HSR Health Services Research

DOI: 10.1111/j.1475-6773.2010.01192.x
RESEARCH ARTICLE

Do Hospitals Alter Patient Care Effort Allocations under Pay-for-Performance?

Lauren Hersch Nicholas, Justin B. Dimick, and Theodore J. Iwashyna

Objective. To determine whether hospitals increase efforts on easy tasks relative to difficult tasks to improve scores under pay-for-performance (P4P) incentives.
Data Source. The Centers for Medicare and Medicaid Services Hospital Compare data from Fiscal Years 2003 through 2005 and 2003 American Hospital Association Annual Survey data.
Study Design. We classified measures of process compliance targeted by the Premier Hospital Quality Incentive Demonstration as easy or difficult to improve based on whether they introduce additional per-patient costs. We compared process compliance on easy and difficult tasks at hospitals eligible for P4P bonus payments relative to hospitals engaged in public reporting using random effects regression models.
Principal Findings. P4P hospitals did not preferentially increase efforts for easy tasks in patients with heart failure or pneumonia, but they did exhibit modestly greater effort on easy tasks for heart attack admissions. There is no systematic evidence that effort was allocated toward easier processes of care and away from more difficult tasks.
Conclusions. Despite perverse P4P incentives to change allocation of efforts across tasks to maximize performance scores at lowest cost, we find little evidence that hospitals respond to P4P incentives as hypothesized. Alternative incentive structures may motivate greater response by targeted hospitals.

Key Words. Medicare, incentives in health care, quality improvement/report cards (interventions)

Payers and policy makers are increasingly turning to pay-for-performance (P4P) and other value-based purchasing strategies in an attempt to control rapidly growing health care costs and improve quality of care. P4P seeks to improve quality by incentivizing hospitals to allocate additional effort toward specific elements of care that are rewarded with bonus payments. However, it is unclear how payers and regulators should design P4P incentives to motivate hospitals to deliver high-quality health care. Fee-for-service payments, including the volume-based payments used in the Medicare program, fail to provide such incentives.

The Centers for Medicare and Medicaid Services (CMS) have been charged with developing value-based purchasing strategies for the Medicare program (CMS 2009).

A recent CMS-sponsored demonstration program, the Premier Hospital Quality Incentive Demonstration (PHQID), was introduced with the goal of significantly improving quality in incentivized areas in response to P4P incentive payments tied to hospital performance measured for five common medical and surgical admissions (CMS 2008). Early evaluations of the Premier demonstration find mixed evidence that the demonstration improved compliance with targeted process of care measures relative to hospital public reporting (without financial incentives) of the same measures (Glickman et al. 2007; Lindenauer et al. 2007). Ryan (2009) finds no improvements in 30-day mortality for Medicare beneficiaries hospitalized with targeted conditions at P4P hospitals. Studies of physician P4P have also identified little response to P4P incentives (Rosenthal et al. 2005; Mullen, Frank, and Rosenthal 2009). Relatively little is known about why P4P strategies fail to meet expectations.

In this paper, we use data from the Premier demonstration to consider a possible explanation for the failure of P4P incentives to motivate improved patient outcomes. We test whether the P4P incentive structure encourages hospitals to maximize the scores used to determine bonus payments by focusing on low-cost, easy-to-improve components of the composite score. We find that P4P hospitals score about 1 percentage point higher than unincentivized hospitals on easy tasks. However, we fail to find consistent evidence that hospitals strategically shift resources to improve scores across three incentivized medical admissions as hypothesized.

BACKGROUND

Premier Demonstration

CMS introduced the PHQID program in October 2003. The demonstration built on a voluntary reporting initiative, the Hospital Quality Alliance (HQA), established by a collaboration between the American Hospital Association

---

Address correspondence to Lauren Hersch Nicholas, Ph.D., M.P.P., Institute for Social Research and Center for Healthcare Outcomes and Policy, University of Michigan, 426 Thompson St, Ann Arbor, MI 48104; e-mail: lnichola@umich.edu. Justin B. Dimick, M.D., M.P.H., is with the Department of Surgery and Center for Healthcare Outcomes and Policy, University of Michigan Medical School, Ann Arbor, MI. Theodore J. Iwashyna, M.D., Ph.D., is with the Division of Pulmonary & Critical Care, University of Michigan Medical School, Ann Arbor, MI.

(AHA), CMS, the Joint Commission on Accreditation of Healthcare Organizations, and several consumer groups (Jha 2005). HQA facilitates public reporting of evidence-based process compliance for three medical classes of admissions (acute myocardial infarction [AMI], heart failure, and pneumonia). Hospital reports are disseminated through the Hospital Compare website (http://www.hospitalcompare.hhs.gov). The Medicare Modernization Act required all hospitals to report to Hospital Compare in October 2004 in order to receive annual payment rate updates.

Hospitals that already subscribed to Premier, a quality reporting and purchasing collective, were invited to participate in this voluntary P4P demonstration program before the introduction of mandatory public reporting. Participants needed at least 30 annual admissions for targeted conditions. Of particular importance from an evaluation perspective, P4P hospitals were already subscribed to a quality reporting service and may be more motivated to improve quality of care than hospitals engaged in reporting only because of the pay-for-reporting efforts. Four hundred and twenty-one hospitals were invited to participate, and 255 completed the 3-year demonstration (Lindenauer et al. 2007).

The Premier demonstration incentivized the three medical admissions targeted by Hospital Compare public reporting and two surgical admissions, coronary artery bypass graft and hip/knee replacement. The P4P demonstration provides bonus payments B for participating hospitals in the top two deciles of a condition-specific composite measure comprised of a subset of observable process and outcomes measures. Outcome measures are included in the two surgical composites and for AMI. Figure 1 details the incentivized measures included in this study, which focuses only on the medical admissions. We analyze process of care measures because hospitals have more certainty about performance on these measures. In an effort to improve hospital quality for all patients, Medicare requires hospitals to report data on all patients treated for the

Figure 1: Classification of Task Difficulty: Public Reporting and Pay-for-Performance Measures for Three Targeted Medical Hospitalizations

Condition	Easy (low per-patient cost)	Difficult (higher per-patient cost)
Acute Myocardial Infarction (AMI)	Aspirin at Discharge Aspirin at Arrival *B*-Blocker at Discharge *B*-Blocker at Arrival	ACE-Inhibitors for Left Ventricular Dysfunction
Heart Failure	Smoking Cessation Counseling Discharge Instructions	Left Ventricular Assessment ACE-Inhibitors for Left Ventricular Dysfunction
Pneumonia	Pneumococcal Vaccination Status Oxygenation Assessment	Blood Culture before Antibiotics Given Timing of Initial Antibiotic

targeted conditions and uses this all-patient data to rank hospitals. However, Medicare pays bonus payments only for Medicare-covered admissions. Hospitals are ranked and paid bonuses separately for each targeted condition.

PQHID hospital quality scores are calculated using a two-stage process for conditions with process and outcome measures. The process component score uses an opportunity model reflecting the number of successfully completed tasks divided by the number of patients eligible for each measure. The outcome component is calculated similarly. The full composite score is a simple average of the process and outcome scores, weighted by the relative number of measures included in each. For example, the full AMI composite score includes eight process measures and one outcome. Thus, the total composite score is $(8/9) \times$ (Total Process Successes/Total Process Opportunities)$+(1/9) \times$ (Total Outcome Successes/Total Outcome Opportunities).

Composite scores are calculated without adjustments for the difficulty or potential impact on patient health of component measures. Although a hospital would have to have more resources to ensure that a left ventricular assessment was performed on an eligible heart failure patient, each of those encounters would contribute one success and one opportunity to the process composite.

Hospitals in the highest decile of composite score for each condition receive an annual bonus of 2 percent of the Medicare diagnosis-related group (DRG) payment for Medicare-covered admissions with the incentivized condition. Hospitals in the second-highest performance decile receive a 1 percent bonus. Hospitals that fail to improve above the lowest quintile of initial performance by the end of the third year face fines of 2 percent of DRG payments.

Hospital Response to P4P Incentives

Payers anticipate that P4P incentives will alter hospital behavior. Because the majority of hospitals eligible for the P4P payments are nonprofit, we follow Horwitz and Nichols (2007) and conceptualize the hospital's problem as choosing a level of quality that maximizes an objective function containing quality and other priorities such as total service volume and revenue. Chosen levels of quality vary across conditions. We assume that hospitals already engage in quality improvement efforts that generate positive return on investment (across monetary or nonmonetary elements of the objective function). P4P offers additional incentives (payments or fines) to indirectly motivate a higher level of quality than hospitals would otherwise select. This is achieved through process compliance and, for some conditions, inclusion of patient outcomes in the composite score.

For the P4P bonuses to motivate changes in hospital process compliance and outcomes, expected bonuses must outweigh the opportunity cost of improvement to the hospital. Hospitals can change performance on two dimensions for each measure: the number of eligible patients and the number of successes. In practice, admissions patterns for the medical conditions targeted by P4P and public reporting would be difficult to manipulate.[1] We posit that performance pay will motivate greater effort on tasks that can be completed at lowest effort and monetary cost to maximize net benefit from the bonus. In the case where all patients are eligible for all measures, a hospital's expected bonus payment would be the same for an increase in the percentage of AMI patients receiving smoking cessation counseling (which can be accomplished by distributing an antismoking booklet during patient registration) or a same-sized improvement in inpatient survival, which may require changes on multiple tasks and utilize additional resources.

To illustrate the trade-offs across easy and difficult tasks, let N_A be the number of patients eligible for measure A, S_A be the number of successes on measure A, and C_A be the per-patient cost of achieving a success in measure A. For any measure, the hospital faces the choice of resource allocation to achieve a number of successes. If the goal is to maximize the P4P score while minimizing cost, we would expect substitution from high cost of success activities to low cost of success activities. Consider the case where the hospital is choosing between allocating enough resources to obtain S_A successes of S'_A successes. The change in the composite score for a change in the number of successes is $(S'_A - S_A)/N_A$, because the number of eligible patients is the same. The cost of those additional successes is $C_A \times (S'_A - S_A)$. Thus, the cost for a 1 percentage point improvement in one's composite score by doing better on measure A is $\{C_A \times (S'_A - S_A)\}/\{(S'_A - S_A)/N_A\}$, which simplifies to $N_A \times C_A$. The key decision-making variable is the ratio of the costs across alternative measures of making equivalent changes in score. Whenever $(N_A \times C_A)/(N_B \times C_B) < 1$, we expect hospitals to substitute toward A away from B if they are solely seeking to maximize their P4P benefit, as the improvement in the score per expenditure is greater under A than under B.

However, not all hospitals face the same incentives to substitute. Because bonus payments and fines are much more relevant for hospitals at the tails of the initial performance demonstration, response to P4P should be concentrated among initially high- and low-performing hospitals. We also note that all hospitals face incentive to improve process compliance scores during the study period due to the CMS public reporting requirements that also began at the onset of P4P. Hospitals may gain or lose volume if patients and payers

respond to posted quality information. Thus, the relevant P4P effect is improvement above and beyond secular trends related to public reporting, driven either by improvements in process of care or improvements in record keeping. Reduced effort on costly tasks may have adverse consequences for the hospital in ways that are not directly related to P4P bonus payments such as diminished reputation. P4P hospitals may balance multiple incentives by concentrating improvements among low-cost processes.

Data and Methods

Data

This study uses Hospital Compare measures collected under the CMS Reporting Hospital Quality Data for Annual Payment Update initiative and cover Fiscal Years 2003–2005 (reported in 2004–2006). Hospital Compare measures are posted with a 9-month lag. Data are available for 243 Premier hospitals and 3,100 non-Premier hospitals. Because public reporting and P4P begin simultaneously, pre-P4P performance data are unavailable. We use Hospital Compare data to assess the effect of P4P on hospital process compliance relative to public-reporting only.

Sample restrictions described below are used to reduce the likelihood that our estimated P4P effects are driven by unobserved differences between P4P and comparison hospitals. We first limit the sample to hospitals reporting to Hospital Compare in all 3 years of the P4P demonstration. We include hospitals with at least 30 admissions for each of the incentivized medical conditions in all 3 years of data. Critical access hospitals, which tend to be small hospitals in rural areas and receive cost-based reimbursement from Medicare, are excluded from the sample. These restrictions generate an analytic sample with 145 P4P hospitals (the treatment group) and 1,089 comparison hospitals. We exclude 98 small P4P hospitals that do not have sufficient sample size in all years.

We augment the Hospital Compare data with survey data from the 2003 AHA Annual Survey. AHA survey data include baseline hospital characteristics, including teaching status and the ratio of registered nurses to admissions, which may reflect hospitals' interests and abilities to comply with evidence-based measures. We also control for the percentage of admissions covered by Medicare, because P4P bonuses will be larger for hospitals which are more reliant on Medicare.

The AHA survey also asks hospitals whether they are involved in quality reporting or improvement efforts as of the 2003 survey. Absent preintervention

compliance data, this variable helps to isolate a control group that is engaged in some form of quality measurement at baseline. Three-quarters of P4P and non-P4P hospitals report participating in quality reporting in 2003. Our preferred control group for the 145 P4P hospitals is the 842 "early adopter" hospitals that are already engaged in some form of reporting as of the 2003 survey. By comparing P4P hospitals to early adopters, we minimize bias related to differential knowledge of or engagement in quality measurement and improvement between P4P and comparison hospitals at baseline.

Our analysis is limited to 13 Hospital Compare measures covering the three initial conditions, which are consistently reported during the initial years of P4P. Process-of-care measures are the proportion of eligible patients receiving each recommended treatment, and they range between 0 and 100. We calculate overall composite scores for each condition and condition-specific composites for easy and hard processes. Composite scores are calculated following PQHID methodology as an opportunity model, which is the proportion of opportunities where the appropriate measure was provided.

Methods

A panel of physician health services researchers, including a cardiologist, critical care physician, and a surgeon, classified hospital efforts on incentivized tasks as easy or difficult to improve (Figure 1). Panelists were instructed to classify tasks that would impose minimal additional per-patient costs as easy to improve and those that would impose additional costs, for example, by requiring additional staff time (either from existing staff or new hiring) as difficult to improve. Hospital Compare data are used to create composite hospital performance scores separately for each of the three conditions and for easy and difficult tasks within conditions. Hospitals are assigned quintiles of initial performance based on where their process compliance composite score falls in the P4P hospital process compliance distribution in Year 1.

We estimate random effects regressions of hospital process compliance with easy and difficult tasks during the first 3 years of P4P using generalized least squares regression.

$$P_{\mathrm{hte}} = \alpha \mathrm{P4P_h} + \beta Q\mathrm{h} + \gamma(\mathrm{P4P_h} \times Q_{\mathrm{h}}) + \tau H_{\mathrm{h}} + \rho_{\mathrm{T}} + \delta \mathrm{P4P_h} \times T + a_{\mathrm{h}} + \varepsilon_{\mathrm{hte}} \quad (1)$$

The dependent variable P_{hte} is the in average process compliance in hospital h in year t on easy (difficult) tasks e. P4P is an indicator for participation in the

P4P demonstration; Q_h is a vector of dummy variables indicating the hospital's condition-specific initial performance quintile relative to the omitted median quintile; $P4P_h \times Q_h$ is a vector of interaction terms which allow the P4P response to vary with initial hospital ranking; H is a vector of baseline hospital characteristics from the AHA survey; T is a vector of year fixed effects relative to the first year; $\delta P4P_h \times T$ is an interaction term which allows the time effect to vary for P4P and non-P4P hospitals; a_h is a hospital random effect uncorrelated with other variables; and ε_{hte} is an error term. We test the hypothesis that P4P incentives motivate hospitals to increase efforts on easy tasks and decrease efforts on difficult tasks. We examine whether this response is concentrated among hospitals that are more likely to receive bonus payments or face larger potential bonuses.

Equation (1) is estimated twice for each of the three incentivized conditions. Within each condition, the model is estimated separately for the easy and difficult composites scores. Hospital rankings Q_h are condition specific. Our preferred specification compares P4P hospitals only to those hospitals which were already engaged in quality reporting in 2003. In addition to testing for heterogeneous response to P4P incentives by initial level of performance, we estimate a second set of regressions that considers the effect of hospital size. The volume regressions include indicator variables for hospitals in the lowest and highest quartiles of condition-specific volume.

RESULTS

Compliance improved for all reported performance measures between Year 1 and Year 3 of the P4P demonstration project both in hospitals receiving financial incentives and in other hospitals that were only subject to public reporting. P4P hospitals experience larger unadjusted gains on some but not all targeted measures (Table 1).

As shown in Table 1, P4P and reporting-only hospitals increased performance across both easy and difficult measures. Overall gains are nearly identical for P4P and early adopter non-P4P hospitals for AMI (3.5 percentage points versus 3.2). Early adopter hospitals actually make larger gains in use of ACE-inhibitors for left ventricular systolic dysfunction (LVSD), which is classified as a difficult task. P4P hospitals do exhibit larger gains in composite scores for both heart failure (7.8 versus 6.8) and pneumonia (11.5 versus 10.1 percentage points) relative to the early adopter non-P4P hospitals. In contrast to the predicted behavior for P4P hospitals to reduce efforts on difficult tasks,

Table 1: Unadjusted Average Process Compliance by Pay-for-Performance (P4P) and Public-Reporting-Only Hospitals, Fiscal Years 2003–2005

	P4P Hospitals			*Reporting Early Adopters*			*Public-Reporting Only*		
	Year 1	*Year 3*	*Gain*	*Year 1*	*Year 3*	*Gain*	*Year 1*	*Year 3*	*Gain*
AMI easy composite	93.3	96.5	3.2	92.5	95.5	3.0	91.9	95.3	3.4
	(4.8)	(3.0)		(6.1)	(3.9)		(6.7)	(4.1)	
Aspirin at arrival (E)	95.0	97.2	2.2	94.7	96.5	1.8	94.5	96.4	1.9
	(4.3)	(2.9)		(5.1)	(3.4)		(5.1)	(3.4)	
Aspirin at discharge (E)	94.1	96.8	2.6	92.9	95.5	2.6	92.2	95.2	3.0
	(6.6)	(4.3)		(8.1)	(5.2)		(9.1)	(5.8)	
β-Blockers at arrival (E)	79.6	88.5	8.8	89.9	93.6	3.7	78.2	85.3	7.1
	(14.2)	(10.7)		(9.2)	(5.5)		(17.9)	(11.6)	
β-Blockers at discharge (E)	92.5	96.8	4.2	91.2	95.6	4.5	90.2	95.2	5.0
	(6.8)	(3.7)		(8.8)	(4.8)		(9.6)	(5.3)	
ACE-inhibitors for LVSD (difficult measure)	90.7	94.9	4.2	79.2	86.0	6.7	89.1	93.4	4.3
	(7.2)	(4.3)		(16.9)	(11.2)		(9.8)	(5.6)	
Heart failure (HF) easy composite	59.1	78.3	19.2	53.3	72.3	19.0	51.7	71.4	19.7
	(24.8)	(13.8)		(25.4)	(17.5)		(26.3)	(17.8)	
Discharge instructions (E)	56.0	74.5	18.5	49.7	68.4	18.8	47.9	67.4	19.5
	(27.0)	(16.3)		(26.9)	(19.9)		(28.0)	(20.2)	
Smoking cessation advice (E)	75.7	95.8	20.1	69.1	90.5	21.3	68.1	90.1	22.0
	(20.8)	(6.7)		(26.3)	(11.9)		(26.8)	(12.1)	
HF difficult composite	86.2	93.5	7.3	85.0	91.6		84.3	91.1	6.8
	(7.4)	(4.8)		(9.3)	(5.4)		(9.5)	(5.9)	
ACE-inhibitors for LVSD	89.0	95.5	6.4	88.1	94.0	6.0	87.3	93.5	6.2
	(7.7)	(4.6)		(9.7)	(5.4)		(10.0)	(6.1)	
Left ventricular assessment	78.5	88.2	9.8	76.2	85.5	9.4	75.8	85.1	9.3

continued

Table 1. *Continued*

	P4P Hospitals			*Reporting Early Adopters*			*Public-Reporting Only*		
	Year 1	*Year 3*	*Gain*	*Year 1*	*Year 3*	*Gain*	*Year 1*	*Year 3*	*Gain*
	(12.2)	(8.3)		(14.0)	(8.6)		(14.2)	(9.1)	
Pneumonia easy composite	98.9	99.8	0.9	98.8	99.7	0.9	98.7	99.7	1.0
	(2.1)	(0.5)		(2.7)	(0.7)		(2.6)	(0.8)	
Pneumonia vaccination status (E)	51.6	80.9	29.3	43.3	71.9	28.6	42.1	70.6	28.5
	(24.3)	(13.0)		(25.8)	(18.0)		(25.5)	(18.8)	
Oxygenation assessment (E)	98.9	99.8	0.9	98.8	99.7	0.9	98.7	99.7	0.9
	(2.1)	(0.5)		(2.7)	(0.7)		(2.6)	(0.8)	
Pneumonia difficult composite	66.2	83.8	17.6	63.7	79.2	15.5	63.2	78.6	15.4
	(10.3)	(6.9)		(11.3)	(9.2)		(11.3)	(9.4)	
Blood culture preantibiotics	83.4	92.1	8.6	82.3	89.9	7.6	82.3	89.7	7.4
	(7.8)	(5.4)		(9.5)	(6.4)		(9.8)	(6.5)	
Antibiotic timing	68.7	80.3	11.6	68.5	77.4	8.8	68.3	77.0	8.6
	(9.5)	(7.6)		(11.6)	(10.3)		(11.6)	(10.3)	
% Medicare patient days	0.51			0.50			0.50		
	(0.11)			(0.13)			(0.13)		
RN:patient ratio	0.02			0.02			0.02		
	(0.01)			(0.01)			(0.01)		
Quality reporting	0.78						0.76		
	(0.41)						(0.43)		
Teaching hospital	0.38			0.28			0.27		
	(0.49)			(0.45)			(0.45)		
Observations	145			842			1,089		

Note. Standard deviations in parentheses. (E) indicates easy tasks. Summary statistics of process compliance for P4P and public-reporting-only hospitals reporting on at least 30 patients with targeted admissions annually. Public-reporting column includes early adopters.

AMI, acute myocardial infarction; LVSD, left ventricular systolic dysfunction; RN, registered nurse.

incentivized hospitals make larger gains on hard tasks for both heart failure and pneumonia than comparison hospitals do.

Table 2 reports regression results from the first set of random effects regressions comparing P4P hospitals to public reporting early adopters. P4P hospitals score higher on easy tasks than control hospitals for AMI ($\alpha = 0.93$ percentage points, SE = 0.36) and heart failure ($\alpha = 3.12$, SE = 2.68), and pneumonia ($\alpha = 0.05$, SE = 0.21), though only the AMI effect is statistically significant. The differences between P4P and control hospitals for difficult tasks are small and insignificant. The P4P coefficient for heart failure is negative ($\alpha = -0.44$, SE = 0.90) as expected, but positive for heart attack ($\alpha = 0.44$, SE = 1.48) and pneumonia ($\alpha = 1.04$, SE = 0.72). The $P4P_h \times$ time effects are positive and statistically significant for the hard pneumonia composites, indicating that P4P hospitals improve more rapidly on difficult tasks than unincentivized hospitals, contrary to our expectations.

The regression evidence confirms our observation from the descriptive statistics; hospitals generally did not respond to P4P incentives as expected. Point estimates are small in magnitude; for example, the 0.93 percentage point increase in the easy AMI composite represents about 1 percent of the Year 1 mean score.

The P4P incentives in PHQID are most relevant for high and low performers. Contrary to our expectations, we fail to find statistically significant effects for P4P hospitals at either end of the initial quality distribution relative to hospitals with average scores. In sensitivity analysis, we fail to observe significant P4P effects in models estimated separately by quintile. We compare P4P hospitals to early-adopter hospitals because we are concerned that other unobserved hospital characteristics, such as motivation to improve and prior improvement activity, correlate with P4P status and generate biased estimates. Early adopter public reporting hospitals have somewhat higher initial composite quality scores for all three incentivized conditions. In sensitivity analysis, models are estimated using the full public-reporting sample as a control group (Table SA1). Our results are essentially unchanged, though P4P coefficients are slightly larger in magnitude.

P4P incentives may be more salient for larger hospitals that are eligible for larger potential bonus payments. Table 3 reports regression results controlling for hospital volume and a volume $\times P4P_h$ interaction. We first omit the initial performance quintiles, which were insignificant determinants of process score in the first set of regressions. P4P main effects are positive and statistically significant for both AMI and Heart Failure easy tasks ($\alpha_{AMI} = 1.24$, SE = 0.43; $\alpha_{HF} = 5.2$, SE = 2.52). While the P4P effect remains small and statistically insignificant for the easy pneumonia composite, P4P hospitals

Table 2: Pay-for-Performance (P4P) Participation, Initial Performance and Hospital Process Compliance, Early Reporters Only: Fiscal Years 2003–2005

	Heart Attack		*Heart Failure*		*Pneumonia*	
	Easy	*Hard*	*Easy*	*Hard*	*Easy*	*Hard*
P4P	0.93*	0.44	3.12	−0.44	0.05	1.04
	(0.36)	(1.48)	(2.68)	(0.90)	(0.21)	(0.72)
Quintile 1	−5.64**	−7.93**	−20.37**	−7.15**	−0.61**	−10.25**
	(0.35)	(1.00)	(1.36)	(0.49)	(0.17)	(0.60)
Quintile 2	−1.65**	−2.78**	−9.76**	−1.85**	−0.18	−4.48**
	(0.26)	(0.91)	(1.29)	(0.37)	(0.12)	(0.53)
Quintile 4	1.72**	2.51**	5.15**	2.01**	0.21*	4.11**
	(0.19)	(0.86)	(1.46)	(0.41)	(0.09)	(0.49)
Quintile 5	2.96**	7.07**	14.11**	4.69**	0.27**	10.32**
	(0.20)	(0.80)	(1.37)	(0.36)	(0.09)	(0.53)
P4P × Quintile 1	0.04	−2.07	2.82	2.4	0.31	1.04
	(0.78)	(2.46)	(3.62)	(1.30)	(0.27)	(1.47)
P4P × Quintile 2	−1.09	−0.01	−1.6	1.52	−0.2	0.73
	(0.62)	(2.01)	(3.69)	(1.12)	(0.32)	(1.09)
P4P × Quintile 4	−0.89*	−0.53	0.42	−0.32	−0.12	−1.1
	(0.41)	(1.58)	(3.26)	(1.01)	(0.18)	(1.08)
P4P × Quintile 5	−0.45	0.25	2.33	0.74	−0.06	−1.14
	(0.39)	(1.53)	(3.09)	(0.98)	(0.18)	(0.99)
Year 2	1.69**	4.20**	8.31**	4.43**	0.64**	8.46**
	(0.15)	(0.61)	(0.67)	(0.26)	(0.08)	(0.28)
Year 3	3.02**	6.73**	19.03**	6.66**	0.95**	15.46**
	(0.19)	(0.65)	(0.85)	(0.30)	(0.09)	(0.34)
P4P × Year 2	−0.1	0.55	0.84	0.35	0.1	2.24**
	(0.33)	(1.45)	(1.69)	(0.56)	(0.18)	(0.66)
P4P × Year 3	0.21	2.11	0.07	0.59	−0.04	2.21**
	(0.40)	(1.51)	(2.13)	(0.62)	(0.19)	(0.81)
Teaching hospital	0.69**	−0.45	−2.70**	1.19**	0.26**	−0.1
	(0.16)	(0.55)	(0.98)	(0.27)	(0.08)	(0.35)
% Admissions Medicare	0.0002	−0.01	0.13**	−0.02	−0.006	0.10**
	(0.007)	(0.02)	(0.04)	(0.01)	(0.003)	(0.01)
RN per admissions	0.17	0.15	0.38	0.30	0.04	−0.21
	(0.09)	(0.33)	(0.59)	(0.16)	(0.04)	(0.20)
R^2	0.49	0.08	0.42	0.32	0.09	0.69

Notes. Robust standard errors in parentheses. 987 hospitals engaged in some form of voluntary public reporting prior in 2003.

*Significant at 5%.

**Significant at 1%.

RN, registered nurses.

Table 3: Pay-for-Performance (P4P) Participation, Volume and Hospital Process Compliance among Early Reporters, Fiscal Year 2003–2005

	Heart Attack		*Heart Failure*		*Pneumonia*		*Heart Attack*		*Heart Failure*		*Pneumonia*	
	Easy	*Hard*	*Easy*	*Hard*	*Easy*	*Hard*	*Easy*	*Hard*	*Easy*	*Hard*	*Easy*	*Hard*
P4P	1.24**	0.76	5.20*	0.99	0.19	2.20*	1.00**	0.24	2.57	−0.4	0.23	1.23
	(0.43)	(1.36)	(2.52)	(0.74)	(0.16)	(0.99)	(0.34)	(1.49)	(3.08)	(0.92)	(0.20)	(0.79)
Year 2	1.30**	3.74**	8.04**	4.23**	0.62**	8.73**	1.32**	4.02**	8.33**	4.33**	0.61**	8.61**
	(0.17)	(0.62)	(0.84)	(0.27)	(0.09)	(0.30)	(0.17)	(0.62)	(0.80)	(0.27)	(0.09)	(0.29)
Year 3	2.67**	6.32**	18.83**	6.48**	0.93**	15.66**	2.68**	6.55**	19.04**	6.57**	0.92**	15.56**
	(0.19)	(0.65)	(0.96)	(0.30)	(0.10)	(0.35)	(0.20)	(0.66)	(0.93)	(0.30)	(0.09)	(0.35)
P4P × Year 2	−0.04	0.89	3.64	0.36	0.06	2.15**	0.05	1.02	2.27	0.16	0.04	2.01**
	(0.37)	(1.56)	(2.09)	(0.65)	(0.17)	(0.68)	(0.37)	(1.52)	(1.94)	(0.64)	(0.16)	(0.69)
P4P × Year 3	0.23	2.39	2.47	0.6	−0.08	2.19**	0.32	2.54	1.34	0.42	−0.09	2.05*
	(0.42)	(1.57)	(2.41)	(0.70)	(0.18)	(0.83)	(0.42)	(1.55)	(2.32)	(0.68)	(0.17)	(0.83)
Teaching hospital	1.31**	0.89	−0.11	1.97**	0.20*	−1.37*	0.44**	−0.53	−2.53*	1.12**	0.25**	−0.09
	(0.25)	(0.65)	(1.33)	(0.38)	(0.08)	(0.61)	(0.15)	(0.56)	(1.01)	(0.28)	(0.08)	(0.35)
% Medicare admits	0.002	−0.05*	0.15**	0.006	−0.001	0.17**	0.003	0.03	0.14**	0.004	−0.007*	0.10**
	(0.01)	(0.02)	(0.02)	(0.007)	(0.002)	(0.01)	(0.004)	(0.02)	(0.02)	(0.01)	(0.002)	(0.01)
RN per admission	0.35*	0.53	1.01	0.52*	0.05	0.05	0.04	0.11	0.40	0.29	0.04	−0.22
	(0.15)	(0.39)	(0.82)	(0.21)	(0.04)	(0.37)	(0.09)	(0.33)	(0.59)	(0.16)	(0.04)	(0.20)
Low volume quarterly	−0.97**	−1.09	−0.46	−0.31	−0.07	0.8	−1.21**	−0.78	−0.01	−0.17	−0.1	0.47
	(0.30)	(0.93)	(1.17)	(0.44)	(0.09)	(0.42)	(0.24)	(0.84)	(1.03)	(0.39)	(0.09)	(0.34)
High volume quarterly	0.94**	1.14*	0.35	0.22	0.06	0.000	0.70**	0.21	−0.06	0.11	0.07	0.38
	(0.18)	(0.56)	(0.89)	(0.30)	(0.07)	(1.06)	(0.18)	(0.55)	(0.90)	(0.29)	(0.08)	(1.14)

continued

Table 3. *Continued*

	Heart Attack		*Heart Failure*		*Pneumonia*		*Heart Attack*		*Heart Failure*		*Pneumonia*	
	Easy	*Hard*	*Easy*	*Hard*	*Easy*	*Hard*	*Easy*	*Hard*	*Easy*	*Hard*	*Easy*	*Hard*
P4P × low volume	−1.56*	−0.89	3.94	0	−0.35	1.12	−0.78	1.29	2.38	−0.29	−0.44	−0.22
	(0.73)	(2.23)	(3.11)	(1.09)	(0.27)	(1.13)	(0.55)	(1.90)	(2.62)	(0.97)	(0.29)	(0.79)
P4P × high volume	−0.76*	−1.51	−3.76	0.01	−0.02	2.14	−0.66	−1.02	−1.75	0.29	0.07	1.9
	(0.37)	(1.31)	(2.40)	(0.66)	(0.19)	(1.60)	(0.35)	(1.22)	(2.35)	(0.66)	(0.21)	(1.70)
Initial performance quintile and interaction	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.15	0.06	0.14	0.16	0.06	0.35	0.51	0.2	0.41	0.42	0.09	0.69

Note. Robust standard errors in parentheses. 987 hospitals engaged in some form of voluntary public reporting prior in 2003.

*Significant at 5%.

**Significant at 1%.

RN, registered nurses.

exhibit significantly higher performance on the difficult pneumonia tasks ($\alpha_{PN} = 2.20$, SE = 0.99).

Magnitudes of the P4P point estimates are reduced when we reintroduce initial performance quintiles Q_h and $P4P_h \times Q_h$. Only the AMI easy P4P effect remains statistically significant, indicating a 1 percentage point higher process compliance score among P4P hospitals relative to public reporting only. P4P hospitals also improve compliance with hard measures of pneumonia care by an additional 2 percentage points in each of the second 2 years of the demonstration, the only significant difference in performance over time between P4P and comparison hospitals. The full P4P effect for the heart failure easy and pneumonia hard measures, including all interaction terms, is also both statistically insignificant and inconsistently signed for most combinations of hospital size, year, and initial performance quintile.

We sought additional evidence as to whether hospitals strategically substitute toward easy tasks in order to improve their scores. In Table 4, we examine the distribution of relative numbers of eligible patients across measures to understand the potential for effort substitution across targeted tasks. Hospitals have, on average, 5.9 times as many patients eligible for an aspirin at admission for AMI (an easy measure) as are eligible for an ACE-inhibitor among those with LVSD (difficult measure). This implies that if the average hospital faces marginal costs to provide an ACE-inhibitor for those with LVSD that are >5.9

Table 4: Task Substitution Ratios of Difficult versus Easy Tasks

Difficult	*Easy*	*Mean*	*Median*	*25th Percentile*	*75th Percentile*	*75/25 Ratio*
ACE-I for LVSD	Aspirin at admission	5.9	4.5	3.0	7.1	2.4
ACE-I for LVSD	Aspirin at discharge	5.0	4.4	3.4	5.8	1.7
ACE-I for LVSD	β-Blocker at admission	5.0	3.8	2.4	5.9	2.5
ACE-I for LVSD	β-Blocker at discharge	5.1	4.5	3.6	5.8	1.6
Left ventricular assessment	Smoking cessation	0.1	0.1	0.1	0.2	2.2
Left ventricular assessment	Discharge instructions	0.7	0.8	0.4	0.8	1.9
ACE-I for LVSD	Smoking cessation	0.4	0.4	0.3	0.6	2.2
ACE-I for LVSD	Discharge instructions	2.2	2.1	1.4	2.7	1.9
Blood cultures	Vaccination status	1.0	1.1	1.5	0.8	0.5
Initial antibiotics	Vaccination status	0.6	0.7	0.8	0.6	0.7
Blood cultures	Oxygen assessment	1.8	1.8	2.7	1.3	0.5
Initial antibiotics	Oxygen assessment	1.1	1.2	1.2	1.0	0.8

LVSD, left ventricular systolic dysfunction.

times the marginal costs of aspirin at admission, they should substitute efforts from the hard to the easy measure in order to maximize the P4P composite. It is implausible that the marginal cost ratio is not >5.9 for the average hospital in practice, but substitution is not observed to have occurred. For some task pairs, the easy:difficult ratio is <1. Unless the harder task was substantially cheaper (at the margin) than the easy task, we would expect score-maximizing hospitals to have fully substituted toward the easier task by Year 3.

In regression analyses, we confirm that hospitals which face a lower marginal cost ratio for substitution (and therefore greater incentives to substitute) were not more likely to substitute toward easier tasks under P4P. We estimate our comprehensive specification of equation (1) including the full set of initial performance, year, and volume P4P interactions separately for each of the incentivized tasks (Table SA2). Among individual measures, the P4P main effect is statistically significant only for two of the easy AMI measures (aspirin at arrival and discharge) and one of the easy pneumonia measures (vaccination status). While hospitals in the highest quintile of performance score do not differentially respond to P4P incentives, hospitals in the lowest performance quintile for heart failure care exhibit higher scores for one easy (smoking cessation counseling, $\alpha = 6.8$ percentage points, SE = 2.7) and one difficult measure (left ventricular assessment, $\alpha = 3.0$ percentage points, SE = 1.46).

We conducted additional sensitivity analyses to confirm our results. Our main findings—that P4P is associated with a 1 percentage point gain in compliance for easy AMI tasks but not related to performance on heart failure or pneumonia measures—are robust across multiple specifications. Findings persist when we reestimate equation (1) using the natural logarithm of compliance score as the dependent variable and in a seemingly unrelated regression model with the change in score as the dependent variable, which allows the error terms to correlate across conditions.

DISCUSSION

Despite limited empirical evidence of its effectiveness, public and private payers continue to view P4P as a promising vehicle for quality improvement and cost savings (Petersen 2006; IOM 2007). If P4P strategies are to achieve these goals, however, they must motivate hospitals to respond in the desired manner. To aid understanding of the apparent failure of P4P programs to motivate changes in health care quality, we tested whether hospitals rationally responded to incentives created by the PHQID. Despite incentives to game

the system and boost scores at low cost, we found that hospitals display no consistent shift in efforts to easier tasks.

Previous studies evaluating P4P yield mixed results. In a national study using clinical registry data, Glickman et al. (2007) find an improvement on some processes of care for AMI but no significant impact on a composite of all processes or on risk-adjusted mortality. Perhaps the most comprehensive study evaluating the Premier P4P program was conducted by Ryan (2009) using national Medicare data. Ryan demonstrated no impact of P4P on risk-adjusted mortality and 90-day Medicare payments for all five incentivized conditions. Our study extends prior work to show that when there is a response, the efforts are concentrated among easy tasks. P4P incentives do not appear to motivate hospitals to improve on difficult tasks. Importantly, we also demonstrate that the improvement in easy tasks does not come at the expense of decreased effort on the more difficult tasks.

There are several limitations to our analysis. Because preperiod data are unavailable, we may underestimate the total P4P effect on effort allocation. Although we control for many observable hospital characteristics, participation in the Premier demonstration was nonrandom and other unobserved factors may simultaneously influence P4P participation and task allocation. We only observe a subset of incentivized tasks, so we are unable to assess, for example, whether P4P hospitals allocated more or less effort to distributing smoking cessation brochures to AMI patients (easy) or ensuring that they received thrombolytics within 30 minutes of arrival.

We lack comprehensive information about whether hospitals are participating in other P4P or public reporting efforts during the study period. While it is likely that some hospitals are also involved in programs run by local payers, including state Medicaid agencies, these programs are unlikely to alter our results because they tend to be small in scope. The current literature lacks examples of P4P programs that led to meaningful differences in hospital performance, so it is unlikely that our results are driven by other programs.

Our results have important implications for payers and policy makers considering ways to expand the role of P4P in reimbursement. We note that nonresponse to P4P incentives is the optimal response for many hospitals when incentives are based on relative performance rankings. Hospitals with average composite scores are unlikely to qualify for bonus payments, so there is no expected return on investments in improved process compliance. We cannot rule out the possibility that hospitals do not change effort allocations to maximize bonus scores because changes in efforts in incentivized tasks would adversely affect overall quality of care or another hospital objective would

suffer. Other incentives and market factors such as reputation and private payer expectations likely balance out explicit incentives for gaming introduced by P4P.

Policy makers should consider the relative difficulty of incentivized tasks when designing P4P programs. The heavy representation of process compliance measures in the PHQID scoring methodology provides incentives for hospitals to improve their scores by devoting additional efforts to easy tasks. This danger would characterize any composite score methodology that does not adjust for task difficulty or relative return. This is especially important when composite quality measures include both processes and outcomes. A rational hospital would improve its scores by increasing compliance on easy processes of care, rather than focusing on improving outcomes, which many agree is the "gold standard" for documenting the effectiveness of quality improvement. Our results highlight the need to consider the difficulty of tasks when creating performance measures.

Policy makers also need to consider the size of the incentives in P4P. Hospitals may not have responded to the P4P incentives to improve their scores by focusing on easy tasks because the bonus payments are small. Two percent of DRG payments, the maximum bonus, is between U.S.$300 and U.S.$500 for Medicare patients. Uncertainty about the probability of bonus receipt would further reduce the expected value of bonus payments to hospitals assessing the costs and benefits of response to the P4P incentives. Although P4P reporting requirements cover all admissions for targeted conditions, bonus payments are only made for the roughly two-thirds of admissions experienced by Medicare beneficiaries. In contrast, the financial incentives associated with public reporting involve a two percentage point reduction in the annual payment rate update across all conditions for noncompliance. CMS provides implicit incentive payments for lower quality outcomes through the outlier payment system, which reimburses additional hospital costs for the sickest and longest-staying patients (including those triggered by hospital-acquired conditions and complications), and the potential for readmissions.

Hospitals could raise comparable levels of revenue by modestly increasing patient volume by attracting new patients (possibly by signaling high quality) or readmitting patients postdischarge (particularly among lower quality hospitals). Even when hospitals can improve P4P scores at very low marginal cost, the response is modest. While it is beyond the scope of this paper to assess the level of bonus payment that would motivate hospital response, our findings suggest that the Premier payments were inadequate to generate changes on intended or unintended dimensions.

In conclusion, our evaluation of the PHQID program highlights lessons for P4P incentive design. First, the incentive mechanism should be relevant for all points in the quality distribution. Second, the quality score should align scientific knowledge about the production process (process measures) with economic incentives to improve or maintain high quality (patient outcomes). Finally, the program must provide a large enough bonus payment to trigger provider response. Our findings suggest that the financial rewards (up to 2 percent of DRG payments for Medicare patients) are insufficient to motivate hospitals to behave strategically as predicted by P4P's motivating logic. Incentive payments large enough to motivate hospital response may exceed public and private payers' willingness to pay for higher quality. Future demonstrations could assess hospital response to bonus payments of larger sizes.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: The authors acknowledge funding from the National Institute on Aging (Nicholas, AG000221-17), The National Heart, Lung and Blood Institute (Iwashyna, K08HL091249), and the Agency for Healthcare Research and Quality (Dimick, 5K08HS017765-02). The findings and conclusions are those of the authors and do not necessarily represent the official views of any of the funding agencies. The authors appreciate comments from participants at the 2009 National Bureau of Economic Research Summer Institute and the 2009 AcademyHealth Health Economics Interest Group Meeting.

Disclosures: None.

Disclaimers: None.

NOTE

1. For example, Ryan (2010) finds scant evidence that the Premier demonstration caused hospitals to reduce service to minority patients.

REFERENCES

Centers for Medicare and Medicaid Services (CMS). 2008. "Premier Hospital Quality Incentive Demonstration Fact Sheet" [accessed April 23, 2009]. Available at http://www.cms.hhs.gov/HospitalQualityInits/Downloads/HospitalPremierFactSheet200806.pdf

Centers for Medicare and Medicaid Services (CMS). 2009. "Roadmap for Implementing Value Driven Healthcare in the Traditional Medicare Fee-for-Service Program" [accessed March 4, 2010]. Available at https://www.cms.gov/QualityInitiativesGenInfo/downloads/VBPRoadmap_OEA_1-16_508.pdf

Glickman, S. W., F. Ou, E. R. DeLong, M. T. Roe, B. L. Lytle, J. Mulgund, J. S. Rumsfeld, W. B. Gibler, E. M. Ohman, K. A. Schulman, and E. D. Peterson. 2007. "Pay for Performance, Quality of Care and Outcomes in Acute Myocardial Infarction." *Journal of the American Medical Association* 297 (21): 2373–80.

Horwitz, J., and A. Nichols. 2007. What Do Non-Profits Maximize? Non-Profit Hospital Service Provision and Market Ownership Mix. NBER Working Paper 13246.

Institute of Medicine. 2007. *Rewarding Provider Performance.* Washington, DC: National Academy Press.

Jha, A. 2005. "Care in U.S. Hospitals—The Hospital Quality Alliance Program." *New England Journal of Medicine* 353 (3): 265–74.

Lindenauer, P. K., D. Remus, S. Roman, M. B. Rothberg, E. M. Benjamin, A. Ma, and D. W. Bratzler. 2007. "Public Reporting and Pay for Performance in Hospital Quality Improvement." *New England Journal of Medicine* 356 (5): 486–96.

Mullen, K. J., R. G. Frank, and M. B. Rosenthal. 2009. Can You Get What You Pay for? Pay-for-Performance and the Quality of Healthcare Providers. NBER Working Paper 14886.

Petersen, L. A., L. D. Woodard, T. Urech, C. Daw, and S. Sookanan. 2006. "Does Pay-for-Performance Improve the Quality of Health Care?" *Annals of Internal Medicine* 145 (4): 265–72.

Rosenthal, M. B., R. G. Frank, Z. Li, and A. M. Epstein. 2005. "Early Experience with Pay-for-Performance from Concept to Practice." *Journal of the American Medical Association* 294 (14): 1788–93.

Ryan, A. 2009. "Effects of the Premier Hospital Quality Incentive Demonstration on Medicare Patient Mortality and Cost." *Health Services Research* 44 (3): 821–42.

———. 2010. "Has Pay-for-Performance Decreased Access for Minority Patients?" *Health Services Research* 45 (1): 6–23.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Table SA1: Pay-for-Performance Participation and Hospital Process Compliance, Voluntary Hospital Compare Reporters: Fiscal Years 2003–2005.

Table SA2: Pay-for-Performance Participation and Hospital Process Compliance among Early Reporters, Fiscal Years 2003–2005, Individual Measure Regression Coefficients.