, we introduce our natural experiment and discuss the features of our data. In Section 4, we describe our estimation strategy for evaluating the effect of P4P on the underlying dimensions of clinical quality, presenting the results in Section 5. We offer concluding remarks in Section 6.

2. A model of provider response to pay-for-performance

■ Consider a principal-agent model in which the agent (e.g., physician medical group) chooses how much to invest in quality q, which is unobservable to the principal (payer, e.g., insurance company, which may or may not be acting on behalf of its patients). Quality may have several dimensions, that is, $q = (q_1, \ldots, q_J)$. In our model, we abstract from the issue of quantity of services provided and focus solely on the determination of quality. Let $B(q)$ denote the benefit to the principal when the agent chooses quality level q, where B itself may be unobservable to the principal. Let $C(q)$ denote the cost to the agent of producing quality at level q, where C is weakly increasing and strictly convex. Costs can be fixed (e.g., a one-time investment in information technology, such as an automated reminder program) or variable (e.g., doctor time or effort).

The principal observes a set of signals (quality indicators) $y = (y_1, \ldots, y_K)$ that depend in part on q but do not fully reveal the agent's choice of quality provided,

$$y = \mu(q) + \varepsilon, \tag{1}$$

where $\varepsilon_k \mid q \sim F_k, \quad k = 1, \ldots, K,$ with $E[\varepsilon_k \mid q] = 0$ and $E[\varepsilon_k \varepsilon_{k'} \mid q] = 0$. Let μ_{jk} denote $\partial y_k / \partial q_j$, which reflects the marginal increase in the expected value of measure y_k resulting from an increase in quality dimension q_j. We assume that μ is fixed and taken as given by the provider. In other words, we assume that providers cannot "game" the measures, for example by selecting only patients with favorable attributes. The concern that P4P could encourage "cream skimming" is widespread, and the measures we examine were chosen to minimize opportunities for patient selection.[1] For the most part, the measures we examine are diagnostically narrow process measures; that is, they evaluate actions taken by providers and so they rely little on inputs from patients (who are all commercially insured in our setting). In addition, the measures are audited by the National Committee for Quality Assurance.

In our model, the measures can only increase (in expected value) if one or more of the underlying quality dimensions changes. If two measures y_k and $y_{k'}$ both depend positively on q_j, then we say a commonality exists in the production of measures y_k and $y_{k'}$. An example of this is the automated reminder program, which may increase the number of patients screened for diseases or examined for follow-up care, regardless of specifics regarding patient population or administration technique of a particular test/exam.

Let $R(y)$ denote the compensation of the agent. In the benchmark case, where compensation does not depend on quality, $R(y) = r_0$. Then the agent chooses q to minimize cost:

$$\frac{\partial C}{\partial q_j} = 0, \quad j = 1, \ldots, J. \tag{2}$$

Note that in a capitated environment the provider may save money by providing quality (e.g., screening for some health problems may be cost-effective if the resultant costs of care are high).[2] Unless $C(q) = -B(q)$, the agent sets q lower than the efficient level. This suggests there is room for improvement if R can depend on q, even if indirectly through y.

Now assume that a target-based P4P bonus scheme is instituted, in which the agent is rewarded additionally on y_k only if y_k reaches a predetermined absolute target level T_k, for $k = 1, \ldots, K$:

$$R(y) = r_0 + \sum_{k=1}^{K} r_k \mathbb{I}(y_k \geq T_k).$$

Assume that the agent is risk neutral, and maximizes expected profits

$$\begin{aligned} E[R(y)] - C(q) &= r_0 + \sum_{k=1}^{K} r_k \Pr(y_k \geq T_k) - C(q) \\ &= r_0 + \sum_{k=1}^{K} r_k [F_k(\mu(q) - T_k)] - C(q), \end{aligned}$$

where F_k is the cumulative density function of $\varepsilon_k, \ k = 1, \ldots, K$. The first-order condition is

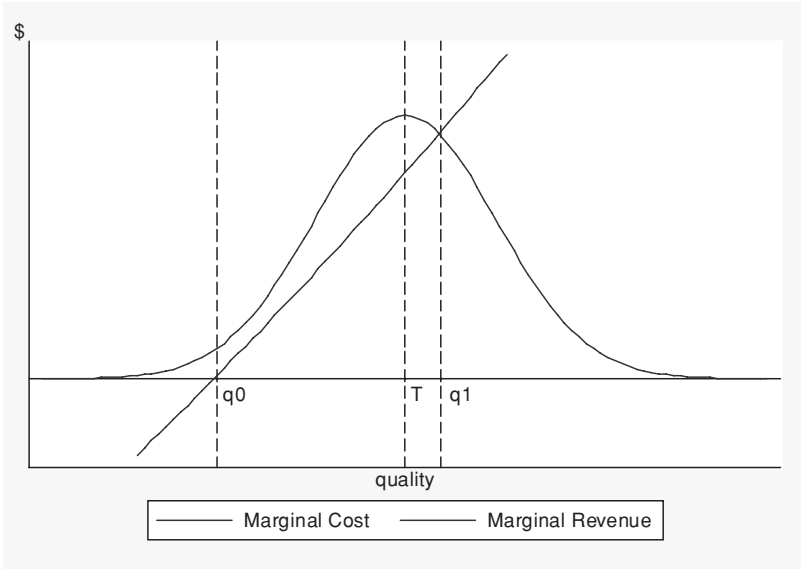
$$\frac{\partial C}{\partial q_j} = \sum_{k=1}^{K} r_k \mu_{jk} f_k(\mu(q) - T_k), \quad j = 1, \ldots, J. \tag{3}$$

[1] Shen (2003) found that performance-based contracting encouraged Maine's Office of Substance Abuse to selectively drop harder-to-treat patients. Similarly, Dranove et al. (2003) found that public reporting of cardiac surgery outcomes encouraged selection against sicker patients.

[2] We can allow for some altruism on the part of providers, that is, providers maximize $R(y) + \alpha B(q) - C(q)$, but this does not change our results qualitatively, as long as providers are imperfect agents, for example, $\alpha < 1$.

FIGURE 1

QUALITY DETERMINATION UNDER P4P



This simply states that medical groups choose q by setting the marginal cost of quality improvement equal to the expected marginal revenue from increasing q. Ignoring cross-partial effects in the cost function, if $r_k\mu_{jk} \geq 0$, for all k, and $r_k\mu_{jk} > 0$ for at least one k, then quality along dimension j will increase as a result of P4P, as the right-hand side of (3) is greater than zero. Figure 1 illustrates the effect of P4P in the simple case of $J = K = 1$ and $y = q$. Initial quality, $q0$, is the value of q for which the marginal cost of quality improvement is zero. Assume that target-based P4P is introduced where the target T is set above initial quality. Under P4P, quality increases to $q1$, where the marginal cost curve intersects the marginal revenue curve assuming a symmetric distribution for ε (e.g., the normal distribution). If f is symmetric, then marginal (expected) revenue is greatest just at the target, where $q = T$.

A common criticism of target-based P4P programs is that the target structure discourages very low performers and very high performers from improving. Figure 1 illustrates this clearly. As the absolute value of the distance $q - T$ increases, the marginal revenue from P4P goes to zero, so there is very little incentive to improve. On the other hand, P4P will have its largest impact at some level of initial quality strictly less than the target level. To see this, consider a linear marginal cost curve $\partial C/\partial q = -q_{0i}/c + cq$, where providers differ in their initial quality q_{0i} only. Because f is decreasing in absolute distance from T, $q_{1i} - q_{0i}$ is maximized at $q_{1i} = T$, which implies that P4P has its greatest effect for providers with an initial quality of $q_{0i} = T - rf(0)/c < T$. Note that this level is decreasing in r and increasing in c; that is, as the bonus amount increases (or, as the marginal cost curve flattens) lower-performing providers find it increasingly worthwhile to improve in response to P4P.

Ignoring initial differences in quality, the marginal benefit to increasing q_j can be decomposed into μ_{jk}, the marginal increase in observed measure y_k, and r_k, the price received for each additional unit of y_k, $k = 1, \ldots, K$. A P4P scheme favors quality dimension q_j relative to $q_{j'}$ if $\sum_{k=1}^{K} r_k(\mu_{jk} - \mu_{j'k}) > 0$ (assuming the overall probabilities of reaching the targets are the same). In general, however, $\partial^2 C/\partial q_j \partial q_{j'} \equiv C_{jj'} \neq 0$, so that changing quality along some other dimension $j' \neq j$ will shift the marginal cost curve up or down depending on the sign of $C_{jj'}$. If $C_{jj'} > 0$ (quality dimensions j and j' are substitutes) and if P4P places a large premium on

quality dimension j', then $\partial C/\partial q_j$ may shift up enough to reduce quality dimension j to a level lower than its initial level before P4P was instituted. Note that the model predicts that it is relative prices $r_k\mu_{jk}$ that matter; it is not necessary for $r_k = 0$ for P4P to induce a negative response on measure y_k if y_k largely reflects a quality dimension j that is weakly reflected in other highly rewarded measures.

Finally, the model predicts that μ plays a crucial role in determining which measures will change, and in which directions, as a result of P4P. Suppose, for example, that we add a new measure y_{K+1}, but y_{K+1} is not rewarded by P4P. Assume there are two dimensions of quality, and that P4P strongly rewards the first dimension. Then y_{K+1} will increase if the increase in y_{K+1} due to the increase in q_1 is not offset by the decrease in y_{K+1} due to the decrease in $q_2(\mu_{1,K+1}\Delta q_1 > \mu_{2,K+1} \mid \Delta q_2|)$. In other words, we can predict that the unrewarded measure y_{K+1} will increase in response to P4P if we have *a priori* reason to believe that it is strongly related to the quality dimension(s) determining the rewarded measure set (or, in the case of differential bonuses, the more lucratively rewarded set). Similarly, if y_{K+1} is weakly related or unrelated to the more lucrative quality dimensions, we may expect it to respond negatively to P4P. Certainly, if we believed *a priori* that y_{K+1} should be strongly related in terms of underlying quality to measures for which we observe a negative response to P4P, then we would expect y_{K+1} to respond negatively as well. These theoretical insights will provide guiding intuitions for the empirical results below.

3. Setting

■ We use data from published performance reports of multispecialty medical groups in California and the Pacific Northwest contracting on a capitated basis with a network HMO, PacifiCare Health Systems.[3] PacifiCare is one of the nation's largest health plans, ranked fifth in commercial enrollment by Atlantic Information Systems in 2003. PacifiCare has been collecting quality information on its providers since 1993, although it did not begin making the reports public until 1998. Many of the measures are adapted from the Healthcare Effectiveness Data and Information Set (HEDIS), developed by the National Committee for Quality Assurance (NCQA) and the accepted standard in quality measurement.

In March 2002, PacifiCare of California announced that, as part of a new quality incentive program (QIP) starting in July 2003, it would begin paying quarterly performance bonuses based on selected quality measures published in the reports. Because the reports measured performance over the preceding year with a lag of 6 months, the first payout in July 2003 corresponded to patient care which took place between January 1, 2002 and December 31, 2002. We obtained data from 17 quarterly performance reports issued between July 2001 and July 2005, corresponding to patient care delivered between January 1, 2000 and December 31, 2004. Table 1 summarizes the time structure of our data. Because the provisions of the QIP were not incorporated into the contracts with most of the groups until July 2002, the earliest we may be able to detect a response would be in the April 2003 report (the 8th quarter in our data set). Eligibility was based on the size of the Commercial (CO) and Secure Horizons (SH; covered by Medicare) patient population. Initially 172 medical groups were eligible for the program, with 70 additional groups in the second year.

PacifiCare set targets for five clinical measures at the 75th percentile of performance in the preceding year (2001), and eligible groups received a quarterly bonus of $0.6795 per SH member for each target met or exceeded. Thus, a group with 2,183 SH members (the average number of SH members in 2002) could receive a potential bonus of up to $7,417 quarterly, or $29,667 annually, if it met all five clinical targets.[4] Table 2 lists the clinical quality measures rewarded

[3] Under capitation, healthcare providers are paid a fixed amount periodically for each enrolled patient. Individual medical groups may choose to pay or reimburse their member physicians differently.

[4] The program also rewarded performance on five service measures, which were calculated from patient satisfaction

TABLE 1 Time Structure of the Data

			Period of Care Covered																			
			2000				2001				2002				2003				2004			
t	Report		Ja	A	Ju	O	Ja	A	Ju	O	Ja	A	Ju	O	Ja	A	Ju	O	Ja	A	Ju	O
1	Jul	2001	x	x	x	x																
2	Oct	2001		x	x	x	x						*QIP 1*		*IHA 1/QIP 2*				*IHA 2/QIP 2*			
3	Jan	2002			x	x	x	x														
4	Apr	2002				x	x	x	x													
5	Jul	2002					x	x	x	x												
6	Oct	2002						x	x	x	x											
7	Jan	2003							x	x	x	x										
8	Apr	2003								x	x	x	x									
9	Jul	2003									x	x	x	x								
10	Oct	2003										x	x	x	x							
11	Jan	2004											x	x	x	x						
12	Apr	2004												x	x	x	x					
13	Jul	2004													x	x	x	x				
14	Oct	2004														x	x	x	x			
15	Jan	2005															x	x	x	x		
16	Apr	2005																x	x	x	x	
17	Jul	2005																	x	x	x	x

Notes: In January 2002, the IHA announced it would begin making annual performance-based payments to participating CA groups in mid-2004 for care delivered in 2003. In March 2002, PacifiCare announced its own program, the QIP, which would begin making quarterly payments in mid-2003 corresponding to care delivered from January 2002. Practically, the first year of the QIP corresponded to care delivered between January 2002 and September 2003, and the second year of the QIP corresponded to care delivered between January 2003 and September 2004. See text for details.

by the QIP with their corresponding thresholds. Table 3 presents the mean and median potential bonuses that providers could earn if they met or exceeded these thresholds. Summary statistics for the clinical measures, by region and year, are reported in the Appendix. After 1 year, PacifiCare added five clinical quality measures and readjusted the bonus calculation scheme to allow for a second tier of performance, set at the 85th percentile of the preceding year (2002) and worth twice as much as the first tier. However, the QIP was quickly overshadowed by a much larger P4P effort launched by the Integrated Healthcare Association (IHA) after its first year.

The IHA is a nonprofit statewide coalition of health plans, physician groups, hospitals, and purchasers. Six California health plans—Aetna, Blue Cross of California, Blue Shield of California, CIGNA Healthcare of California, Health Net, and PacifiCare—agreed to pay bonuses to participating California medical groups for performance on a common measure set. These health plans began paying annual bonuses in mid-2004 for patient care delivered in 2003. (A seventh plan, Western Health Advantage, joined the program in its second year.) Table 2 reports the IHA measure sets for 2003 and 2004. Note that the IHA added appropriate asthma medication, but otherwise paid on the same measures as the QIP in its first year. Unlike the QIP, the IHA program was announced a year before it went into effect. In the absence of the QIP, we could have seen whether medical groups improved quality in anticipation of the implementation date. As a result, we cannot disentangle the "IHA anticipation effect" from the pure impact of the QIP. We take January 2003 to be the start date for the IHA initiative, corresponding to the October 2003 report (the 10th quarter in our data), recognizing that we cannot tell when providers actually started responding to the IHA, if they did so before this date.

surveys, as well as six hospital patient safety measures, which were essentially structural quality measures. We ignore this aspect of the program and concentrate solely on clinical quality as measured by process and outcome measures.

TABLE 2 Clinical Measures

Measure	Commonalities in		QIP Thresholds			IHA Measure Set	
	Prod.	Pop	Year 1 (2002) (%)	Year 2 (2003–2004) Tier 1 (%)	Year 2 (2003–2004) Tier 2 (%)	2003	2004
Cervical cancer screening rate among women ages 21–64	IS	W	51.0	60.3	63.8	Yes	Yes
Breast cancer screening rate among women ages 52–69	IS	W	70.6	71.3	73.7	Yes	Yes
Hemoglobin A1c testing rate among diabetics ages 31+	IS	D	72.0	76.8	80.9	Yes	Yes
Childhood immunization rate among children age 2[a]	IS		45.0	72.2	76.2	Yes	Yes
LDL cholesterol testing rate, coronary disease patients and/or diabetics[b]	IS	H/DH	71.4	68.1	72.4	Yes	Yes
Appropriate asthma medication rate, ages 5–56	MD	A		75.0	77.5	Yes	Yes
Preferred antibiotic usage rate in cases of bronchitis or pharyngitis	MD	G		55.6	61.5		
Antidepressant medication management rate, ages 18+	MD			45.6	50.0		
Hospital readmission rate (% of inpatients readmitted within 30 days) (↓)		G		2.8	2.0		
Avoidable hospitalization rate (preventable with optimal outpatient care) (↓)		G		7.2	5.6		
Chlamydia screening rate among women ages 16–26	IS	W					Yes
Eye exam rate among diabetics ages 31+	IS	D					
ACE inhibitor usage rate for congestive heart failure (SH only)	MD	H					
Appropriate use of antibiotics (% of antibiotics prescribed in appropriate cases)	MD	G					
Cholesterol-lowering drugs (% of patients on statin managed properly)	MD	H					
Asthma-related emergency room visits, ages 2–44 (↓)		A					

Notes: Categories for commonalities in production are: identification/scheduling (IS), doctor effort/time (MD). Categories for commonalities in patient population are: asthma (A), diabetes (D), heart (H), women's health (W), and general population (G).

[a]This measure is not comparable before/after year 2 due to changes in method of calculation. In year 1, the threshold for childhood immunization was set higher than the 75th percentile of the preceding year, which was 11.9%.

[b]This measure is not comparable before/after year 2 due to changes in population.

The successive introduction of the QIP and IHA programs provides a unique opportunity to examine the responses of medical groups to different aspects of P4P programs. First, when the other plans in the IHA coalition adopted P4P, this dramatically increased the size of potential bonuses (on the order of 10 times for the average group). Together, the health plans participating in the IHA program accounted for an average of roughly 60% of capitated revenues of the California

TABLE 3 Distribution of QIP Quarterly Potential Bonus for Clinical Measures

	Year 1		Year 2			
			Tier 1		Tier 2	
	Per Measure ($)	×5 Targets ($)	Per Measure ($)	×10 Targets ($)	Per Measure ($)	×10 Targets ($)
Medical groups with at least 100 SH members and at least 1000 CO members						
Example: group with 2000 SH members	1,359.00	6,795.00	450.00	4,500.00	900.00	9,000.00
Minimum	27.18	135.90	22.50	225.00	45.00	450.00
Median	914.61	4,573.05	285.53	2,855.25	571.05	5,710.50
Mean	1,414.90	7,074.50	452.02	4,520.15	904.03	9,040.30
Standard deviation	1,590.99	7,954.95	520.11	5,201.10	1,040.22	10,402.20
Maximum	10,088.54	50,442.70	3,212.78	32,127.75	6,425.55	64,255.50
Medical groups with less than 100 SH members and at least 1000 CO members						
Minimum	0	0	156.75	1,567.50	313.50	3,135.00
Median	0	0	533.70	5,337.00	1,067.40	10,674.00
Mean	0	0	717.00	7,169.95	1,433.99	14,339.90
Standard deviation	–	–	712.81	7,128.10	1,425.62	14,256.20
Maximum	0	0	4,037.70	40,377.00	8,075.40	80,754.00

Notes: For groups with at least 100 SH members and 1000 CO members, quarterly potential bonus per measure is calculated by multiplying SH membership by 3*.2265 in year 1 of the program, by 3*.15*.5 for tier 1 (75th–85th percentile) in year 2, and by 3*.15 for tier 2 (greater than 85th percentile) in year 2. Groups with less than 100 SH members were not eligible for the QIP in year 1; however, those with at least 1000 CO members were eligible in year 2 for a potential per-measure bonus of (CO membership) multiplied by 3*.1*.5 for tier 1 or by 3*.1 for tier 2.

medical groups.[5] Total performance payments from IHA-affiliated groups (including payments for nonclinical and non-IHA performance measures) amounted to more than $122.7 million in 2004 and $139.5 million in 2005. PacifiCare's QIP accounted for only 16% of the total payout in 2004, and only 10% in 2005. The IHA program was not just bigger in terms of absolute dollar amounts, it also made performance bonuses attainable for the lower-performing groups, as the biggest payers such as Blue Cross and Blue Shield made payments to groups above the 20th and 30th percentiles, respectively. Although the measure set was common across health plans, each plan individually decided on the size and structure of the awards it offered. In particular, PacifiCare and Health Net were the only plans to use absolute thresholds for determining payment; the rest of the plans based their payments on relative rankings of providers. (See Damberg et al. [2005] for more details on the IHA program; in addition, the IHA's Financial Transparency Reports are publicly available at www.iha.org.) Thus, part of the increase in dollars paid can be attributed to the fact that PacifiCare had stricter requirements (i.e., higher thresholds).

The interaction of the QIP and IHA programs also provides a unique opportunity to examine the responses of medical groups when measure sets diverge. In the first 6 months of P4P, California medical groups were paid small bonuses for performance on five measures which rely primarily on identifying patients in appropriate risk groups and successfully scheduling patient visits.[6] The IHA program increased the size of the bonuses for these identification/scheduling measures, while at the same time PacifiCare added five new measures which rely primarily on doctors' prescribing and managing the right medications (as well as outcomes, which, theoretically, could be controlled with optimal outpatient care). In other words, these measures could potentially be improved by focusing on interventions at the doctor level.

Thus, we can estimate responses to P4P when one type of measure is rewarded more or less than others (where "type" refers to measures grouped on commonalities in production). As

[5] Glied and Zivin (2002) provide evidence that, in a mixed payment environment, healthcare providers respond to the incentives of their modal patient. Unfortunately, we do not have data on PacifiCare or IHA's share of total enrollment, so we cannot evaluate the effect of the increased "salience" of the program.

[6] Generally the measures do not correlate highly. However, cervical cancer screening, HbA1c testing, and chlamydia screening are all highly correlated with one another, on the order of 0.5–0.7, lending some support to our hypothesis that these measures may have similar production technologies despite differences in patient population.

we saw in Section 2, in theory even a rewarded measure could decrease in response to a P4P program that provides substantially higher rewards to other measures (a relative price effect). If this is the case, then it underscores the fact that payers considering implementing P4P should take into account any other existing or proposed incentive programs. In the next section, we describe the empirical specifications that we estimate and explain how they relate to our hypotheses about providers' responses to P4P.

4. Empirical strategy

■ To examine healthcare providers' responses to the introduction of P4P in California, we use longitudinal data on 14 clinical quality measures, 9 of which were rewarded by one or more health plans at some point during the period we study.[7] All but one of our measures are rates, for which we have data on both numerators and denominators (where the denominator represents the number of PacifiCare patients enrolled in the medical group who are clinically indicated to receive a screening or treatment). We restrict our sample to medical groups with complete data on one or more measures reported in the July 2001 to July 2005 performance profiles published by PacifiCare. Note that some measures are not available for all 17 quarters due to definition changes and the introduction of new measures. We consider only those measures reported at least two quarters before the first wave of P4P began. Note that we also observe a number of mergers between medical groups in our sample. In these cases, we combine the numerators and denominators for groups that eventually merge with one another, so that these groups are treated as one entity throughout our time frame.

We would like to estimate the effects of P4P on unobserved quality q_{it}, for medical groups $i = 1, \ldots, N$, at time $t = 1, \ldots, T$, but we are restricted to estimating the effects on observed performance measures y_{kit}, $k = 1, \ldots, K$, which reflect unobserved quality as in equation (1). We hypothesize that q_{it} is multidimensional, but that the measures reflect primarily one of two dimensions of quality: identification/scheduling (IS) and physician-level care (MD). If we restrict our analysis to the California medical groups, we start with the following equation:

$$q_{jit} = \alpha_0^j + \alpha_1^j \, QIP1_t + \alpha_2^j (QIP2 \cdot IHA1)_t + \alpha_3^j (QIP2 \cdot IHA2)_t + \alpha_4^j t, \quad j = IS, \; MD,$$

where $QIP1_t$, $(QIP2 \cdot IHA1)_t$, and $(QIP2 \cdot IHA2)_t$ are mutually exclusive dummy variables denoting which P4P regime, if any, was in effect at time t. Note that we cannot separate out the effects of the IHA initiatives from the effects of PacifiCare's adjustment of its own QIP measure set and bonus structure in its second year. We assume a linear time trend and estimate the effects of P4P as breaks in this time trend. Then for a given dimension j, measure k can be written $y_{kit} = \mu_k(q_{jit}) + \varepsilon_{kit}$. With the exception of asthma-related emergency room visits, all of our measures are proportions, so we observe $0 \leq y_{kit} \leq 1$. Because we have information on the numerators n_{kit} of our outcome variables, we assume that $n_{kit} y_{kit}$ is distributed Binomial(n_{kit}, p_{kit}), where $p_{kit} = \mu_k(\cdot)$. A natural choice for the link function μ_k is the cdf of a known distribution function. We let $\mu_k(z) = \Lambda(\lambda_k z)$, where $\Lambda(z) = \exp(z)/(1 + \exp(z))$, the logistic function.[8] For asthma-related ER visits, we assume a Gaussian distribution with an identity link function $(\Lambda(z) = z)$. Note that we cannot separately identify $\alpha^j \lambda_k \equiv \beta^k$. Thus, we estimate the following reduced-form equation for each measure k.[9]

[7] We exclude LDL cholesterol testing due to changes in population (adding diabetic patients to coronary artery disease patients) at the beginning of year 2. We also exclude antidepressant medication management due to lack of preperiod data.

[8] To evaluate the robustness of our results, we also estimate the models using a power function specification $(\Lambda(z) = z^m, \; m = 1/2 \text{ and } m = 2)$. The results under these alternative specifications are quantitatively and qualitatively similar to the results we present here.

[9] Because of the wide variety of measures, no medical group has observations on every measure, so we cannot restrict the sample to be the same across regressions. If groups with mammography scores are generally different from those with antibiotics scores, then these differences may be reflected in the regression coefficients.

$$y_{kit} = \Lambda\left(\beta_0^k + \beta_1^k\, QIP1_t + \beta_2^k(QIP2 \cdot IHA1)_t + \beta_3^k(QIP2 \cdot IHA2)_t + \beta_4^k t\right) + \varepsilon_{kit}. \qquad (4)$$

We estimate (4) by the generalized estimating equation (GEE) method of Liang and Zeger (1986). GEE is an extension of the quasi-likelihood-based generalized linear model to autocorrelated panel data. GEE is a pooled model, that is, the estimates are "population averaged," as medical group random effects are averaged out across the sample. Whereas in random effects models the coefficients represent conditional effects for given values of the random effects, GEE estimates can be interpreted as marginal effects. GEE also has an advantage over random effects models because the parameter estimates are consistent even if the correlation structure is misspecified (Liang and Zeger, 1986). With this in mind, we estimate our models using an AR(3) structure for the error term, a flexibly parametric specification, even though the overlapping nature of our data induces at least an MA(4) error term. In reality, the error structure is probably a mixture between MA and AR processes. Note that we cannot estimate a completely nonparametric specification for the working correlation matrix because we do not have enough observations to estimate each element of R (e.g., for 17 quarters, there are $17 * (17 - 1)/2 = 136$ unique elements in the working correlation matrix).[10]

We exclude two measures, hemoglobin A1c testing and cholesterol-lowering drugs, from this analysis as we only observe two quarters of preperiod data. For the remaining measures, we tested the assumption of a linear time trend by creating a dummy variable for each quarter before P4P was introduced ($t < 8$) and regressing each measure on quarter t and the dummies. We then performed an F test of joint significance of the coefficients on the quarter dummies and dropped those measures where the p-value from this test was less than 0.10. The measures which failed this test were: preferred antibiotics, avoidable hospitalizations, ACE inhibitor usage, and appropriate antibiotic usage. This leaves us with eight measures for the California-only before-after analysis.

To control for shocks to our time series during the post-P4P period, we also make use of a comparison group consisting of medical groups in the Pacific Northwest (i.e., Washington and Oregon) also contracting with PacifiCare, and reporting the same measures, but not under any pay-for-performance program. To estimate this difference-in-differences (DID) approach, we modify (4) as follows:

$$y_{kit}^g = \Lambda\left(\begin{array}{c}\beta_0^k + \beta_1^k\, QIP1_t^g + \beta_2^k(QIP2 \cdot IHA1)_t^g + \beta_3^k(QIP2 \cdot IHA2)_t^g \\ + \sum_{\tau=1}^{T} \gamma_\tau^k d_\tau + \delta^{gk}\end{array}\right) + \varepsilon_{kit}^g, \qquad (5)$$

where g indexes treatment and "control" groups. The primary identifying assumptions of this method are: (i) that the treatment was randomly assigned to one group over the other, and (ii) that the treatment and control groups are influenced by the same variables over time (or, more generally, that quarter-to-quarter changes are roughly the same for both groups). (See Meyer [1995] for a more complete inventory of threats to the validity of DID models.) These are both strong assumptions. For example, if P4P was instituted in California instead of the Northwest because California groups were expected to be more responsive, then DID overestimates the causal effect of P4P for the average medical group. More generally, we require $E[\varepsilon_{kit}^g \mid QIP1_t^g, (QIP2 \cdot IHA1)_t^g, (QIP2 \cdot IHA2)_t^g] = 0$.

Closely related is our second identifying assumption, which states that the treatment and control groups are subject to the same shocks over time. This assumption is crucial to maintaining DID's advantage over simple before-after comparisons. If the "control" group experiences some, say positive, shock in the postperiod which is not experienced by the treatment group, then DID estimates of the effect of P4P will be biased downward. In general, if the dependent variables for the treatment and control groups move together in the preperiod, then we may have more

[10] We also estimated a range of models with different AR(p) specifications, as well as a stationary($T - 1$) specification for measures with enough observations. We found that the magnitude of the estimates was not very sensitive to the specification of R, but the standard errors tended to increase the more parameters estimated.

faith in our estimates. However, this occurrence does not definitively point to the validity of this assumption.

Note that, in the DID model, it does not matter whether the preperiod levels are different across treatment and control groups as long as the measures move together. However, as we use proportions data, this may pose a problem if either of the time series is at a ceiling or floor. (Note the ceiling is not necessarily 1, as some factors, such as patient compliance, may be beyond the control of the medical group.) For example, if the ceiling for cervical cancer screening is 70%, then DID on levels would understate the effect of P4P. In general, any difference in levels between treatment and control groups is cause for concern when using proportions data, as an identical change in a dependent variable can have drastically different effects on the proportion, depending on initial values. The GEE model solves this problem by estimating DID on the log odds ratios of the measures. In general, failing to account for nonlinearity is more important the more dispersed the observed values over the [0,1] line, especially at the extremes.

To test for parallel movement in the preperiod trends for CA versus NW medical groups, we interacted the quarter dummies with a dummy for the control group, and regressed each measure on the interacted and noninteracted quarter dummies and control group dummy. We then performed an F test of joint significance of the coefficients on the interacted dummies, and dropped those measures where the p-value from this test was less than 0.10. This led us to drop one measure: diabetic eye exam. This leaves us with 13 measures for the DID analysis.

5. How did providers respond to pay-for-performance?

■ In this section, we present our estimates of equations (4) and (5) on each of our clinical quality measures, which vary in terms of their predicted responses to P4P. These effects are estimated both relative to a linear time trend and using the DID approach. The results are generally robust to inclusion of quadratic, cubic, and quartic time trends. In the DID models, we assume a completely nonparametric specification for the influence of time by replacing the parametric time trend in the CA-only models with fixed effects for each time period. This flexibility comes at the cost of assuming that, on average, quarter-to-quarter changes are identical for medical groups in California and our control group, the Pacific Northwest. In our discussion below, we note where we think there is reason to believe this assumption may not hold. We present our estimates of equations (4) and (5) for each measure in Table 4.

Despite the small sample size, the results are fairly sharp, and in cases where we find null effects the point estimates are generally small in magnitude (less than 1 percentage point), lending confidence that they are not merely reflective of low power. We find that some of the measures rewarded by P4P improve when the program is introduced, and they improve even more when the bonuses are increased. These measures are ones that we predict, *a priori*, to share some commonality in production, namely they rely on identification/scheduling (IS) for improvement. By contrast, measures that we hypothesize to depend more on doctor time/effort (MD) tended to fall with the introduction of P4P. These included some MD measures which were actually rewarded by one or both programs. Although this result is surprising at first glance, it is consistent with the fact that both programs emphasized the IS dimension over the MD dimension in determining their measure sets. We do not uncover any important spillovers along the lines of shared population or disease groups.

□ **Can one payer make a difference?** The P4P program introduced first in our California sample, PacifiCare's QIP, paid on four measures in our data set: cervical cancer screening, breast cancer screening, hemoglobin A1c (HA1c) testing for diabetics, and childhood immunization. The threshold for childhood immunization was the only one set above the 75th percentile of the preceding year. In fact, the maximum immunization rate in 2000 was 15.38%, well below the threshold of 45%. As the expected bonus on childhood immunization was essentially zero for all medical groups in our sample, we might therefore consider any changes in childhood

TABLE 4 Estimates of Effect of Pay-for-Performance on Clinical Quality Measures

	QIP 1			IHA 1/ QIP 2			IHA 2/ QIP 2		
Measure	Paid	(1) CA	(2) DID	Paid	(1) CA	(2) DID	Paid	(1) CA	(2) DID
Paid IS measures									
Cervical cancer screening	$	−0.285	−0.043	$$	3.625***	3.499***	$$	8.812***	6.009**
		(0.357)	(0.907)		(1.202)	(1.373)		(1.737)	(2.367)
Breast cancer screening	$	0.237	−1.067	$$	1.169*	0.118	$$	1.193	1.283
		(0.380)	(0.737)		(0.675)	(1.068)		(0.767)	(1.184)
Hemoglobin A1c testing	$		1.357	$$		−3.756*	$$		1.916
			(2.388)			(2.083)			(2.351)
Childhood immunization	$	−0.471	3.155**	$$	−1.092**	2.078*	$$		
		(0.385)	(1.365)		(0.485)	(1.196)			
Paid MD measures									
Appropriate asthma medication	No	−1.591**	−0.635	$$	−1.884	1.548	$$	−7.970***	−2.270
		(0.696)	(3.097)		(1.157)	(3.434)		(1.407)	(3.580)
Preferred antibiotic usage	No		1.402	$		−2.830*	$		−3.443**
			(1.181)			(1.670)			(1.670)
Positive spillovers?									
Chlamydia screening	No	−1.922**	−2.506**	No	−2.706***	−5.264***	$$	2.090*	−1.625
		(0.520)	(1.103)		(0.957)	(0.1613)		(1.221)	(2.314)
Diabetic eye exam	No	0.758*		No	−0.661		No		
		(0.410)			(0.505)				
Negative spillovers?									
ACE inhibitor for CHF	No		0.598	No		0.924	No		−0.448
			(0.562)			(0.706)			(0.884)
Appropriate use of antibiotics	No		−0.583	No		−2.123	No		−4.048*
			(1.490)			(1.484)			(2.249)
Cholesterol-lowering drugs	No		0.195	No		0.112	No		0.353
			(0.220)			(0.341)			(0.366)
Intermediate outcomes									
Hospital readmission	No	−0.129***	−0.175	$	−0.187**	−0.196	$	0.174*	−0.338*
		(0.050)	(0.134)		(0.094)	(0.181)		(0.098)	(0.186)
Avoidable hospitalization	No		−0.151	$		0.088	$		−0.265
			(0.548)			(0.678)			(0.700)
Asthma-related ER visits	No	−0.160***	−0.165	No	−0.205**	−0.274	No	−0.269***	0.082
		(0.053)	(0.013)		(0.082)	(0.185)		(0.084)	(0.177)

Notes: $ denotes in QIP measure set, $$ denotes in both QIP and IHA measure sets. CA models (specification 1) are estimated with linear time trend. DID models are (specification 2) are estimated with quarter and region fixed effects. All models assume binomial distribution with logit transformation except asthma-related ER visits, which assumes Gaussian distribution with identity link function. Robust standard errors are in parentheses, clustered on medical group. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

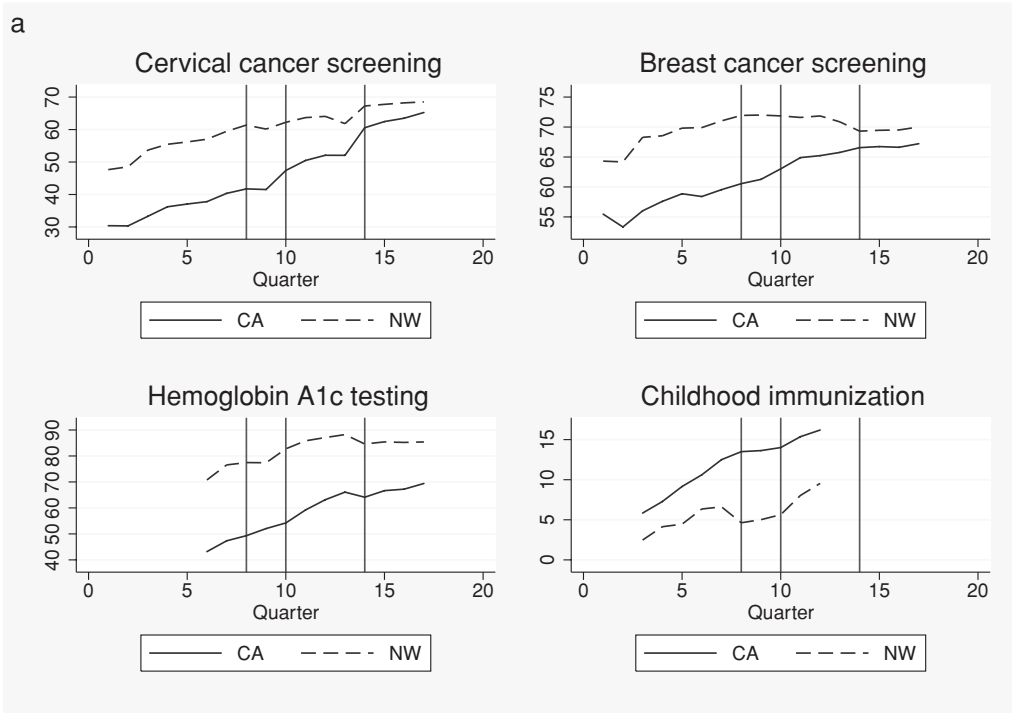
immunization to be "spillover effects," as though childhood immunization were an unpaid measure.[11] Figure 2a plots the average values of these measures, by region and quarter. The Appendix also reports distributional statistics for the measures.

Recall that the QIP was announced in March 2002 ($t = 7$) and incorporated into the contracts of most groups by July ($t = 8$), even though it paid for care delivered from January ($t = 6$). We take $t = 8$ to be the starting point for the QIP. Recall also that we cannot distinguish an "anticipation effect" from the announcement of the IHA initiative in January 2002, $t = 6$. Any anticipation of the IHA initiative will tend to bias our estimate of the effect of the QIP away from zero. Setting

[11] This assumes that medical groups respond to P4P only insofar as it affects them financially. That is, they do not redirect resources toward "rewarded" measures simply because the program draws attention to those measures.

FIGURE 2

AVERAGE QUALITY BY REGION AND QUARTER



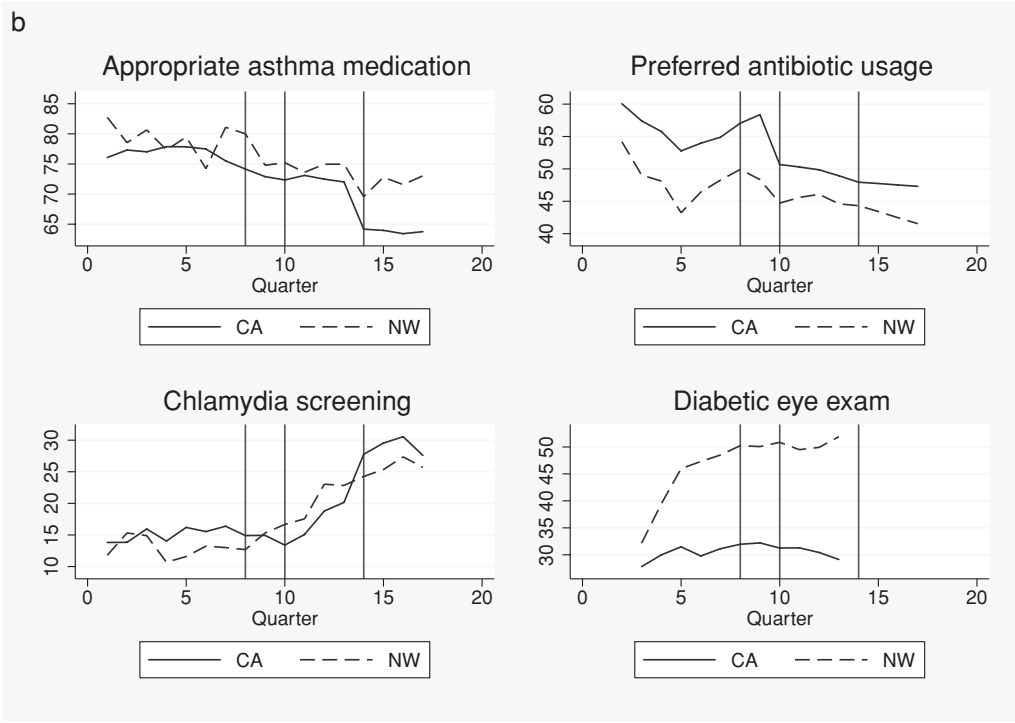
aside anticipation, the starting point for the IHA initiative is clear, at $t = 10$ (confounded with the second year of the QIP), with the second year of the IHA program beginning at $t = 14$. Note that we only observe two quarters where the QIP is in effect before the IHA initiative begins. To the extent that it takes longer for changes in quality to be reflected in the indicators, our estimate of the QIP effect is biased toward zero.

Table 4 reports estimates of the marginal effects of P4P, estimated on the California sample only (specification 1) imposing a linear time trend, and compared to medical groups in the Pacific Northwest (specification 2). None of the three paid measures (excluding childhood immunization) was estimated to be significantly affected by the QIP. By contrast, we estimate a 3 percentage point effect of the QIP on child immunization rates. However, Figure 2a illustrates that this effect is entirely driven by a dip in childhood immunization rates in the Northwest in 2002. If that dip was region specific, then DID overestimates the effect of P4P in California. (This is true in general, as it violates our identifying assumption that treatment and control groups are subject to the same quarterly shocks.) This estimate highlights the danger of relying solely on differences in differences, even when the treatment group tracks the control group reasonably well in the preperiod. Because the CA-only and DID estimates differ so dramatically for childhood immunization, we are reluctant to draw conclusions from these results.

Our results are consistent with Rosenthal et al. (2005), who combine the QIP and year 1 of the IHA program into one P4P indicator. They estimate the effects of P4P on cervical cancer screening, breast cancer screening, and HA1c testing, and find a positive significant effect for cervical cancer screening only, relative to the control group. Their estimate of a 3.6 percentage point effect on cervical cancer screening is very close to our estimate of 3.5–3.6 percentage points, which we attribute entirely to the IHA program, initiated 6 months into the P4P regime. The IHA program linked P4P to plans that accounted for 60% of providers' revenues, resulting in dramatically higher payments to medical groups participating in P4P after 2003.

FIGURE 2

CONTINUED



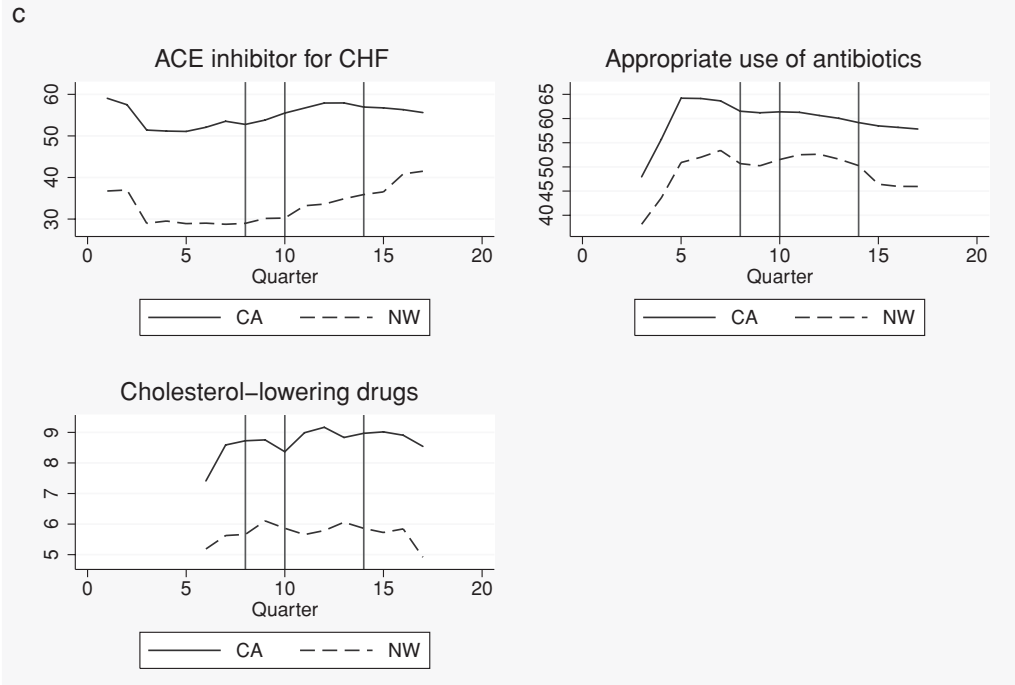
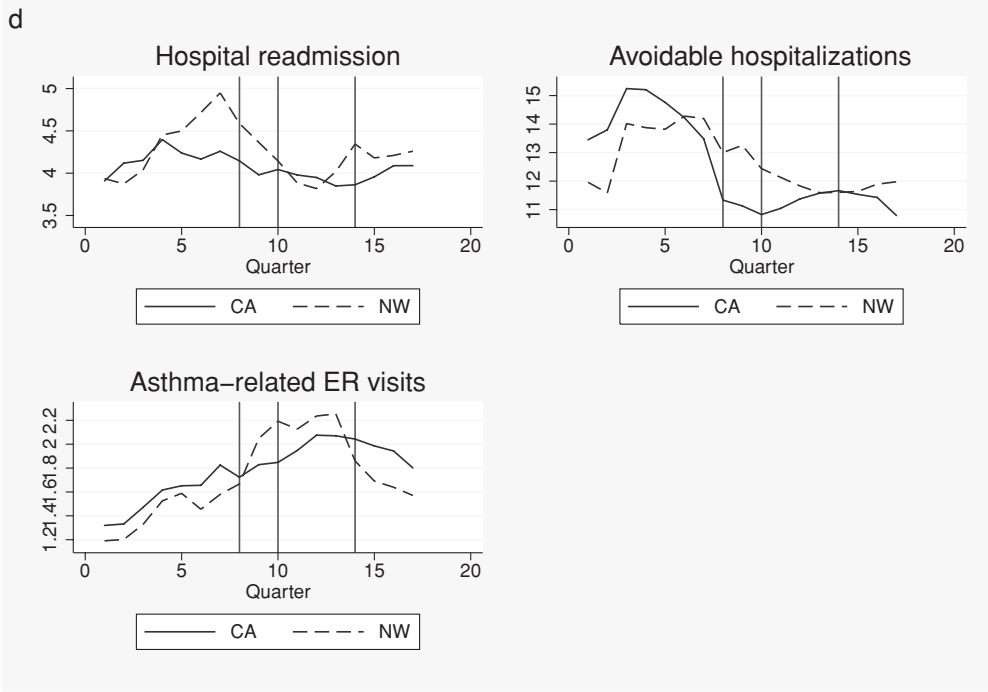


FIGURE 2

CONTINUED



□ **Does paying more matter over time?** As noted earlier, PacifiCare accounts for approximately 15% of total capitated revenues of the medical groups we examine, compared with 60% for the combined IHA plans. Thus, the introduction of the IHA plan represents a sizeable increase in potential revenue from P4P. All of the above measures were included in the IHA measure set introduced the following year. Looking at the time series for California only in Figure 2a (solid lines), cervical cancer screening appears to be the only measure to depart (positively) from trend around the time the IHA initiative was instituted. In fact, it appears to make a second jump around $t = 14$, when the second phase of the IHA initiative was instituted. Neither breast cancer nor childhood immunization rates appear to respond to P4P, from the time series, and, if anything, HA1c testing rates appear to dip down around $t = 14$.

This is confirmed in Table 4. The CA-only model estimates the effects of the IHA program on cervical cancer screening rates to be 3.6 and 8.8 percentage points in years 1 and 2 of the program, respectively. The CA-only model also estimates a small effect of 1.17 percentage points on breast cancer screening rates in year 1, which is significant at the 10% level. Comparing the CA groups to those in the Northwest, we still see a positive effect of the IHA program on cervical cancer screening rates. The DID model reports estimates that are roughly the same size as the CA-only estimates at 3.5 and 6 percentage points, for the first and second years of the IHA program, respectively.

As before, the results for child immunization rates are inconclusive, as the CA-only and DID estimates differ so dramatically. The DID estimates for hemoglobin A1c testing are also inconclusive. We find a statistically significant *negative* impact in the first year of the IHA program, but the estimates for the other P4P regimes are positive and insignificant, suggesting caution in interpreting the result.

In addition to the above four measures, appropriate asthma medication was included in the IHA common measure set starting in 2003. Figure 2b plots average appropriate asthma medication rates, by region, over time. Surprisingly, it is immediately apparent that there is a

sharp 8 percentage point drop in asthma medication rates going into the second year of the IHA initiative relative to preperiod performance, which seems to be stable, or trending slightly downward, leading into the postperiod.[12] The CA-only estimates are consistent with the graphical evidence, with large estimated impacts. The estimated difference is reduced to −2 percentage points relative to the control group in the DID model, which is not statistically significant. Note that we only have seven observations from the NW control group on asthma medication.

Even though asthma medication is included in the IHA performance measure set, it is one of six such measures, where the other five line up along the identification/scheduling (IS) dimension, according to our hypothesis of commonality in the production of clinical quality. From the medical groups' perspective, even if this one measure is rewarded, profit maximization may imply substitution away from the MD quality dimension toward IS-directed improvements, thereby increasing performance on the other five measures. If we see decreases in other MD measures which were unpaid, this may provide evidence that medical groups are responding to P4P by substituting away from (relatively) unrewarded toward rewarded dimensions of care. We explore evidence on multitasking below.

Finally, in the second year of the IHA program, chlamydia screening was added to the common rewarded measure set. Because chlamydia screening was not rewarded prior to 2004, we can attribute any changes during the QIP and IHA year 1 periods to substitution or commonality spillovers from other rewarded measures. Figure 2b shows the time series of chlamydia screening for the CA and NW groups, whereas Table 4 reports the model estimates. Even though chlamydia screening declined during the first year and a half of P4P, relative to its trend and compared with rates in the Pacific Northwest, we find that this decline reversed itself when chlamydia screening was added to the measure set in 2004. The CA-only model estimates a positive response to the second year of the IHA; however, this result is not robust in the DID model.

To summarize, we find that the IHA initiative, in contrast to the QIP alone, did motivate changes in some quality measures. We find evidence for positive improvements in cervical cancer screening and chlamydia screening when these measures were rewarded by the IHA. The improvement in cervical cancer screening did not wane after the first year of the IHA program, and indeed almost doubled. A puzzling result is the estimated negative impact of P4P on appropriate asthma medication, even though it was rewarded by the IHA starting in 2003. A possible explanation is that appropriate asthma medication, an MD-level measure, suffered because the IHA measure set emphasized the IS dimension of quality over the MD dimension.

□ **Is there evidence of commonality in multitasking?** In the second year of the QIP, PacifiCare added four measures to the set of rewarded measures (antidepressant medication management was also added, but we do not have preperiod data): appropriate asthma medication, preferred antibiotic usage, hospital readmission, and avoidable hospitalizations. Only one of these measures, appropriate asthma medication, overlapped with the IHA measure set, meaning that it was "worth" approximately 10 times these other measures to California medical groups after 2003 ($t = 10$). For the latter two measures the medical groups may have little control over performance. However, if P4P really does improve important aspects of outpatient care, then we could see real declines in these measures (recall that these outcomes reflect adverse events, so lower is better), assuming these improvements outweigh any negative impacts on outpatient care. We discuss the impact of P4P on overall health outcomes further below.

Table 4 reports that, compared to the Northwest, preferred antibiotic usage did not change during the QIP but decreased by 2.8 percentage points during the IHA initiative.[13] This provides

[12] In 2005, two health plans, Blue Cross and Health Net, introduced financial incentives for generic prescribing in addition to the clinical measures. This may have hurt appropriate asthma medication because most of the controller medications for asthma are brand name only.

[13] In 2005, some plans introduced incentives for generic prescribing in addition to clinical quality. Note that in the case of preferred antibiotic usage these additional incentives should have reinforced the existing P4P incentives, because most of these drugs are generic.

some evidence that relative benefit-cost ratios matter in terms of provider response to P4P. Even if a measure is rewarded by P4P, if other measures are rewarded significantly more, or if they cost less to improve, then providers may substitute toward the more lucrative measures, causing the measure with the smaller reward to fall.

Relative differences in quality awards may not operate only at the level of individual quality indicators, but also at the level of quality dimensions, if there is commonality in multitasking. (Unfortunately, we do not observe any IS measures which were rewarded by the QIP but not the IHA.) As discussed above, even though appropriate asthma medication was included in the well-paid IHA measure set, we find that it may have actually declined in response to P4P. Even though the measure itself was rewarded highly, the MD dimension of quality that it reflected was only weakly rewarded by the IHA.

The availability of data on unpaid clinical quality measures allows us to examine spillover effects of P4P. These unpaid measures include: diabetic eye exams, ACE inhibitor usage for seniors with congestive heart failure (CHF), appropriate use of antibiotics, management of cholesterol-lowering drugs, and asthma-related ER visits. In addition, chlamydia screening was unpaid until 2004.

Because chlamydia is an IS measure, and shares its population focus (i.e., women's health) with two of the main rewarded measures (cervical cancer screening and breast cancer screening), we expect to see it increase after P4P was introduced. Instead, we find that chlamydia screening rates actually decreased by about 2–5 percentage points during the QIP and the first year of IHA, relative both to its time trend and to the NW control group. This is surprising, as chlamydia screening is positively correlated with other IS measures (e.g., breast cancer screening, HA1c testing). One explanation is that, even though chlamydia screening may be increased by making a general quality improvement, such as instituting an automated reminder program, that increases other, paid measures, the cost to including the criteria for chlamydia screening is still positive, even if it is small.

Similarly, we hypothesize that diabetic eye exam rate shares commonalities in production with other IS measures; in addition, diabetics were one of the patient groups emphasized by both the QIP and IHA efforts. (Diabetic eye exams could be classified as either an IS or MD measure, as the eye exams require some MD effort. Any expected gains from improvements on the IS side may be tempered if there is substitution from the MD side.) Although diabetic eye exam rates did increase slightly (less than 1 percentage point) following the initial introduction of P4P with the QIP, they leveled off and even declined slightly after the larger IHA initiative was introduced (Figure 2b). Despite the potential for positive spillovers, it does not appear that any real gains were made. This may be due to the fact that eye exams require a separate referral to an ophthalmologist.

With the exception of asthma-related ER visits, which we discuss below, the rest of these measures deal with appropriate prescription and management of medications. These MD-level measures were de-emphasized by P4P efforts compared to the IS dimension, so we may expect to find reductions if provider groups responded by substituting away from the MD quality dimension. On the other hand, if spillovers across populations or disease groups are more important than spillovers in production technologies, then we may expect to see increases in performance on measures corresponding to patient populations targeted by P4P (i.e., women, diabetics, and heart patients).

Turning to the heart-related measures, we do not see any convincing changes relative to the control group. This could happen if the commonality in production with the MD dimension puts downward pressure on the measures, while at the same time the commonality in patient population puts upward pressure on the measures. The only heart-related measures we have are MD measures, so we cannot separate out a "heart-related" spillover effect. However, we do have an MD measure which is unrelated to any patient groups emphasized by P4P: appropriate use of antibiotics, which might be expected to fall in response to the QIP and IHA programs. Figure 2c shows time-series plots of these measures. Appropriate use of antibiotics begins a very

slight decline after the introduction of P4P and, compared with the control group, drops 2–4 percentage points after the IHA initiative is introduced. (Note that we only observe five control observations for this measure.)

Unfortunately, these measures do not give a clear-cut picture of response patterns to P4P. One surprising result is the lack of positive spillovers to the other IS measures. We tentatively conclude that, even if some measures may be increased by general quality improvements to shared dimensions, we may not actually see such an increase as there is still a cost to expanding the improvement to encompass those unpaid measures when there is no return. If medical groups are focusing on the measures themselves more than on underlying dimensions of quality, it makes P4P an ineffective tool for motivating general quality improvements.

□ **Are there global effects on health?** Changes in outcome measures, such as avoidable hospitalizations, inpatient readmissions, and asthma-related ER visits, are difficult to interpret given the complexity of production in healthcare markets. An open question is whether these measures assess aspects of quality that doctors have enough control over to respond to P4P incentives. Even if medical groups do not respond to P4P by directly targeting outcome measures, we may still see movement in these measures if the groups did respond for measures important for optimal outpatient care. Asthma-related ER visits, for example, are negatively correlated with appropriate asthma medication in our data, although it seems plausible that these are generally long-run relationships.[14]

Table 4 reports that hospital readmission and asthma-related ER visits did fall significantly relative to their preperiod time trends. These differences are generally insignificant when compared to the NW groups; however, for these measures, it may be harder to maintain the assumption of identical quarterly shocks across regions. Outcomes such as these are complex functions of many factors, and depend a great deal on patient characteristics. For this reason, we view the DID estimates as less reliable.

One problem with interpreting our estimates of the effect of P4P on such outcome measures is that clearly a lag is necessary to allow these outcome measures to reflect underlying changes in the quality of outpatient care. However, it is not obvious how long such adjustments should take. If we assume that a large part of the adjustment takes place within a year, then we may consider our estimates of the coefficients on IHA2 to reflect changes in response to the IHA1 regime.

If we can attribute postperiod increases in health status to P4P, then this may provide some evidence on whether the costs of P4P in potential losses to patients on unrewarded measures are more than offset by the gains from increasing rewarded measures. These differences represent sizeable effects. For example, if we assume an average cost of $5,300 per admission, then an increase (decrease) of 1.5 hospitalizations in 100 amounts to an annual cost increase (savings) of $80 per member, or more than $134 million network-wide.[15] However, without more complete data measuring positive health events/status in addition to these few adverse events, we cannot draw more general conclusions about the global effects of P4P. Although it is difficult to conclude whether P4P had a significant impact on overall health, at least it does not look like the programs resulted in adverse net effects on these measures.

□ **Do providers' responses vary by financial incentive?** Even though the QIP had a negligible effect on providers' performance on average, it may have had an impact on those providers for whom the potential bonus represented a sizeable financial reward. Recall from Table 3 that, in the first year of the program, quarterly potential bonuses ranged from $27 to more than $10,000 per measure among eligible medical groups. Because we know the benefit formula, we can

[14] Avoidable hospitalizations are those that medical experts agree can to a large extent be avoided with optimal outpatient care. These conditions are: angina, asthma, cellulitis, chronic obstructive pulmonary disease, congestive heart failure, diabetes, hypertension, kidney/urinary tract infections, pneumonia, and immunizable conditions.

[15] The estimate of the cost of hospital admissions is taken from Kruzikas et al. (2000).

compute each provider's potential quarterly bonus for achieving the targeted performance level, and estimate the interaction between the potential bonus and the introduction of the QIP on two paid measures: cervical cancer and breast cancer screening. If we can adequately control for the marginal cost of quality improvement, then this should give us the marginal expected improvement in performance per dollar pledged to the QIP program.

Note that the potential bonus depended directly on the number of PacifiCare's Secure Horizons (SH), or Medicare, patients served, rather than total enrollment including commercial members. On average, SH members accounted for about 20% of medical groups' PacifiCare enrollment, ranging from 4.5% to 63.5%. If we assume that PacifiCare accounts for a constant fraction of all managed care enrollment, we can estimate the impact of the size of the bonus while controlling for the size of the practice (as measured by total enrollment). We can also distinguish between scale effects from overall practice size (total enrollment) and number of patients in a given risk group (denominator), which may have different signs if there are returns to scale more generally, but it is harder to manage, say, diabetic patients if there are too many of them.

In addition, recall from Section 2 that initial performance should be related to providers' responses to target-based P4P. In particular, the effect of the program should be greatest at some initial level of performance below the target threshold, and should decrease as the absolute distance between initial and target performance increases. Recall that the QIP thresholds were based on the 75th percentile of performance in the year before it was introduced. We divide initial performance (from calendar year 2000) into four quartiles and estimate their interaction with the QIP. Assigning the fourth (top) quartile as the omitted category, we hypothesize that the coefficients on these interaction terms will be positive and hump-shaped, with a maximum effect just below or below the 75th percentile (i.e., quartile 3 or 2).

Finally, the target structure of the QIP may induce cross-substitution among measures if one measure is just under the target while another measure is safely above its own target level. We pool the lower two quartiles of performance and interact the quartiles for cervical cancer and breast cancer screening. We estimate the effects on responsiveness to the QIP for four interactions: low/middle, middle/low, middle/high, and high/middle, where middle refers to the 3rd quartile, or just below the target. We hypothesize that low or high performance on cervical cancer screening combined with middle performance on breast cancer screening will have a negative effect on cervical cancer and positive effect on breast cancer screenings, and vice versa.

Table 5, panel (1), reports estimates of the QIP interaction models for cervical cancer and breast cancer screening. Unfortunately, given the low power of the QIP incentives, it is difficult to detect statistically significant effects. The estimated interaction between the QIP and potential bonus is actually negative, although statistically insignificant. However, it is important to recall that potential bonus is exactly linearly related to the number of SH patients; as a result, the coefficient could equally plausibly be interpreted as the impact of increasing the SH population, which does not include women screened for cervical cancer and only overlaps slightly with the population of women screened for breast cancer.

Total enrollment has a small, marginally significant impact on cervical cancer screening in the expected direction, but the denominator is not statistically significant for either measure. None of the quantile measures are statistically significant, although they are all positive (relative to top performers). Finally, we obtain mixed results for the estimated interaction *between* measures. Low performance on cervical cancer combined with breast cancer performance just below its target is associated with a strong negative impact of the QIP on cervical cancer screening and a positive impact on breast cancer screening. Only one of the other pairs is consistent with its prediction (middle/high) although it is not statistically significant. The other two interactions are insignificant and have the same sign for both measures. However, note that because the PacifiCare target is so high and the sample size low, it is difficult to distinguish between the truly top performers and those who still need to maintain their performance to stay above the target.

Because the QIP is such a small program and only observed in isolation for 6 months in our sample period, we also estimate models in which we interact baseline performance with

TABLE 5 Estimates from Interaction Models

		(1) QIP Interactions		(2) QIP + IHA Interactions	
	Hypothesis	Cervical Cancer	Breast Cancer	Cervical Cancer	Breast Cancer
QIP1		−1.549	−1.105	−0.443	−2.318**
		(0.988)	(0.691)	(0.894)	(1.046)
× Potential bonus (in thousands)	Positive	−0.311	−0.396	−0.455**	−0.170
		(0.247)	(0.393)	(0.211)	(0.210)
× Total enrollment (in thousands)	Positive	0.203*	0.225	0.205	0.161*
		(0.109)	(0.143)	(0.127)	(0.088)
× Denominator (in hundreds)	Negative	−0.068	−0.166	−0.089	−0.126
		(0.042)	(0.110)	(0.059)	(0.100)
× Quantile 1	Positive, hump-shaped with max at quantile 3/2	2.072	2.034	1.383*	5.831***
		(1.504)	(1.287)	(0.804)	(2.106)
× Quantile 2		1.802	0.036	0.356	1.126
		(1.521)	(0.810)	(1.414)	(0.952)
× Quantile 3		0.075	−0.402	0.818	0.358
		(1.721)	(0.888)	(0.694)	(0.890)
× Low/Middle	Neg/Pos	−3.325***	2.118*		
		(1.172)	(1.215)		
× Middle/Low	Pos/Neg	0.912	0.658		
		(1.104)	(0.713)		
× Middle/High	Pos/Neg	2.371	−5.919		
		(1.783)	(6.358)		
× High/Middle	Neg/Pos	−2.077	−1.661		
		(1.769)	(1.581)		
IHA1		1.923	1.027	5.500**	0.859
		(1.333)	(0.773)	(2.305)	(2.342)
× Total enrollment (in thousands)	Positive			−0.108	0.035
				(0.184)	(0.079)
× Denominator (in hundreds)	Negative			−0.014	−0.086
				(0.113)	(0.116)
× Quantile 1	Negative, increasing in quantile			−1.106	5.312**
				(3.133)	(2.422)
× Quantile 2				−2.846	−0.509
				(2.894)	(1.803)
× Quantile 3				−1.167	−1.245
				(2.680)	(1.485)
IHA2		6.929***	0.884	13.684***	0.160
		(2.326)	(0.910)	(3.312)	(2.436)
× Total enrollment (in thousands)	Positive			−0.187	0.057
				(0.150)	(0.084)
× Denominator (in hundreds)	Negative			0.013	−0.132
				(0.144)	(0.131)
× Quantile 1	Negative, increasing in quantile			−9.028**	7.887***
				(4.370)	(2.273)
× Quantile 2				−3.146	0.331
				(4.968)	(1.848)
× Quantile 3				−1.963	−1.691
				(3.870)	(1.510)
N		170	151	170	151
T		16	16	16	16

indicators for the IHA regime. Recall, however, that unlike the PacifiCare QIP, participating health plans in the IHA program tended to reward *relative* performance with stratified payments above the 20th or 30th percentile. Because meeting higher thresholds results in higher payments, and top performers must remain in the top quartile (rather than simply meeting a preset target based on *prior* performance as in the QIP), we expect the impact of the IHA to be increasing

in initial performance. That is, the interaction terms (with top quartile as the omitted category) should be negative and increasing in quantile. Panel (2) of Table 5 presents estimates from the models with IHA interactions added. (In this specification, we drop the between-measure interactions and focus only on within-measure impacts.) Unfortunately, we do not find conclusive evidence of any systematic differences in the effect of the IHA by baseline performance. Whereas the quantile-IHA interactions are all negative for cervical cancer screening, with the expected slope for the second year of the IHA, these results are generally not significant. At the same time, we find that being in the lowest quintile for breast cancer is associated with the *greatest* response to the IHA program. Unfortunately, as we cannot net out the cost of quality improvement, it is difficult to attribute these effects solely to the reward structure of either program. In particular, although the theory implies that initial low performers should face higher costs, if costs are not related to prior performance then it is possible that the lowest performers actually face the lowest cost and simply needed some small incentive to pick the low-hanging fruit.

6. Implications and conclusion

■ Our results highlight the fact that pay-for-performance may not necessarily have the dramatic and or even predictable effects touted by its enthusiasts. In the intervention in our study, six health plans combined to pay out more than $122.7 million in additional payments to affiliated providers in 2004, and $139.5 million in 2005, receiving a small and mixed return on their investment.[16] In fact, of the six measures initially rewarded by the IHA, only cervical cancer screening showed consistently positive returns, on the order of 3.5–6 percentage points (a 9%–15% increase). When chlamydia screening was added to the IHA measure set in 2004, it began reversing its decline, relative to the gains found in the Northwest group, when P4P was introduced on other measures 2 years earlier.

On the other hand, appropriate asthma medication rates actually decreased by 2–8 percentage points (2.5%–10%) when P4P was introduced in California, even though it was one of the measures in the common rewarded set and therefore linked to significant potential monetary payouts. Preferred antibiotic usage, which was rewarded by the small-scale QIP but ignored by the larger IHA effort, also declined by roughly 3 percentage points (6%), as did appropriate antibiotic usage (4 percentage points, or 6%), which was ignored by both programs. These declines emphasize the importance of understanding relative rewards when constructing P4P programs. In general, if medical groups can improve some measures by substituting resources away from other measures, then the danger exists that, even if some measures are rewarded by P4P, it may not be enough to offset the gains from substitution toward more lucrative measures, or dimensions of quality.

One take-away lesson from our analysis is that the size of the awards matters. In general, we did not detect movement in the measures until the IHA program went into effect, dramatically increasing the rewards for high performance and broadening the salience of pay-for-performance to medical groups well below the 75th percentile, the point in the distribution targeted by PacifiCare. Of 11 process measures, only 2 showed any response before the IHA came in—appropriate asthma medication and chlamydia screening—which, if anything, went down. This negligible response occurred despite the fact that the IHA initiative, known to be a large-scale program, was just on the horizon, going into effect only months after the QIP. Given that the literature on public reporting has found positive effects on measured quality without any direct financial incentives, these results may seem strange. Indeed, there is little evidence that doctors or patients pay very much attention to report cards (Schneider and Epstein, 1996, 1998). A common

[16] One caveat to our analysis is that we are only using data on medical groups contracting with one health plan, PacifiCare. To the extent that these medical groups are not representative of the average participant in the IHA effort, our results are not generalizable to the IHA population as a whole.

explanation is that public reporting operates through nonfinancial channels such as reputation and/or learning (Kolstad, 2008). Along these lines, one criticism of P4P is that it "commodifies" medical care at the expense of doctor professionalism and intrinsic motivation (see, e.g., Gneezy and Rustichini [2000] for experimental evidence suggesting it may be better in some cases to pay nothing rather than a small performance incentive).

We looked for evidence that measures that shared common production technologies and/or patient population groups responded to P4P in the same way. Although we found some evidence that identification/scheduling may be a driving force in the determination of which measures rise and fall in response to P4P, this evidence was complicated by the fact that we did not uncover the expected positive spillovers to unpaid measures such as chlamydia (before 2004) and diabetic eye exam rates. Part of the problem is that, even if providers do respond to P4P by making information technology improvements, such as automating reminder systems, to increase their performance on rewarded measures, they may not make the natural extension to use these IT improvements to increase performance on other measures, even when the cost is small, if there is no obvious return.

When it comes to disease groupings, this problem should not be as large. For example, if diabetics are more likely to come in for their blood sugar tests or cholesterol checks, then it actually may lower costs to combine the eye exam with these visits. We did not find any significant improvement on hemoglobin A1c testing rates in response to P4P, which may be why we do not see the expected spillover to diabetic eye exams (although these usually cannot be done in the same visit). As for other populations, we did not see any positive spillover from women's health measures to chlamydia screening. (Part of this may be because "women's health" is too broad a measure; chlamydia screening only applies to women ages 16–26, which barely overlaps with the recommended ages for cervical cancer screening and does not overlap at all for breast cancer screening.)

Finally, the most important effect of P4P is its effect on health outcomes. Even if providers show that they are willing to substitute away from unrewarded dimensions of quality toward rewarded dimensions, patients may still be better off if the measures providers are substituting toward are ones which we care about and which are important for clinical outcomes. We examine three measures of bad outcomes in our data set, two where groups are rewarded for reductions by the QIP 1 year into P4P. We find mixed evidence on the effect of P4P on outcomes. Two of the three measures showed significant improvements, including unrewarded asthma-related ER visits, while hospital readmissions increased 1 year after the IHA initiative was introduced in California, relative to trend. Yet both overall readmissions and avoidable hospitalizations are not likely closely related to the measured process indicators that only changed modestly. The mechanisms underlying these changes are not well understood and are hard to link to P4P.

In the end, we fail to find evidence that a large P4P initiative either resulted in major improvement in quality or notable disruption in care. In particular, although some paid measures may have improved in response to the program, we do not find any evidence of positive spillovers to other aspects of care. This result casts doubt on the promise of P4P as a transformative mechanism for improving the general quality of the healthcare system. At the same time, even though we fail to find conclusive evidence of negative spillovers in this analysis, the concern that P4P encourages "teaching to the test" should not be dismissed. Given the complex and largely unobservable nature of healthcare quality, we can only study some potential unintended consequences but we cannot confirm or reject the existence of all such effects. Our results suggest caution in moving ahead with P4P and in interpreting the results of future studies. The negative incentives of P4P programs still exist and should be taken seriously given evidence that providers do indeed respond to incentives.

Appendix

Summary statistics are shown for clinical measures, by region and year are shown.

	Med. Denom.		Mean		Standard Deviation		25th perc.		Median		75th perc.		Maximum	
Measure	CA	NW	CA	NW	CA	NW	CA	NW	CA	NW	CA	NW	CA	NW
Cervical cancer screening														
No.g medical groups	170	25												
2000	1194	1135	30.39	47.66	19.71	17.19	11.90	39.72	27.50	52.86	45.52	63.21	74.52	75.63
2001	1285	1194	37.06	56.19	19.37	11.52	21.35	54.77	38.53	58.74	51.86	63.41	78.17	73.36
2002	1327	889	41.53	60.15	19.60	12.29	27.72	57.14	46.82	63.89	56.62	66.46	77.01	73.96
2003	1166	800	52.06	61.82	15.90	12.84	44.44	58.97	54.32	66.39	64.04	69.57	79.25	74.95
2004	553	304	65.26	68.52	10.65	13.52	58.23	64.14	65.90	73.27	73.66	76.58	84.95	82.61
Breast cancer screening														
No. medical groups	151	24												
2000	272	309	55.47	64.31	19.98	15.15	49.22	56.72	60.82	69.06	70.48	75.31	80.80	83.12
2001	302	322	58.87	69.82	17.39	9.94	51.29	65.50	64.58	72.61	71.67	75.90	80.19	78.88
2002	319	325	61.27	72.00	17.33	6.04	57.40	68.87	65.27	72.44	71.88	75.57	82.31	82.10
2003	319	278	65.77	70.86	11.05	11.88	60.00	69.33	68.33	73.07	73.38	77.01	82.55	82.69
2004	277	241	67.22	70.02	9.52	14.30	62.30	69.06	68.54	73.70	74.08	76.28	92.86	82.64
Hemoglobin A1c testing														
No. medical groups	186	32												
2000														
2001														
2002	264	342	52.07	77.39	30.53	29.19	19.05	84.82	64.17	87.72	77.32	89.87	92.11	93.62
2003	202	319	66.08	88.23	24.50	5.18	63.04	84.57	74.89	88.82	82.22	92.46	92.16	94.53
2004	178	183	69.42	85.39	17.52	6.94	64.37	80.98	73.58	87.31	81.25	90.87	94.83	97.26
Childhood immunization														
No. medical groups	133	18												
2000														
2001	64	37	9.15	4.46	8.88	4.58	1.06	0.00	7.69	4.03	14.85	8.14	47.42	15.38
2002	73	33	13.63	5.02	10.49	5.87	4.49	0.00	12.87	2.91	20.00	8.33	41.96	19.64
2003														
2004														

(*Continued*)

Measure	Med. Denom.		Mean		Standard Deviation		25th perc.		Median		75th perc.		Maximum	
	CA	NW	CA	NW	CA	NW	CA	NW	CA	NW	CA	NW	CA	NW
Appropriate asthma med.														
No. medical groups	77	7												
2000	34	32	76.07	82.76	10.95	5.34	68.42	78.38	77.14	83.33	85.42	84.04	94.12	93.33
2001	37	29	77.86	79.47	9.94	9.22	71.79	67.86	80.00	84.75	84.62	86.21	97.22	87.01
2002	46	32	72.90	74.78	7.56	7.45	68.97	67.69	72.82	76.47	78.87	79.63	87.50	85.00
2003	35	42	72.01	74.97	10.27	6.69	66.67	68.00	73.17	76.92	78.57	77.36	95.00	84.62
2004	79	41	63.76	73.04	6.77	10.03	60.66	65.85	64.86	68.42	68.00	83.84	77.46	88.68
Preferred antibiotic usage														
No. medical groups	145	11												
2000														
2001	277	219	52.78	43.23	12.43	10.46	45.45	36.59	52.41	45.21	60.91	48.28	95.00	60.00
2002	448	283	58.37	48.35	9.72	10.26	51.46	39.54	57.94	46.43	64.66	54.63	86.87	69.41
2003	346	271	48.94	44.61	11.40	9.27	42.16	39.17	48.06	43.69	55.58	50.54	83.83	58.28
2004	323	157	47.31	41.54	10.18	10.48	40.24	32.86	46.09	40.22	53.85	52.59	82.39	55.42
Chlamydia screening														
No. medical groups	127	20												
2000	64	60	13.81	11.80	13.69	9.54	2.53	2.94	9.09	9.41	23.81	21.42	55.56	30.00
2001	92	99	16.20	11.59	13.02	8.65	2.70	3.73	15.91	11.41	25.00	18.24	59.23	27.03
2002	93	68	14.94	15.33	11.34	11.14	5.96	4.88	13.04	16.76	21.15	24.86	50.67	33.33
2003	116	69	20.16	22.84	12.75	12.22	10.34	10.66	17.56	24.87	32.06	33.77	53.85	38.20
2004	112	39	27.59	25.69	12.80	11.91	18.39	17.50	28.00	27.12	36.71	35.25	64.10	42.86
Diabetic eye exam														
No. medical groups	185	29												
2000														
2001	274	318	31.48	45.98	15.49	13.09	21.09	40.25	32.88	47.32	43.05	55.63	62.71	69.16
2002	278	360	32.21	50.07	14.74	14.09	22.98	40.92	32.84	51.72	42.82	59.65	68.05	79.35
2003	213	328	29.13	51.91	14.74	14.07	18.75	45.21	29.90	54.85	39.35	61.54	67.86	70.86
2004														

(*Continued*)

Measure	Med. Denom.		Mean		Standard Deviation		25th perc.		Median		75th perc.		Maximum	
	CA	NW	CA	NW	CA	NW	CA	NW	CA	NW	CA	NW	CA	NW
ACE inhibitor for CHF														
No. medical groups	109	23												
2000	83	116	59.06	36.75	8.18	7.99	53.75	30.30	60.87	35.71	64.41	41.33	74.29	57.14
2001	188	332	51.08	28.90	8.16	7.32	44.58	25.22	51.99	27.44	57.14	30.04	65.82	46.15
2002	162	410	53.82	30.15	7.82	6.96	50.40	22.26	55.29	32.09	58.97	36.71	70.00	39.77
2003	132	275	57.94	34.88	6.57	10.17	53.52	23.85	57.89	35.81	61.36	42.95	74.07	58.33
2004	147	157	55.62	41.51	7.59	6.09	51.47	37.80	54.91	40.88	59.46	46.05	84.85	54.21
Approp. use of antibiotics														
No. medical groups	95	5												
2000														
2001	80	73	64.21	50.91	10.49	12.53	58.06	45.21	66.18	46.30	70.64	58.75	88.31	68.12
2002	235	204	61.20	50.24	8.41	6.44	55.69	47.11	61.92	50.00	66.67	55.90	79.87	56.87
2003	202	176	60.05	51.61	9.45	7.86	53.95	44.62	59.72	50.62	66.15	59.09	86.27	60.25
2004	191	134	57.83	45.96	10.78	8.39	50.00	43.45	59.18	46.03	65.05	50.00	83.94	56.50
Manag. of cholesterol drug														
No. medical groups	179	31												
2000														
2001														
2002	200	191	8.75	6.11	3.58	2.59	6.62	4.86	8.42	5.68	10.48	6.28	19.75	13.41
2003	175	209	8.83	6.06	4.19	2.14	6.25	4.39	8.33	5.63	10.87	7.16	24.19	11.63
2004	91	145	8.54	4.92	4.83	2.66	5.48	3.09	7.98	4.65	10.64	6.67	33.33	10.29

(*Continued*)

Measure	Med. Denom.		Mean		Standard Deviation		25th perc.		Median		75th perc.		Maximum	
	CA	NW	CA	NW	CA	NW	CA	NW	CA	NW	CA	NW	CA	NW
Hospital readmission														
No. medical groups	169	27												
2000	499	588	3.91	3.94	1.65	1.60	3.01	3.09	4.11	3.99	5.02	5.00	8.57	7.45
2001	520	831	4.24	4.50	1.72	1.25	3.32	3.66	4.38	4.35	5.20	4.91	9.09	7.02
2002	460	666	3.98	4.36	1.77	2.09	2.70	2.67	4.12	4.55	5.26	5.48	9.14	8.56
2003	413	551	3.85	4.02	1.82	1.61	2.94	2.74	3.88	4.50	4.82	5.09	13.04	6.57
2004	446	566	4.09	4.26	1.84	1.43	3.08	3.45	4.10	4.66	4.90	5.30	11.76	6.78
Avoidable hospitalization														
No. medical groups	174	27												
2000	497	584	13.45	11.96	4.85	3.88	11.02	10.34	13.67	12.93	16.53	14.74	25.00	19.15
2001	552	844	14.76	13.82	5.24	6.31	11.99	10.78	15.16	12.70	17.88	17.06	29.94	28.42
2002	524	741	11.13	13.26	4.37	4.57	7.92	11.90	11.87	13.93	14.08	16.85	20.76	19.86
2003	448	553	11.57	11.59	5.10	3.88	7.66	9.15	12.19	11.62	15.45	14.31	23.68	18.70
2004	445	588	10.79	11.98	4.80	4.46	7.02	10.02	11.58	12.91	14.24	14.53	24.17	19.13
Asthma-related ER visits														
No. medical groups	163	27												
2000	N/A	N/A	1.32	1.19	1.25	0.99	0.33	0.23	1.01	1.02	2.00	1.97	6.35	3.17
2001			1.65	1.59	1.37	1.04	0.61	0.88	1.44	1.39	2.31	2.45	6.42	3.70
2002			1.83	2.05	1.48	1.75	0.78	0.75	1.61	1.35	2.48	3.55	10.54	5.51
2003			2.07	2.25	1.33	1.30	1.03	1.41	1.88	1.99	2.74	2.87	6.44	6.71
2004			1.80	1.57	1.33	1.18	0.90	0.89	1.59	1.39	2.39	2.09	8.04	5.57

Note: Descriptive statistics are calculated from July performance reports.

References

Campbell, S., Reeves, D., Kontopantelis, E., Middleton, E., Sibbald, B., and Roland, M. "Quality of Primary Care in England with the Introduction of Pay for Performance." *New England Journal of Medicine,* Vol. 357 (2007), pp. 181–190.

Damberg, C.L., Raube, K., Williams, T., and Shortell, S.M. "Paying for Performance: Implementing a Statewide Project in California." *Quality Management in Health Care,* Vol. 14 (2005), pp. 66–79.

Doran, T., Fullwood, C., Gravelle, H., Reeves, D., Kontopantelis, E., Hiroeh, U., and Roland, M. "Pay-for-Performance Programs in Family Practices in the United Kingdom." *New England Journal of Medicine,* Vol. 355 (2006), pp. 375–384.

Dranove, D., Kessler, D., McClellan, M., and Satterthwaite, M. "Is More Information Better? The Effects of Report Cards on Health Care Providers." *Journal of Political Economy,* Vol. 111 (2003), pp. 555–588.

Glied, S. and Zivin, J.G. "How Do Doctors Behave When Some (but Not All) of Their Patients are in Managed Care?" *Journal of Health Economics,* Vol. 21 (2002), pp. 337–353.

Gneezy, U. and Rustichini, A. "Pay Enough or Don't Pay at All." *Quarterly Journal of Economics,* Vol. 115 (2000), pp. 791–810.

Holmstrom, B. and Milgrom, P. "Multi-Task Principal-Agent Problems: Incentive Contracts, Asset Ownership and Job Design." *Journal of Law, Economics, & Organization,* Vol. 7 (1991), pp. 24–52.

Institute of Medicine. *To Err Is Human: Building a Safer Health System*. Washington, DC: National Academy Press, 1999.

——. *Rewarding Provider Performance: Aligning Incentives in Medicare*. Washington, DC: National Academy Press, 2006.

Jacob, B. "Accountability, Incentives and Behavior: Evidence from School Reform in Chicago." *Journal of Public Economics,* Vol. 89 (2005), pp. 761–796.

Kolstad, J. "Information and Quality When Motivation Is Intrinsic: Evidence from Surgeon Report Cards." Mimeo, Harvard University, 2008.

Kruzikas, D., Jiang, H., Remus, D., Barrett, M., Coffey, R., and Andrews, R. "Preventable Hospitalizations. Windows into Primary and Preventative Care, 2000. HCUP Fact Book no. 5." AHRQ Publication no. 04-0056. Rockville, MD: Agency for Healthcare Research and Quality, 2000.

Leonhardt, D. "Why Doctors So Often Get It Wrong." *New York Times*, February 22, 2006.

Liang, K. and Zeger, S. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika,* Vol. 73 (1986), pp. 13–22.

Lu, S.F. "Multitasking, Information Disclosure and Product Quality: Evidence from Nursing Homes." Mimeo, University of Rochester, 2009.

McGlynn, E., Asch, S., Adams, J., Keesey, J., Hicks, J., DeCristofaro, A., and Kerr, E. "The Quality of Health Care Delivered to Adults in the United States." *New England Journal of Medicine,* Vol. 348 (2003), pp. 2635–2645.

Meyer, B. "Natural and Quasi-Experiments in Economics." *Journal of Business and Economic Statistics,* Vol. 13 (1995), pp. 151–161.

Rosenthal, M. and Frank, R. "What Is the Empirical Basis for Paying for Quality in Health Care?" *Medical Care Research and Review,* Vol. 63 (2006), pp. 135–157.

——, ——, Li, Z., and Epstein, A. "Early Experience with Pay-for- Performance: From Concept to Practice." *Journal of the American Medical Association,* Vol. 294 (2005), pp. 1788–1793.

Schneider, E. and Epstein, A. "Influence of Cardiac-Surgery Performance Reports on Referral Practices and Access to Care: A Survey of Cardiovascular Specialists." *New England Journal of Medicine,* Vol. 335 (1996), pp. 251–256.

—— and ——. "Use of Public Performance Reports: A Survey of Patients Undergoing Cardiac Surgery." *Journal of the American Medical Association,* Vol. 279 (1998), pp. 1638–1642.

Shen, Y. "Selection Incentives in a Performance-Based Contracting System." *Health Services Research,* Vol. 38 (2003), pp. 535–552.