3. Review of the Pay-for-Performance Literature

The use of P4P in health care emerged in the late 1990s, and between 1999 and 2012, a number of natural experiments testing P4P occurred. On the federal side, CMS started testing the application of P4P in the hospital setting through the Premier HQID and in the physician group practice setting through the PGP demonstration. Much of the published literature related to hospital P4P comes from early and more recent evaluations of HQID. The Premier HQID initially provided incentive payments to hospitals for attaining predetermined performance levels and then evolved to reward both attainment and improvement.

During this same period, private payers began experimenting with P4P that primarily targeted ambulatory care providers (i.e., physician groups and, in some cases, individual physicians). While there have been various small tests of P4P that have yielded very limited information on the impact of P4P, there have also been a few large-scale private sector demonstrations (e.g., Rochester, New York; California; Hawaii; and Massachusetts) that have provided more robust tests of the P4P concept. Several of these early large-scale P4P experiments received start-up funding from the Robert Wood Johnson Foundation's Rewarding Results initiative.

This chapter summarizes our review of the P4P literature to extract information related to each of the research questions that were the focus of this study.

Methods

The goal of the search strategy was to identify all published P4P evaluations. We searched PubMed, including only articles that were published in English and between January 1, 2000, and December 6, 2012. The search terms that we used are listed in Table 3.1. A librarian performed the initial search, which was reviewed by the two senior researchers on the project.

We supplemented the results from this search with additional strategies. We combined the Endnote library for a previous 2007 review of P4P articles[19] with the PubMed search. Several systematic reviews on P4P have been conducted,[9, 12, 96, 108, 125–130] and we reference-mined these reviews on P4P to ensure that key articles were identified. We scanned the titles listed in the reference section of the reviews to identify additional articles for inclusion. Additionally, we cross-referenced relevant articles, conducted ad hoc Google Scholar searches, and conducted a targeted PubMed search for articles published by leading P4P researchers (see Table 3.1). In addition to the search strategies we implemented, the TEP identified several additional studies of P4P that we included in our review.

Search results were catalogued in Endnote software and organized by the following categories: U.S. P4P program evaluations, commentaries/editorials, government documents,

systematic reviews, qualitative evaluations, international evaluations, and background articles. We limited our focus to articles that summarized findings from program evaluations, and we excluded commentaries/editorials or background articles in our abstraction. Using these various search strategies, we identified a total of 1,891 articles for screening, after excluding duplicates (Figure 3.1). After consulting with our TEP, we identified seven additional studies that were published after the December 2012 search date, which we screened for possible inclusion.

Table 3.1. Search Terms Used in Pay-for-Performance Literature Review

Search Terms	Search Engine	Search Dates
PubMed Search Terms: "pay for performance"[tiab] OR P4P[tiab] OR "pay for value"[tiab] OR "financial incentive" OR ((bonus[tiab] OR reward[tiab]) AND (payment[tiab] OR reimburse*[tiab] OR incentive*[tiab]) AND (quality[tiab] OR value[tiab])).	PubMed	January 1, 2000–December 06, 2012
Selected P4P researcher search: Howard Beckman, Kathleen Curtin, Larry Casalino, Adams Dudley, Tim Doran, Ashish Jha, Laura Petersen, Martin Roland, Meredith Rosenthal, Andrew Ryan, Eric Schneider, Rachel Werner, Cheryl Damberg	PubMed Google Scholar	January 1, 2000–December 6, 2012
2007 RAND Hospital P4P Review search terms: "pay for performance" OR "p4p" OR "pay for quality" OR "pay for value" OR "value based purchasing" OR "financial incentives" OR "monetary incentives" OR (bonus* OR reward* OR (incentive reimbursement)) AND "quality" AND "hospital" OR "hospitals"	PubMed, EconLit, CINAHL, Psycinfo, and ABInform	January 1, 1996–June 30, 2007

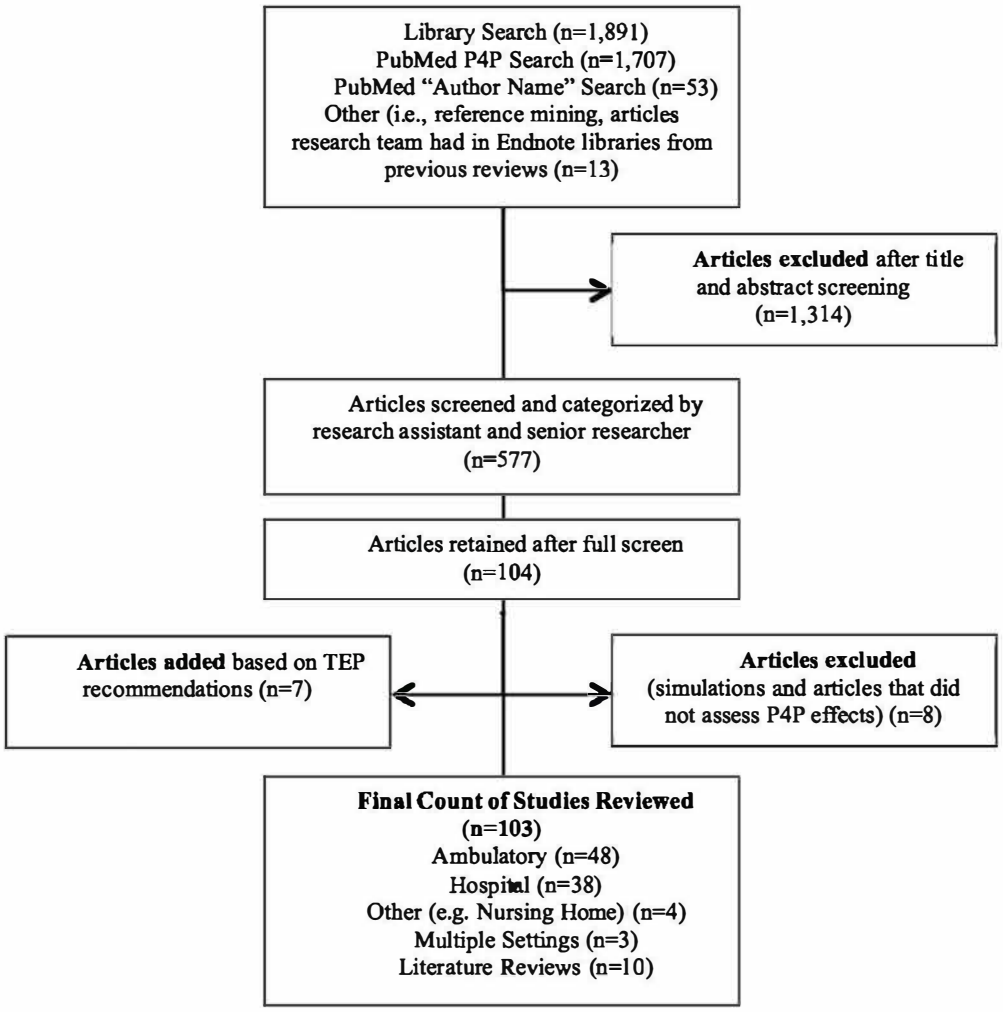
The research assistant on the team (Laura Raaen) conducted an initial screening of titles and abstracts for relevance and content. If there was indecision about whether or not an article was relevant, it was included. A senior researcher (Cheryl Damberg) on the team reviewed the final set of titles and abstracts and excluded those articles that did not examine the effects of implementing P4P. The final count of studies reviewed was 103. Once the list was finalized, the research assistant then abstracted the set of articles. As a quality check, two researchers (Grant Martsolf, Cheryl Damberg) on the project team reviewed the data abstracted by the research assistant.

As described more fully in Chapter One (in the section "Methods and Research Questions"), we rated the methodological quality of each study as follows: **good** indicates a low risk of bias (i.e., the study has strong methods to guard against bias); **fair** indicates a medium risk of bias; and **poor** indicates a high risk of bias. We based the assessment on the strength of the study design, analytic techniques used to control for confounding explanations, intervention characteristics, and conflict of interest/independence of the evaluator. We also graded the strength of the evidence as a whole for each research question using four grade levels:

- **High—**A high degree of confidence that the evidence reflects the true effect. Additional research is unlikely to change the estimate of the effect.
- **Moderate**—Moderate confidence that the evidence reflects the true effect. Additional research may change the estimate or confidence in the estimate of the effect.

- **Low**—Low confidence that the evidence reflects the true effect. Further evidence is likely to change our confidence in the estimate of effect and is likely to change the estimate. A low rating indicates that there is a high risk of bias and residual confounding.
- **Insufficient**—A lack of evidence to estimate the effect(s).

Figure 3.1. Process Used to Identify Articles for Review, Pay-for-Performance



Research Questions

Measuring Performance in Value-Based Purchasing Programs

1. What goals should be set and how should success be defined for VBP programs?

As discussed in Chapter Two (environmental scan of VBP programs), P4P sponsors generally established goals that were high-level (e.g., "improved health," "bend the cost curve") and heavily emphasized clinical quality (27 out of 35 programs). Goals related to cost/affordability

(16 of 35) and patient outcomes (10 of 35) were next most common, and less frequently mentioned were goals related to patient safety, care coordination, patient experience, and infrastructure development. Program sponsors rarely established quantifiable goals to facilitate the ability to measure whether the program was successful; the handful of exceptions were goals related to cost savings targets.

From the literature review, we found mention of the following P4P program goals:

- Increase adherence to heart care clinical guidelines designed to assure patient safety and improve community health.
- Encourage greater quality improvement, particularly among low-performing hospitals.
- Improve evidence-based care and reduce asthma-related emergency department/urgent care visits, admissions, office visits because of acute symptoms, missed school days, missed workdays, and daytime and evening symptoms. Redesign care delivery within primary care practices.
- Improve chronic care treatment for diabetic members and promote the development of office-based systems of care.
- Improve diabetes care quality and outcomes.
- Improve quality and productivity.
- Improve the quality of and access to preventive care services for children.
- Encourage plan members to seek prenatal care in the first trimester of pregnancy.
- Incentivize nonprofit providers to care for high-priority clients in a cost-effective manner.

Strength of Evidence: Not applicable, descriptive only.

2. What are the metrics by which VBP programs can and should be evaluated?

We did not find information to address this question from the literature review. We direct the reader to the TEP's discussion of this question (Chapter Six).

3. Which aspects of VBP are measurable and which are not?

We did not find information to address this question from the literature review. We direct the reader to the TEP's discussion of this question (Chapter Six).

4. What is the relationship between health outcomes and what is measured in VBP programs?

Hospital Measures

We reviewed 13 articles (Tables 3.2 and 3.3) that assessed the relationship between clinical process-of-care measures and patient outcomes. The articles addressed four clinical conditions: AMI (10 articles), pneumonia (7 articles), CHF (6 articles), and major surgeries (2 articles). The articles examined a relatively small number of risk-adjusted or risk-standardized outcome measures. Thirty-day mortality (8 articles) and in-hospital mortality (7 articles) were the most commonly assessed outcomes, while few studies examined complications (2 articles), 30-day readmissions (2 articles), or one-year survival (1 article). The studies typically used cross-

sectional data and examined associations (i.e., correlations) between individual or composite clinical process measures with one or more outcomes, or they measured the process-outcome relationship by comparing outcomes in high versus low performers on process measures. Many of the studies controlled for patient and hospital characteristics in at least some of their analyses.

The three studies examining surgical care had inconsistent, but mostly nonsignificant findings. Bhattacharyya et al.,[131] which was a poorly designed study, did not find a significant difference in inpatient mortality across four tiers of performance on measures for hip and knee arthroplasty, but did find a trend toward higher mortality in the worst-performing tier ($p=0.08$). Stefan and colleagues[132] found that performance scores were weakly associated with readmission rates for orthopedic surgery, but not abdominal, cardiac, or vascular surgery. Nicholas et al.[133] found no consistent relationships between process-of-care measures and 30-day mortality rate or selected complications for six high-risk surgical procedures.

Studies have consistently found a weak relationship between better performance on process measures or composite measures and better patient outcomes for AMI and pneumonia, although the amount of variation in outcomes explained by variation in process measures was low, and the absolute risk reduction of moving from poor-performing hospitals to high-performing hospitals was small. The results were less consistent for CHF. While one study found that hospital performance on process measures was weakly negatively correlated with in-hospital mortality, the three studies examining 30-day mortality had inconsistent results, and the one study examining one-year mortality failed to find an association.

A study by Ryan et al.[71] raised questions about whether observed associations are causal in nature. While many studies controlled for hospital characteristics in multivariable analyses, Ryan, in contrast, included hospital fixed effects, year fixed effects, and interactions between time-variant hospital characteristics and year. The hospital fixed effects adjust for unobservable characteristics that could affect hospital performance on both process measures and outcome measures, such as interest in quality improvement. The resulting observed association between process and outcome is driven by within-hospital changes in performance over time rather than differences in performance between hospitals. While Ryan's models without hospital fixed effects showed negative associations between composite measures of quality and 30-day mortality, these associations reduced in magnitude and were not statistically significant with the inclusion of the fixed effects. While this suggests that the process-outcome relationship is not a causal one, the results are not conclusive. The changes in performance over the three years of data included in the study were small. To the extent that the changes across hospitals were similar, these could be captured by the year fixed effects. Even then, however, the magnitude of any causal relationship would be small.

Results of studies were also somewhat sensitive to analytic decisions. For example, the correlation between inpatient AMI measures and patient outcomes are sensitive to whether or not patients that transferred out of the hospital that was the unit of analysis are included in the analyses; correlations were stronger when these patients were excluded.[125, 134]

Bradley et al.[134] found that better performance on beta-blocker at discharge, aspirin at discharge, timely reperfusion therapy, and a quality composite, but not other AMI process measures was associated with lower risk-standardized 30-day all-cause mortality while aspirin at arrival was the only AMI process measure that was significantly associated with lower risk-standardized in-hospital, all-cause mortality. In contrast, Petersen[125] found that a broader set of AMI measures were associated with lower in-hospital mortality among a small group of hospitals participating in a QI initiative. Werner and Bradlow[135] found that the absolute risk reduction for AMI and pneumonia measures was greater for one-year mortality than 30-day mortality.

Table 3.2. Summary of Studies Examining the Association Between Process and Outcome Measures

	Risk-Adjusted or Standardized Outcomes									
	30-Day Mortality		In-Hospital Mortality		Complications		30-Day Readmissions		1-Year Survival	
Condition-Related Process Measures	# Studies Lower Mortality	# Studies Non-significant Effect	# Studies Lower Mortality	# Studies Non-significant Effect	# Studies Fewer Complica-tions	# Studies Non-significant Effect	# Studies Fewer Readmissions	# Studies Non-significant Effect	# Studies Better Mortality	# Studies Non-significant Effect
AMI										
Beta-blocker use at admission	1	1	1	4					1	
Beta-blocker use at discharge	2		1	2					1	
Aspirin use at admission	1	1	3						1	
Aspirin use at discharge	2		2	1					1	
ACE inhibitor use at discharge		2	2	1						1
Smoking cessation counseling for smokers during admission		1		1						
Timely reperfusion therapy	1			1						
Heparin at admission			1							
Intravenous glycoprotein IIb/IIIa inhibitors at admission			1							
Lipid lowering medication at discharge			1							
AMI composite measures[3]	5[1]		4[2]	1			1	1	1	
CHF										
CHF composite measures[4]	2[1]	1	2	1				1		1
Pneumonia										
Antibiotics timing	1		1	1					1	
Pneumonia composite measures[5]	2[1]	1	2				1		1	
Orthopedic Surgery										
Composites of SCIP and other process measures[6]				1		1	1			
High Risk Surgical Procedures										
Composites of SCIP measures[7]		1[8]				1				

[1] In one study, significant results were no longer observed when hospital fixed effects were included in the model.
[2] In one study, two composites with different weighting of the measures were included in the model. One composite was associated with lower inpatient mortality and one was associated with higher inpatient mortality.
[3] Two different AMI process measure composite measures were used. One included five measures: beta-blocker use at admission, beta-blocker use at discharge, aspirin use at admission, aspirin use at discharge, ACE inhibitor use at discharge. The other composite included these measures plus smoking cessation counseling and timely reperfusion therapy.
[4] Two different CHF process measure composites were used. One included two measures: ACE inhibitor or angiotensin receptor blocker for left ventricular systolic and dysfunction and assessment of left ventricular function. The other composite included these measures plus smoking cessation counseling and discharge instructions.
[5] Two different pneumonia process measure composite were used. One included 3 measures: antibiotics provided within 4 hours or less, pneumococcal vaccination, and oxygenation assessment. The other included these measures plus blood culture prior to antibiotics, appropriate antibiotic, pneumococcal vaccination status, influenza vaccination status, and smoking cessation counseling.
[6] Two different process-of-care composite measures were used for orthopedic surgery. One included 6 measures: metabolic complication avoidance index, hematoma avoidance index, readmission avoidance index, antibiotics administered within 1 hour before incision, antibiotics discontinued within 24 hours of surgery, appropriate antibiotic selection. The other included 9 SCIP measures: prophylactic antibiotic received within 1 hour prior to surgery, prophylactic antibiotic selection, prophylactic antibiotic discontinuation within 24 hours after surgery, cardiac surgery patients with controlled 6 AM postoperative glucose, patients with appropriate hair removal, colorectal surgery patients with immediate postoperative normothermia, recommended venous thromboembolism prophylaxis ordered, recommended venous thromboembolism prophylaxis ordered and received, surgery patients on beta-blocker therapy prior to admission who received a beta-blocker during perioperative period.
[7] Two different SCIP measure composites were used. One included 5 SCIP measures: receipt of prophylactic antibiotics within 2 hours of surgery, discontinuation of prophylactic antibiotics within 24 hours of surgery, selection of correct prophylactic antibiotic, ordering of venous thrombosis prophylaxis, ordering of venous thrombosis prophylaxis within 24 hours of surgery. The other included these measures plus cardiac surgery patients with controlled 6 AM postoperative glucose, patients with appropriate hair removal, colorectal surgery patients with immediate postoperative normothermia, recommended venous thromboembolism prophylaxis ordered and received, surgery patients on beta-blocker therapy prior to admission who received a beta-blocker during perioperative period.
[8] Non-significant effects except abdominal aortic aneurysm, where highest SCIP compliance had lower mortality rates.

Ambulatory Measures

A 2011 systematic review[136] summarized the literature on the relationship between quality indicators and outcomes for diabetes. Of the 24 studies included in the review, three cohort studies and four case-control studies examined the relationship between process measures and outcomes (i.e., disease-related complications, lower extremity amputations, death, and measures of mental and physical health). There was relatively little overlap in the combination of process and outcome measures assessed by the different studies, increasing the challenges of assessing the consistency of results in the literature. For any of the process measures examined, evidence on its relationship to patient outcomes was mixed at best.

In a study by Ryan and Doran,[137] the researchers conducted a retrospective analysis to evaluate the association between improvements in incentivized process and intermediate outcomes among family practices participating in the UK Quality and Outcomes Framework. The study analyzed data from 2004 through 2008 for five conditions: diabetes, coronary heart disease, stroke, epilepsy, and hypertension. The researchers constructed condition-specific composite measures for the process and outcome measures for each year. Longitudinal fixed effects models controlling for composite process performance for all other conditions and year fixed effects were used to estimate the extent to which improvements in incentivized intermediate outcomes were associated with improvements in incentivized process measures. The study showed that a 10 percentage point increase in the process composite was associated with an increase in intermediate outcome performance of 3.16 percentage points for diabetes, 4.32 percentage points for coronary heart disease, 7.60 percentage points for stroke, 7.24 percentage points for epilepsy, and 7.16 percentage points for hypertension. In other words, the amount of the increase in the intermediate outcome composite due to the change in the process composite ranged from 17 percent for hypertension to 34.7 percent for stroke.

A study by Kralewski and colleagues[138] found an association between low-density lipoprotein (LDL) testing and the number of avoidable emergency department visits and hospital admissions among 133,704 diabetic Medicare beneficiaries in 234 group practices. Group practices that performed LDL testing for all diabetic patients significantly reduced the number of unnecessary emergency department visits and hospital admissions compared with group practices that did not test all patients. However, the study did not randomly assign beneficiaries to practice groups with differing structural characteristics, and certain practice characteristics were associated with outcomes variables. The number of support services available on site was associated with both avoidable emergency department visits and hospital admissions, while larger practice size, more nurse practitioners, and more physician's assistants relative to the number of physicians were associated with more avoidable hospitalizations. Government owned practices, community health centers, and physician-owned practices were associated with few avoidable hospitalizations.

Nursing Home Measures

We identified only one study that examined the relationship between process-of-care measures in the nursing home setting and outcome measures for long-stay residents.[74] This was a well-designed longitudinal study that used nursing home fixed effects to assess whether changes in performance on process measures was associated with changes in performance on outcome measures. Approximately one-third of the improvements in the percentage of nursing home patients in moderate or severe pain were due to changes in process measures. None of the improvement in other outcome measures (e.g., pressure sores in low risk or high risk residents) appeared to be due to improvements in process measures. However, there was less than a two-percentage point change in most of the process measures 2000–2009. The exceptions were the percentage enrolled in pain management program (9.0 percentage point change) and percentage receiving preventive skin care (9.43 percentage point change).

Strength of Evidence: Low. A number of studies have attempted to examine the association between receipt of clinical processes and outcomes; however, the findings from these studies are inconclusive. Many of the studies suffer from problems that limit their ability to be able to detect an effect. Studies that attempt to examine the relationship between clinical process measures and outcomes in observational settings face numerous challenges and, if not addressed, can result in incorrect conclusions. The challenges include (1) the population of patients to whom the measure is applied in practice may differ significantly in terms of clinical, demographic, or socioeconomic factors from the patients who were enrolled in the randomized clinical trial (RCT) that served as the basis for the recommended clinical process, and therefore may not achieve the same level of benefit as patients in the RCT; (2) the analyses are under-powered because of too little variance between providers or over time in process measures for the types of outcomes that are readily available (e.g., mortality, readmissions); and (3) a small maximum possible difference in outcomes found in the RCT which, in practice, is even smaller and hard to detect after controlling for potential confounding variables. Given these challenges, the fact that most currently published process-outcome studies could not find an effect is not surprising.

Table 3.3. Articles Examining Relationship Between Performance on Pay-for-Performance Measures and Patient Outcomes

Reference	Setting	Study Design	Program Measure(s)	Patient Outcome(s)	Findings	Assessment of Methodological Quality
Bhattacharyya et al., 2009[131]	Hospital	Cross-sectional analysis of correlation between composite quality score for hip and knee surgery and patient outcomes among the subset of the 260 HQID hospitals that participated in the hip and knee portion of the program in 2004/2005 (actual number of hospitals not reported). Hospitals were placed into 1 of 4 tiers based on composite performance score: top 10% (tier 1); second decile (tier 2); top 50% but not in top 2 deciles (tier 3); bottom 50% (tier 4).	• Composite measure capturing 3 process measures and 3 intermediate outcome measures • Data for 4 of the 6 individual measures were only available for those hospitals with performance in top 50% of HQID hospitals	• Inpatient mortality after hip and knee arthroplasty • Iatrogenic complications • Urinary tract infections	• Higher-tier hospitals did not have lower complications or urinary tract infections. • No significant difference in hip and knee arthroplasty associated mortality across the hospital tiers, but was a trend toward a higher rate of mortality in tier 4 hospitals ($r = 0.116$; $p = 0.088$). • All hospitals with mortality > 2.0% were in tiers 3 and 4.	**Poor:** Data on 4 of 6 measures used in composite only available for top 50% of performers. Mortality and complications not available for all hospitals. Limited variability in quality composite led to arbitrary placement into tiers. Lack of control for confounders.

Reference	Setting	Study Design	Program Measure(s)	Patient Outcome(s)	Findings	Assessment of Methodological Quality
Bradley et al., 2006[134]	Hospital	Cross-sectional analysis of correlation between CMS/Joint Commission AMI core process measures and hospital-level, risk-standardized measures of patient outcomes using January 2002–March 2003 Medicare claims data from 962 hospitals participating in the National Registry of Myocardial Infarction. Hospital-level performance was estimated using hierarchical generalized linear models as well as crude process rates. Main analysis included patients transferred out; these were excluded in secondary analyses	• 7 AMI process measures and a composite quality score	• Risk-standardized 30-day all-cause mortality • Risk-standardized in-hospital mortality	• Risk-standardized 30-day all-cause mortality significantly, but weakly, correlated with beta-blocker at discharge (r=–.16, p<.001), aspirin at discharge (r=–.18, p<.001), timely reperfusion therapy (r=–.18, p<.001), and the quality composite (r=–.25, p<.001), but not with other process measures (beta-blocker at admission, aspirin at admission, ACE inhibitor at discharge, smoking cessation counseling). • Amount of variation in 30-day mortality explained by process measures ranged from 0.1% to 3.3%; the measures jointly explained 6% of variation. • Aspirin at admission was weakly associated with risk-standardized in-hospital, all-cause mortality (r=–.12, p<.05); other measures, including the composite, were not.	Fair

Reference	Setting	Study Design	Program Measure(s)	Patient Outcome(s)	Findings	Assessment of Methodological Quality
Glickman et al., 2009[139]	Hospital	Assessed association between AMI and CHF process measures and inpatient mortality measures after AMI among 1,351 hospitals participating in Hospital Compare that had at least one patient eligible for AMI measures and one eligible for CHF measures, at least 25 treatment opportunities across all measures, and could be merged with American Hospital Association data on hospital characteristics and Joint Commission data on risk-adjusted inpatient mortality after AMI. Hospital-level multivariable logistic regression assessed association for each scoring system with inpatient survival (1-inpatient mortality) in subsequent year, controlling for hospital-level academic affiliation, geographic location, population density, bed size, presence of percutaneous coronary intervention and cardiac surgery.	• 8 AMI process measures • 4 CHF process measures • Two sets of composite adherence scores assigned different weights to individual measures. • Opportunity model • Principal components analysis used to place measures into one of two groups (clinical cardiac activities and administrative cardiac activities). Adherence was calculated with more weight given to measures with greater opportunity for improvement	• Risk-adjusted inpatient mortality after AMI	• In a model with both clinical and administrative cardiac activities composite, higher clinical cardiac activities were associated with higher inpatient survival (OR=1.13, p<.001), while higher scores for administrative cardiac activities were associated with worse inpatient survival (OR=0.96, p<.001). • When separate composite measures were included for AMI and CHF, AMI performance was associated with improved survival (OR 1.09, p<.001) while the CHF composite was associated with lower inpatient survival (OR 0.98, p<.05).	**Poor:** Outcome measures was risk-adjusted inpatient mortality after AMI, but analyses included quality measures for heart failure patients. In addition, analyses included quality measures for care delivered at discharge, which would not affect inpatient mortality rates

Reference	Setting	Study Design	Program Measure(s)	Patient Outcome(s)	Findings	Assessment of Methodological Quality
Jha et al., 2007[140]	Hospital	Cross-sectional analyses assessed association between condition-specific composite and morality using Hospital Quality Alliance data from April 1, 2004–March 31, 2005, linked with American Hospital Association data on hospital characteristics and 2003 Medicare Provider and Analysis Review (MEDPAR) discharge data for calculating outcomes. Patients received in transfer or transferred to another hospital were excluded. Patient-level multivariable logistic regressions accounting for clustering of patients within hospitals controlling for patient demographics, comorbidities using Elixhauser method, and hospital characteristics were used to estimate the probability of death stratified by hospital's performance on Hospital Quality Alliance measures (by quartiles). The number of hospitals included in analyses ranged from 1,965 for AMI to 3,270 for pneumonia.	• 10 Hospital Quality Alliance process measures were used to create summary performance scores for three clinical conditions: • 5 AMI process measures • 2 CHF process measures • 3 pneumonia process measures	• Risk-adjusted inpatient mortality for patients with primary diagnosis of AMI, CHF or pneumonia	• Significant trend for lower performance being associated with higher mortality for each condition (AMI p<.001; CHF p=.005; pneumonia p<.001). • Compared with hospitals in the bottom quartile of performance, hospitals in the top quartile had ~1% lower mortality for AMI, 0.4% for CHF, and 0.8% for pneumonia. • In multivariable analyses, patients discharged from a hospital in top quartile of Hospital Quality Alliance performance for each condition had a lower odds of dying than patients discharged from hospitals in the bottom quartile performance (AMI: OR=0.91, 95% CI=0.86, 0.96; CHF: OR=0.92, 95% CI=0.88, 0.98; pneumonia: OR=0.90, 95% CI=0.86, 0.95).	**Poor**: The data used to generate mortality rates predates the data on quality measures, which may not reflect the quality of care delivered at the time of the inpatient mortality data. Quality composites used in analyses included measures of care delivered at discharge, would not affect inpatient mortality rates.

Reference	Setting	Study Design	Program Measure(s)	Patient Outcome(s)	Findings	Assessment of Methodological Quality
Jha et al., 2011[111]	Hospital	Cross-sectional analysis of relationship between hospital quality of process-of-care measures, costs and mortality using the 2007 Hospital Compare data, 2005 MEDPAR data linked with the 2005 Medicare Beneficiary file, 2007 American Hospital Association data, 2007 information on hospital-specific cost-to-charge ratios, disproportionate share hospital (DSH) index[a] and ratio of interns and residents to beds, 2007 Area Resource File with county-level socioeconomic information, and the 2008 Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey. Hospital-level risk-adjusted cost ratios (actual to expected costs), quality composite scores, mortality rates, and HCAHPS scores were estimated. Four groups of hospitals were identified: those in the highest quartile of performance and lowest quartile of cost (best), those in the lowest quartile of performance and highest quartile of costs (worst), those in the highest quartile of performance and highest quartile of costs, those in the lowest quartile of performance and lowest quartile of costs.	• Process-of-care measures for AMI, CHF, pneumonia and prevention of surgical complications. • Summary scores were created for each condition using the Joint Commission's methodology for those hospitals.	• 30-day risk adjusted mortality rate for patients hospitalized with AMI, CHF, and pneumonia.	• AMI patients admitted to low-quality hospitals had a higher probability of death than those admitted to the "best" hospitals (low cost, low quality OR=1.12; high cost, low quality OR=1.10; analysis of variance p-value=.005). • Pneumonia patients also had a higher probability of death when admitted to low-quality hospitals (low cost, low quality OR=1.19; high cost, low quality OR=1.07; analysis of variance p-value<.001). • No significant difference observed for CHF.	**Fair**

Reference	Setting	Study Design	Program Measure(s)	Patient Outcome(s)	Findings	Assessment of Methodological Quality
Krumholz et al., 2013[141]	Hospital	30-day readmissions and 30-day mortality were identified for a cohort of aged Medicare beneficiaries with an index hospitalization with a primary diagnosis of AMI, CHF, or pneumonia between July 1, 2005, and June 30, 2008. 30-day all-cause risk-standardized readmission rate (RSRR) and risk-standardized mortality rate (RSMR) were estimated for each hospital using hierarchical logistic regression models that adjusted for patients demographic and clinical characteristics and accounted for patient clustering within hospitals, and had hospital-specific random effects. For each condition, hospitals were considered high performers if they were in the lowest quartile for RSMR and RSRR and lower performers if they were in the highest quartile for both. Analysis included 4506 hospitals for AMI, 4767 hospitals for CHF, and 4811 hospitals for pneumonia.	Not applicable	For AMI, CHF, and pneumonia • 30-day all-cause risk-standardized mortality rates (RSMRs) • 30-day, all-cause, risk-standardized readmission rates (RSRRs)	• Overall, there was no association between RSMR and RSRRs for AMI or pneumonia. • There was a negative association between RSMRs and RSRRs for CHF (r=–.17, 95% CI –.20 to –.14).	**Good**

Reference	Setting	Study Design	Program Measure(s)	Patient Outcome(s)	Findings	Assessment of Methodological Quality
Nicholas et al., 2010[133]	Hospital	Cross-sectional analysis of SCIP measures reported on Hospital Compare data Jan 1, 2005–Dec 31, 2006, and patient outcomes derived from MEDPAR data for patients with 1 of 6 high-risk surgical procedures (abdominal aortic aneurysm repair, aortic valve repair, coronary artery bypass graft, esophageal resection, mitral valve repair and pancreatic resection) using hierarchical linear models to assess associations. Models controlled for hospital-level procedure volume and patient characteristics and comorbidity using the Charlson comorbidity index, whether the admission was scheduled, emergent or urgent, zip code-level median income, year of admission and hospital random effects. Hospitals were placed in low (bottom quintile of performance), medium (middle three quintiles of performance) and high (top quintile of performance) compliance groups based on opportunity composite score. Analyses included 2,189 hospitals.	• 2 SCIP measures in 2005: • An additional 3 measures were included in 2006 • An opportunity composite score was created	• 30-day risk-adjusted postoperative mortality rate, venous thrombo-embolism, and surgical site infection.	• In univariate analyses, there were no significant associations between process measures and mortality except for aortic valve replacement where hospitals with highest SCIP compliance had lower mortality rates. • In multivariate analyses, neither high nor low compliance hospitals were significantly different from hospitals with middle compliance; nor did high and lower compliance hospitals have different mortality rates from one another. • Unadjusted complication rates were lower among hospitals in the lowest compliance quintile than those in the highest compliance quintiles. Results were not significant in multivariate analyses.	Good

Reference	Setting	Study Design	Program Measure(s)	Patient Outcome(s)	Findings	Assessment of Methodological Quality
Peterson et al., 2006[125]	Hospital	The association between process-of-care measures for patients presenting with symptoms consistent with acute coronary syndrome to 350 hospitals participating in the "Can Rapid Risk Stratification of Unstable Angina Patients Suppress Adverse Outcomes with Early Implementation of the American College of Cardiology/American Hospital Association Guideline" (CRUSADE) National Quality Improvement Initiative between January 1, 2001, and September 30, 2003, and in-hospital mortality was examined using Pearson correlation coefficients and Cochran-Armitage test for trend. Adjusted mortality rates were estimated using hierarchical generalized linear mixed models adjusting for patient characteristics, comorbid conditions, and a patient's propensity to be treated at a top quartile center.	• 9 cardiac process-of-care measures • Opportunity model composite was created	• In-hospital mortality	• Improved performance on process measures was significantly, though modestly, associated with lower in-hospital mortality (ranging from –.12 to –.36) (p<.05) except for beta blocker within 24 hours and beta-blocker at discharge, which were not significant. • Composite measure of quality was negatively associated with in-hospital mortality (r=–.30, p<.001). • The adjusted in-hospital mortality rate for hospitals in the top quartile was 6.31% versus 4.15% for hospitals in the 4th quartile (OR=0.81, p<.001).	Fair

Reference	Setting	Study Design	Program Measure(s)	Patient Outcome(s)	Findings	Assessment of Methodological Quality
Popescu et al., 2009[142]	Hospital	The association between AMI process measures 2004–2006 and risk-adjusted 30-day mortality for 2005 was assessed for 2761 hospitals reporting AMI measures to the Hospital Compare database. Hospitals were categorized as high adherence (top decile of performance on AMI measures for 3 consecutive years), low adherence (lowest decile of performance for 3 consecutive years), or intermediate performance (all other hospitals in sample). 30-day mortality rates for AMI patients were estimated using multivariable mixed models controlling for patient sociodemographic characteristics and comorbidity as well as hospital random effects.	• 5 AMI process measures • Opportunity model composite was created	• 30-day mortality	• Mean AMI performance varied significantly across the three groups p<.001). • Low-performing hospitals had higher unadjusted 30-day mortality rates (23.6% vs. 17.8% vs. 14.9%, p< 0.001). • Differences persisted after adjusting for patient characteristics (16.3% vs. 16.0% vs. 15.7%; P 0.02).	Fair

Reference	Setting	Study Design	Program Measure(s)	Patient Outcome(s)	Findings	Assessment of Methodological Quality
Quattromani et al., 2011[143]	Hospital	Cross-sectional analysis of 95,704 adult emergency department admissions with a principal diagnosis of pneumonia from 530 hospitals in the 2007 Hospital Healthcare Cost and Utilization's National Inpatient Sample linked with hospital-level data on the timely receipt of antibiotics and American Hospital Association data. Hospitals were placed in quartiles based on their timely receipt of antibiotics performance. A population-averaged logistic regression model controlled for patient demographics and comorbid conditions, weekend admission, and accounting for correlation of patients within hospitals.	• Receipt of first dose of antibiotics within 4 hours of arrival at hospital	• All-cause inpatient mortality	• No significant associations found; compared with the lowest-performing hospitals, the risk-adjusted OR of mortality was 0.89 (95% CI = 0.77 to 1.02) in the highest-performing time-to-first-antibiotic-dose quartile, 0.94 (95% CI = 0.82 to 1.08) in the second quartile, 0.91 (95% CI = 0.79 to 1.05) in the third quartile.	Fair

Reference	Setting	Study Design	Program Measure(s)	Patient Outcome(s)	Findings	Assessment of Methodological Quality
Ryan et al., 2009[78]	Hospital	Medicare inpatient claims and Hospital Compare process-of-care measures for 2004–2006 were used to assess relationship between the process measures and risk-adjusted patient outcomes. One model estimated the relationship between performance and the log of risk-adjusted mortality, controlling for hospital characteristics, year and hospital characteristics - year interactions. The second model included hospital fixed effects to capture unobserved characteristics as well as year and hospital characteristics interacted with year. Excluded from analysis were transfer patients and hospitals with less than 10 patients for each measure.	• 5 AMI process measures • 2 CHF process measures • 3 pneumonia process measures • Two methods for creating composites were used: • The weighted sum of z-scores for process measures for each diagnosis • The z-score of the unweighted sum of each process measure for each diagnosis	• Risk-adjusted 30-day mortality for AMI, CHF, and pneumonia	• Based on the models with hospital characteristics, a one standard-deviation increase in process measure composite was associated with a 9% reduction in mortality for AMI ($p<.01$), 1.5% reduction for CHF ($p<.05$) and 1.9% reduction for pneumonia ($p<.01$). • Associations no longer significant when hospital fixed effects included in the models. • These results are supported by finding that while small process performance improvements from 2004 to 2006, there were not similar changes in mortality.	**Good**

Reference	Setting	Study Design	Program Measure(s)	Patient Outcome(s)	Findings	Assessment of Methodological Quality
Stefan et al., 2013[132]	Hospital	The association between Hospital Compare process quality measures and 30-day readmission for patient with AMI, CHF, or pneumonia and those undergoing major surgery in 2007 was examined using Spearman rank correlations. Data were obtained from the Quality Improvement Organization Clinical Data Warehouse. 30-day readmission rates were estimated using the same technique as CMS for the Hospital Compare website, with hierarchical generalized linear models accounting for patient clustering within hospitals, adjusted for patient characteristics, zip-code level median income, comorbidities, discharge disposition, number of admissions in previous year, and length of stay relative to median length of stay for that condition. A ratio of predicted to expected readmission rate was calculated for each hospital for each condition. Hospitals were placed into quartiles based on performance score for each condition and the absolute difference in mean risk-standardized readmission rates of hospitals in the highest and lowest quartiles of performance calculated.	• 8 AMI process measures • 7 pneumonia process measures • 4 CHF process measures • 9 SCIP measures • Two sets of composite adherence scores used. (1) an opportunities composite and (2) an appropriate care composite (i.e., did patients receive all care processes for which they were eligible?)	• Condition-specific 30-day risk standardized readmission rate (only for those also included in process-of-care measures)	• Higher performance scores were significantly, but weakly correlated with lower readmission rates for pneumonia (r=–.07, p<.0001), AMI (–.10, p<.0001) and orthopedic surgery (r=-.06, p<.003), but not heart failure, abdominal surgery or cardiac and vascular surgery. • Results very similar whether opportunity model or appropriate care composite used. • Multivariable models with process measures and hospital characteristics explained a very small amount of total variation in hospital-level readmission rates. • The difference in mean risk-standardized readmission rates between hospitals in the 1st and 4th quartiles of process performance significant for AMI, but difference in readmission rates only 0.3 percentage points.	**Good**

Reference	Setting	Study Design	Program Measure(s)	Patient Outcome(s)	Findings	Assessment of Methodological Quality
Werner and Bradlow, 2006[135]	Hospital	Examined correlation between Hospital Quality Alliance 10 measure starter set from Hospital Compare for 2004 and hospital-level patient outcomes calculated using 2004 MEDPAR data and risk adjusted using the Elixhauser method, patient characteristics, and whether the admission was emergent or elective in 3657 hospitals using. Hospitals were grouped into thirds based on average 1-year risk-adjusted mortality rate for each condition. A Bayesian approached was used to assess relationship between composite measures, individual performance measures and condition-specific outcomes. The relationship between hospital performance and outcomes were estimated controlling for hospital characteristics.	• 5 AMI process measures • 2 CHF process measures • 3 pneumonia process measures • Two composite measures created • Opportunity model composite • An "all or none" measure that identified hospitals that performed above the 75th percentile on every measure they reported and hospitals that performed below the 75th percentile on every measure reported	• Condition-specific inpatient mortality • Condition specific 30-day mortality • 1-year risk adjusted mortality rates	• Adjusting for hospital characteristics, hospitals in the 75th percentile had significantly lower inpatient mortality than those performing in the 25th percentile for each condition's composite measure and most of the individual measures. • The absolute risk reduction (ARR) was small, ranging from .001 for CHF to .005 for both AMI and pneumonia. • Results were similar for 30-day mortality. • Results for 1-year mortality were significant for AMI and pneumonia, but not for CHF. • Comparing hospitals performing above the 75th percentile on all measures to those performing below the 25th percentile on all measures, the ARR for AMI ranged from 0.008 (p=.06) for inpatient mortality to 0.18 (p=.008) for 1-year mortality. • The ARR for pneumonia was .014 (p<.001) in inpatient mortality, .003 (p=.00) for 30 day mortality and 0.13 (p<.001) for 1 year mortality.	**Good**

Reference	Setting	Study Design	Program Measure(s)	Patient Outcome(s)	Findings	Assessment of Methodological Quality
Kralewski et al., 2012[138]	Ambulatory care	Cross-sectional study of 133,703 Medicare patients with diabetes treated by 234 group practices in 2009. Patients were attributed to the practice where they received the plurality of their care. Claims data were used to assess lab testing, emergency department use, hospitalizations and total costs. Practice structural characteristics were obtained from the 2009 practice survey of the Medical Group Management Association. Regression analysis was used to assess association between measures and risk-adjusted outcomes.	• LDL lab test during the past year	• Inappropriate emergency department use • Avoidable hospitalizations • Costs per patient with diabetes	• LDL testing for an additional one percentage point of diabetics in the practice was associated with reduced per capita costs of $51 (p<.001), fewer primary care treatable emergency visits (p<.001) and few avoidable hospitalizations (p<.001).	Fair

Reference	Setting	Study Design	Program Measure(s)	Patient Outcome(s)	Findings	Assessment of Methodological Quality
Ryan and Doran, 2012[137]	Ambulatory care	Retrospective analysis of the amount of improvement in incentivized intermediate outcomes was a result of improvements in incentivized process measures for diabetes, coronary heart disease, stroke, epilepsy, and hypertension using 2004–2008 data from a panel of family practices participation in the UK's Quality Outcomes Framework. Data on practice performance was linked to patient and practice characteristics and community-level Index of Deprivation. The number of included practices ranged from 3864 (epilepsy) to 6822 (diabetes). "Opportunities model" composite measures were created for each year separately for process and outcomes measures for each condition for each practice. Longitudinal fixed effects models controlling for composite process components performance for all other conditions and year fixed effects were used to estimate the extent to which improvements in incentivized outcomes were due to improvements in incentivized process measures. Separate models were run for each diagnosis. Standard errors accounted for clustering at the practice level.	• 10 diabetes process measures • 5 coronary heart disease process measures • 3 stroke process measures • 2 epilepsy process measures • 1 hypertension process measure	• Intermediate outcomes • 4 for diabetes • 2 for coronary heart disease • 2 for stroke • 1 for epilepsy • 1 for hypertension	• A 10 percentage point increase in process composite was associate with an increase in the outcome performance of 3.16 percentage points for diabetes, 4.32 percentage points for coronary heart disease, 7.60 percentage points for stroke, 7.24 percentage points for epilepsy and 7.16 percentage points for hypertension. • The amount of increase in the outcome composite due to the change in the process composite was 29.6% for diabetes, 25.6% for coronary heart disease, 34.7% for stroke, 29.1% for epilepsy, and 17.7% for hypertension.	**Good**

Reference	Setting	Study Design	Program Measure(s)	Patient Outcome(s)	Findings	Assessment of Methodological Quality
Sidorenkov et al., 2011[136]	Multiple settings	Systematic review of literature indexed on MEDLINE and Embase up through May 1, 2010, that focused on relationship between quality indicators and outcomes for diabetes care. Studies were classified as high, medium, or low quality. 24 studies were identified, 17 of which evaluated intermediate outcomes. Of the studies assessing "hard" outcomes, 3 were cohort and 4 were case-control studies	• Adequate drug treatment • visits and exams • HbA1c tests • other or composite tests/exams	• Hospitalizations • Treatment-related complications, • Disease-related complications, hospital • Readmissions, • Microvascular complications or lower extremity amputations • Macrovascular complications • Death • Composite physical and/or mental health score	• Few associations between process measures and outcome measures were identified. One study showed adequate drug treatment of patients hospitalized for diabetes was associated with fewer treatment-related complications, but another study[144] found no association with readmission rates. • A medium-quality cohort study found HbA1c testing was associated with decreased macrovascular complications and kidney disease, but not microvascular complications or death.[145] • Lipid testing was associated with fewer lower extremity complications, while eye exams were not. • A high-quality study showed a composite measure that captured HbA1c testing, eye exams, LDL screening and nephropathy monitoring was associated with better mental health status but not physical health status as measured by the SF36.[146]	Good

Reference	Setting	Study Design	Program Measure(s)	Patient Outcome(s)	Findings	Assessment of Methodological Quality
Werner et al., 2013[74]	Nursing home	Assessed the extent to which changes in nursing home process measures account for changes in outcome measures among 16,623 nursing homes reporting data from 2000 to 2009 for the Online Survey, Certification, and Reporting and nursing home Minimum Data Set. Analyses included facility fixed effects, time-varying facility characteristics, indicator for quarter of the year to capture seasonal effects, and quarter interacted with process measures.	• 6 process measures focused on pain management, written bladder training program, preventive skin care, receiving tube feeds, mechanically altered diets, assist devices while eating	• 4 outcome measures focused on long-stay residents with moderate or severe pain, catheter inserted and left in their bladder, pressure sores, or significant weight loss	• Approximately one-third of the improvements in the percentage of nursing home patients in moderate or severe change were due to changes in process measures. • None of the improvements in other outcome measures appeared to be related to improvement in process measures.	**Good**

NOTE: Not all of the studies listed in the table were conducted in the context of a P4P experiment; rather, the measures that were the focus of the study are typically found within P4P programs.

[a] DSH hospitals are those that receive compensation through Medicare for treating a disproportionate number of indigent patients.

Results of Performance in Value-Based Purchasing Programs

5. Based on the metrics used to date, have VBP programs facilitated improvements in quality and value?

We identified 50 studies that examined the effect of P4P on performance on clinical quality. In this section, we discuss the findings of studies that addressed performance on clinical processes-of-care, while we address performance on health outcomes and costs in sections 5a and 5b, respectively. We summarized P4P impact studies that examine effects on disparities in care, spillover effects, and unintended consequences under the research questions that focus on these issues. Synthesizing the evidence across these studies was challenging because of the heterogeneous nature of the studies and programs; the studies also used different variables of interest, study periods, incentive structure, and analysis designs. In addition, some of the studies were poorly described, which made it difficult to understand key aspects of the study, such as the methods used and the duration of the intervention. We organize the presentation of findings by setting of care. All of the results listed were significant at $p \leq 0.05$ unless otherwise noted.

Pay-for-Performance Programs Focused on Physicians or Physician Groups

Thirty-nine studies examined the impact on clinical process measures of P4P programs that targeted physicians or physician groups (Table 3.4). The studies evaluated a wide range of P4P programs executed by various sponsors. The researchers who evaluated these programs used a variety of analytic designs of varying methodological rigor. We deemed only seven of the 39 studies to be of good quality, 15 of fair quality, and 17 of poor quality. A number of the poor-quality studies were very small-scale tests of performance incentives, with no comparison groups, short study duration, or which tested an intervention that no one expected to be permanent.

The studies that we deemed as "good" tended to have multiple years of data, focused on large ongoing national or regional efforts, and used methodologies such as difference-in-differences or instrumental variable models to address confounding that might result from unobserved variable bias. The studies with stronger designs found generally modest positive results for treatment, screening, and prevention measures, while one study had a mix of positive and negative results:

- Fagan et al.[40]—Based on two years of data, mixed results were observed in the trends on five incentivized measures between nine physician practices that received incentives from a large national managed care organization and comparison practices. P4P practices had significant improvement compared with non-P4P practices on one measure (influenza vaccine: OR=1.79), had significant reductions on two measures (HbA1c testing: OR= 0.44; LDL screening: OR=0.62), and were no different on one measure (eye exam for diabetes).
- Rosenthal et al.[10]—In a P4P program within PacifiCare, a large health plan in California, cervical cancer screening rates went up significantly for the P4P practices relative to non-

P4P practices by ~4 percentage points over the course of three years. Mammography and HbA1c testing rates were unchanged.

- Mullen et al.[42]—Also in a P4P program sponsored by PacifiCare in California, no improvement was observed on any incentivized measures related to screening (cervical cancer, breast cancer), prevention (childhood immunizations), chronic disease care (HbA1c testing, asthma medication), or appropriate antibiotic usage relative to comparison practices in the Pacific Northwest over a five-year period.
- Chien et al.[69]—No significant improvement was found on any of three diabetes measures (HbA1c, lipid, and dilated eye exam rate) over a five-year period in the New York Medicaid P4P program.
- Chien et al.[22]—Small but statistically *insignificant* improvements (seven percentage points) in immunization rates were observed for the first three years of the New York Medicaid P4P program, but a statistically significant improvement (11 percentage points) was observed using five years of data.
- Bardach et al.[147]—A one year small randomized study of 42 primary care clinics in New York found modest, statistically significant improvements in antithrombotic prescription for patients with diabetes or ischemic vascular disease (12 percent for intervention vs. 6.1 percent for control, p=0.001; blood pressure control (9.7 percent vs. 4.3 percent, p=0.01; smoking cessation interventions (12.4 percent vs. 7.7 percent, p=0.02). No significant difference was found for cholesterol control. Intervention and control groups had subsidized electronic health records (EHRs) and quarterly quality feedback reports.
- Petersen et al.[148]—A randomized trial in Veterans Health Administration hospital-based primary care clinics, which compared the effects of physician-level incentives, practice-level incentives, both, or none. During the 16-month study period, performance improved for the three intervention groups; however, the study found that only physician-level financial incentives resulted in significantly greater blood pressure control or appropriate response to uncontrolled blood pressure compared with the control group. None of the incentives led to greater use of guideline-recommended medication or increased incidence of hypotension compared with controls.

Similarly, studies deemed to be of fair quality generally found positive, although in at least one case mixed, results for diabetes, screening, and prevention measures. Of the fair-quality studies, the vast majority included some type of comparison group, but the studies were of short duration, did not adequately account for unobservable confounding factors, or were limited in sample size or geographic region. The five studies that included diabetes measures were as follows:

- Chen et al.[48]—In the Hawaii Medical Service Association P4P program, P4P practices were significantly more likely than non-P4P practices to deliver all recommended care (HbA1c and LDL testing) (OR=1.2) among patients who saw a P4P providers for three straight years.
- Chen et al.[50]—In the Hawaii Medical Service Association P4P program, P4P practices were significantly more likely than comparison practices to deliver HbA1c screening (ranging from two to seven percentage point improvements) based on four years of data.
- Pearson et al.[5]—P4P was not associated with regular improvements in diabetes scores over a three-year period among five Massachusetts health plans' P4P programs. Of the 15

potential diabetes measures (four different measures across five different P4P programs), three improved significantly and two got significantly worse under P4P.

- Rosenthal et al.[52]—In a cross-sectional comparison of P4P practices and comparison practices using four years of data, P4P practices experienced significantly higher performance on all four diabetes process measures of quality, with the largest differences observed in microalbumin screening (18 percentage points).
- Levin-Scherz et al.[45]—In a P4P program within a large integrated delivery system, P4P practices experienced significant improvement compared with non-P4P practices on four diabetes measures ranging from roughly two to 19 percentage points across a three-year period.

Six other fair studies examined P4P's effects on screening and prevention measures. These studies generally found positive results, although in at least one study the results were mixed:

- Chen et al.[50]—In the Hawaii Medical Service Association P4P program, P4P practices were significantly more likely than comparison practices to deliver cervical cancer screening and varicella vaccinations across four years of reporting (ranging from one to seven percentage points). However, for other screening rates, the results were a mix of positive and negative results. For mammography rates, the improvements were insignificant in year 2 and 4 of the program, while a small significant difference was observed in in the third year (0.8 percentage points). Colorectal cancer screening rates declined significantly in year 2 and three and increased significantly in year 4 (ranging from ~ negative two to positive two percentage points).
- Chung et al.[33]—In an RCT, frequency of bonus payment did not affect delivery of preventive care over a one-year test period. However, despite the strong design (RCT), the study was of short duration, had a relatively small sample number of providers (n=117 physicians in a single center), and was restricted to a single geographic region.
- Fairbrother et al.[23]—In an RCT, P4P practices improved their immunization rates significantly (by five to eight percentage points) compared with comparison practices. Despite the strong design, the study was of short duration (one year), included a small number of providers (60 physicians in nine clinics), and was restricted to single geographic region.
- Pearson et al.[5]—P4P was not associated with regular improvements in scores for breast cancer, cervical cancer, or chlamydia screening over a three-year period among practices exposed to five different Massachusetts health plans' P4P programs. Of 19 potential diabetes measures (seven different measures across five different P4P programs), two measures experienced greater improvements at P4P practices compared with non-P4P practices, whereas two measures experienced greater improvements at the non-P4P practices compared with the P4P practices.
- Gavagan et al.[51]—In a large network of community health centers, there was no evidence for a clinically significant effect of P4P on breast and cervical cancer screening and immunizations.
- Rosenthal et al.[52]—In a cross-sectional comparison of P4P and non-P4P practices using four years of data, P4P practices had significantly better performance on cervical (3.9 percentage points) and breast cancer screening (2.2 percentage points) than non-P4P practices.

One study estimated the effect of receiving recommended care across 11 indicators of screening, other preventive care (e.g., immunizations), and chronic disease care (e.g., diabetes, heart disease, asthma) and estimated the probability of delivering any single recommended care process (rather than look at the results for each measure independently):

- Gilmore et al.[25]—In a P4P program sponsored by the Hawaii Medical Service Association, there was a significant positive association between having seen only P4P program-participating providers and receiving recommended care across the six years (OR: 1.06–1.27).

Finally, one study of fair quality focused on cardiovascular care,[149] one on smoking,[47, 49] one on well-child visits,[44] and one on hypertension,[24] These studies similarly found generally positive effects of P4P on quality.

The studies that we deemed to be of poor quality tended to focus on a small number of physician practices, included no comparison group, or were simply cross-sectional comparisons of P4P participants and nonparticipants in a single year. Many of these studies also consisted of preliminary evaluations or "alpha tests" of P4P concepts rather than evaluations of fully implemented programs. Nearly every poor-quality study found that P4P was significantly associated with higher levels of quality, and many reported substantial effect sizes. Seven of these studies[8, 26–29, 41, 46, 150] included diabetes measures. All of these studies found significant improvements on common diabetes indicators ranging from seven to 45 percentage points over a one- to four-year period. For example:

- Chung et al.[26]—In the Hawaii Medical Service Association P4P program, HbA1c testing increased significantly from 52 percent to 80 percent over four years.
- In a pre-post evaluation of a P4P program at Intermountain Health Care, Larson[151] found that HbA1c testing increased significantly from 79 percent to 91 percent over the course of five years.

Two of the poor-quality studies[30, 31] found that P4P was significantly associated with improvements in documentation, counseling, and referrals related to the use of tobacco products, as follows:

- Amundson et al.[30]—Physician practices exposed to a P4P program sponsored by a large health plan in Minnesota significantly increased the rates at which they provided advice to patients about quitting tobacco use from 32 percent in the pre-period to 53 percent across four years.
- Hung et al.[31]—Smokers in 89 practices participating in a joint Robert Wood Johnson Foundation–AHRQ P4P program were 27 times more likely to be referred to smoking cessation counseling compared with those in comparison practices in a single cross-sectional year.

Finally, two of the poor-quality studies investigated the effect of P4P on screening or related treatment rates; two focused on cancer[32, 33] and two focused on sexually transmitted diseases.[33, 34] These two studies found statistically significant improvement that was small to moderate in size

for screening or medication rates ranging from three to seven percentage points over a one- to three-year time period. The remaining poor-quality studies focused on various clinical conditions (e.g., sinusitis),[35] asthma,[36] depression,[37] and hospital care.[38, 39] All of these found varying degrees of impact on screening and prescribing measures of 20 to 40 percentage points across one to three years.

Strength of Evidence: Low. Although there are a large number of studies that have evaluated the impact of P4P on clinical quality, only seven were of good methodological quality. Across all the studies, findings were generally positive, but among the strongest studies, there were no or relatively small improvements in performance. Studies with the weakest research designs showed consistently significant and large positive effects; however, because these studies relied on cross-sectional data or did not use a comparison group, it is not possible to disentangle any observed improvements due to P4P from secular trends in improvement that were occurring more broadly due to other interventions (e.g., public reporting, QI support). A number of the studies also suffer from being small-scale interventions of short duration that were not intended to continue after the experiment, which might have affected the response to the incentive.

Pay-for-Performance Programs Focused on Hospitals

We found 11 studies that examined the effect of P4P on clinical quality (i.e., process measures) in the hospital setting (Table 3.5). Six of the 10 studies examined the effect of the CMS HQID, of which five were of good methodological quality. We deemed one additional study to be of good quality, which evaluated the impact of a program in Massachusetts that used the same measures and incentive methodology as CMS HQID. All of the results listed were significant at $p \leq 0.05$ unless otherwise noted.

The CMS HQID program was executed in two phases. Phase I spanned Q4 2003 to Q3 2006, while Phase II spanned Q4 2006 to Q3 2009. Different payment models marked the two phases. In Phase I, hospitals were eligible to receive a 2 percent bonus on Medicare reimbursement by performing in the top decile on a composite quality measure for each of the clinical conditions incentivized in the HQID. In Phase 2, hospitals could receive bonuses based both on performance (i.e., attainment) as well as improvement[152] as per the following performance categories:.

- A "Top Performer Award," given to hospitals with scores in the top 20 percent of all HQID hospitals in the current year.
- An "Attainment Award," given to hospitals with composite scores exceeding the median from HQID hospitals for the two years prior.
- An "Improvement Award," given to hospitals scoring above the median of HQID hospitals in the current year and also ranking within the top 20 percent in terms of quality improvement among HQID hospitals.

Of the HQID studies deemed to be of good quality, the findings are generally positive but modest. Two of the good-quality studies to evaluate the first phase of the program are Glickman et al.[53] and Lindenauer et al.[59] Another paper of good quality[54] investigated the extent to which hospitals responded to incentives by working on the "easiest" measures. At the time of these studies, virtually all of the hospitals reimbursed under the Inpatient Prospective Payment System (IPPS) were reporting their data into CMS for the purposes of public reporting of results through the Reporting Hospital Quality Data for Acute Payment Update system. Consequently, it is difficult to separate the effect of P4P from other incentives hospitals faced, namely pay-for-reporting and public-display-of-performance results. The HQID studies found the following:

- Glickman et al.[53] focused on six measures of AMI across the first three years of HQID. The study found a significantly higher rate of improvement for two of the six incentivized measures at P4P hospitals relative to comparison hospitals: aspirin at discharge (OR 1.31 vs. 1.17) and smoking cessation counseling (OR 1.50 vs. 1.28). The study found no significant difference in a composite measure of the six incentivized measures.
- Lindenauer et al.[59] focused on estimating the incremental effect of P4P on performance for measures of AMI, CHF, and pneumonia, as well as an overall composite measure, across the first two years of HQID. When comparing the differences between P4P and pay-for reporting hospitals, the study found that P4P hospitals achieved greater improvement in all the composite process measures, with differences ranging from 4.1 percentage points for pneumonia to 5.2 percentage points for CHF. For the overall composite measure, the difference in the change was 4.3 percentage points. However, when the authors controlled for baseline performance volume, and all hospital characteristics, the effects fell substantially, ranging from 1.9 percentage points (AMI) to 3.5 percentage points (pneumonia). For the overall composite measure, the effect was 3.4 percentage points. The authors also investigated the individual measures that constituted the composites. On these measures, P4P hospitals showed significantly greater improvement relative to comparison hospitals on seven of the 10 individual measures, using the raw comparison in changes. Four of five measures of AMI improved between three and ten percentage points. One of two CHF measures improved by five percentage points. Two of three pneumonia measures improved between four and 10 percentage points.
- Nicholas et al.[54] investigated the extent to which P4P induced hospitals to address measures that were easier to comply with, while ignoring measures that were more difficult to comply with. This is a potential unintended consequence of P4P programs. To do this, they used an expert panel to classify measures of AMI, CHF, and pneumonia care as either "easy" or "hard." They found that P4P hospitals did not improve on the "easy" tasks more than non-P4P for CHF or pneumonia. However, P4P hospitals did improve more on "easy" tasks for AMI compared with non-P4P hospitals by around one percentage point. They found no effect for hard measures.

A study of fair quality by Grossbart,[153] focusing specifically on hospitals with the Catholic Healthcare Partners system, found participating hospitals improved their overall composite scores by 9.3 percentage points versus 6.7 percentage points at comparison hospitals, and participating hospitals improved on CHF scores by 19.2 percentage points compared with 16.7

percentage points in nonparticipating hospitals. There was no significant difference for AMI or pneumonia. However, it is important to note that this study compared only four Catholic Healthcare Partners hospitals that self-selected to participate in HQID with six hospitals in that system that were not participating. The small study size limits the generalizability of this study and the methodology does not adequately control for bias.

Two studies,[56, 90] which we deemed to be of good quality, investigated the effect of CMS HQID across the entire life of the program:

- Werner et al.[56] found that, over the first three years of the HQID, participating hospitals had greater performance on an overall composite measure of AMI, CHF, and pneumonia than hospitals that did not participate. After five years, the two groups' scores were virtually identical.
- Ryan et al.[90] found that, in both phases, P4P hospitals improved more than non-P4P hospitals on all three composite measures of AMI, CHF, and pneumonia care (a difference of one to two percentage points); however, P4P hospitals improved less in phase II than phase I, compared with non-P4P hospitals. The difference was significant for CHF and pneumonia, but not AMI.

Another study by Ryan and Blustein,[55] deemed to be of good quality, evaluated the Massachusetts Medicaid P4P program and found no effect of P4P for pneumonia or surgical infection prevention in the two years after the onset of the program.

Two other studies in the hospital setting that were unrelated to HQID[45, 57, 60] were deemed to be of fair quality:

- Calikoglu et al.[57]—Hospitals in Maryland were exposed to a state-run P4P program and experienced improvement in only one of 19 process measures (influenza vaccine), which increased by roughly five percentage points more than the national trend from 2009–2011.
- Herrin et al.[60]—In this study, hospitals in the Baylor Health Care System provided financial incentives to administrators for improving quality. Hospitals increased their compliance significantly faster than comparison hospitals on two of seven measures over four years: aspirin at discharge (OR=2.94) and pneumonia vaccination (OR=1.53).

We classified two studies as having poor design, and these[154–156] investigated the effect of P4P on hospital quality in the context of private health plans or delivery systems. These studies tended to lack a comparison group, be based solely on cross-sectional comparisons across hospitals, or be a single institution case study. The results from these studies were generally positive. These studies found:

- Atkinson et al.[154]—In a single integrated delivery system in New York state, an overall composite measure of quality showed a steady increase over time from 78 percent in the first quarter of 2004 to 93 percent in the first quarter of 2008.
- Atkinson et al., Berthiaume et al.[154, 156]—Within a P4P program executed by the Hawaii Medical Service Association, four of 13 hospitals attained 85 percent adherence to the Get with the Guidelines–Coronary Artery Disease (GWTG-CAD) performance measures in a single cross-sectional examination (one year).

Strength of Evidence: Low. All of the studies focused on P4P in the hospital setting found modest but often statistically insignificant effects, regardless of methodological quality. Because most of the studies of P4P in the hospital setting are of a single intervention (i.e., Premier HQID), it is unknown what effects hospital P4P might have under different design structures.

P4P Programs in Other Settings

We found only one study that evaluated the effect of P4P on clinical quality (i.e., process measures) for settings other than hospitals or physician groups. This study, which we deemed to be of fair quality, evaluated P4P in the substance abuse setting. In an RCT, the study looked at the effect of providing $100 to addiction counselors for every patient that attended at least five treatment sessions.[61] Over a two-year period, the program was associated with a significant increase in the proportion of patients completing five treatment sessions.

Strength of Evidence: Insufficient. There is a lack of evidence regarding the use of P4P in other health settings to say what the impacts might be.

5a. What improvements in health outcomes attributable to VBP can we expect, and over what time horizon?

The majority of P4P impact studies investigated the effect of P4P on clinical process-of-care measures; only a small number of studies have investigated the effect of P4P on outcomes. The studies provide very little information related to what we might expect regarding the impact of P4P on health outcomes and the time horizon within which we might expect to see that impact. These studies focused on a small number of measures for which an effect could reasonably be observed in a short time horizon (e.g., intermediate outcomes rather than long-term health outcomes). Intermediate outcomes are important markers or predictors of long-term and health outcomes (e.g., readmissions, hospitalizations, mortality, stroke, AMI, foot amputations). Below, we summarize the findings from the literature on the effect of P4P on measures of health outcomes in P4P programs. All of the results listed were significant at $p \leq 0.05$ unless otherwise noted.

P4P Programs Focused on Physicians or Physician Groups

We found 11 studies that examined the impact of P4P on outcomes related to physicians or physician groups. Most of the studies reporting effects on outcomes focused on intermediate outcomes related to diabetes (e.g., HbA1c and LDL levels). Only one of the 12 studies was rated as good quality:

- Chien et al.[69]—In a New York Medicaid plan-sponsored P4P program, changes in the percentage of patients with LDL control as well as changes in emergency department use and hospitalizations were not significantly different than comparison practices over a five-year period.

Four fair- or poor-quality studies also focused on intermediate outcomes for diabetes:

- Lester et al.[46]—In a P4P program within Kaiser Permanente in California, HbA1c control improved (47 percent to 70 percent) during the ten-year period (no p-value reported).
- Coleman et al.[27]—No significant improvement in HbA1c control was observed in a P4P program in a large network of community health centers over a single intervention year.
- Larsen et al.[29]—In a P4P program in Intermountain Health Care, the percentage of diabetes patients with HbA1c <7.0 increased and those with an HbA1c score >9.5 decreased, while the average HbA1c scores went down from 8.1 to 7.3 over a five-year period. Additionally, the percentage of patients with LDL<130 mh/dl increased (no p-values reported).
- Chung et al.[33]—In three medical groups in California, the proportion of patients whose blood sugar, blood pressure, and lipid levels were under control improved by two to four percentage points across one year.

One other study investigated the effect of P4P on smoking quit rates, and one studied depression:

- Roski et al.[47]—In an RCT, the smoking quit rate and sustained abstinence was 22.4 percent for patients in the P4P group and 19.2 percent for patients in the control group over one year. However, this difference was not statistically significant.
- Unutzer et al.[37]—The hazard ratio for achieving depression treatment response was 1.73 among 29 integrated behavioral health care clinics two years post P4P program intervention compared with pre-program implementation; meaning that patients were 73 percent more likely to respond to treatment in the post period compared with the pre period. The study was a pre-post examination, with no comparison group.

Finally, four studies focused on long-term or final health outcomes. Only one of the studies was of good quality:

- Rosenthal et al.[70]—A P4P program targeted at pregnant members of a union health plan and their prenatal care providers found a significant reduction in the odds (0.45) of neonatal intensive care unit (NICU) admissions but no significant reduction in low birth weight.

We rated the other three studies as fair or poor:

- Serumaga et al.[24]—In a UK P4P program, no effect was observed on the incidence of stroke, AMI, renal failure, CHF, or all-cause mortality over an eight-year period.
- Leitman et al.[39]—In a P4P program executed within a single large medical center, P4P was associated with no measurable change in 30-day mortality or readmission over four years.
- Chen et al.[48]—In a P4P program sponsored by the Hawaii Medical Service Association, patients were 25 percent ($p<.05$) less likely to be hospitalized if they were continuously attributable to a P4P provider for the entire three years of the intervention.

Strength of Evidence: Insufficient. There is a lack of evidence regarding the effect of P4P in physician practices related to health outcomes.

P4P Programs Focused on Hospitals

We identified six studies of the impact of hospital P4P programs on measures of clinical outcomes. Five of the six studies assessed mortality (either inpatient or 30-day) and one used quality-adjusted life years. We categorized three of the six studies to be of good methodological quality:

- Glickman et al.[53]—There was no evidence that in-hospital mortality improvements were greater at P4P hospitals compared with hospitals exposed only to public reporting using four years of data.
- Sutton et al.[72]—Risk-adjusted mortality for the conditions included in the P4P program decreased significantly compared with hospitals that were not exposed to P4P (1.3 percentage points) 18 months after program introduction. This study focused on a P4P program in the UK that was modeled after the CMS HQID.
- Ryan[71]—There was no evidence that P4P had a significant effect on risk-adjusted 30-day mortality for AMI, CHF, pneumonia, or CABG using seven years of data.

The three studies that we deemed to be of fair or poor quality found:

- Herrin et al.[60]—In a P4P program that provided financial incentives to administrators in the Baylor Health Care System for improving quality, no significant difference was observed over a four-year period in in-hospital mortality between P4P hospitals and a random selection of non-Baylor hospitals reporting to the Joint Commission.
- Jha et al.[73]—There was no evidence that HQID led to a decrease in 30-day mortality using seven years of data.
- Nahra et al.[157]—Over a three-year period, a P4P program administered by a single health plan in Michigan led to improvements in quality-adjusted life years of between 733.3 and 1,701.2. However, the estimate of program benefit was calculated without a comparison group.

Strength of Evidence: Insufficient. There is a lack of evidence regarding the effect of P4P in hospitals related to health outcomes.

P4P Programs Focused on Other Settings

One study,[74] which we rated as good, evaluated five states' Medicaid nursing home P4P programs and found that three of six outcome measures (the percentage of residents who were physically restrained, in moderate to severe pain, and developed pressure sores) improved in P4P sites between 0.3–0.5 percentage points relative to comparison sites one year post program implementation. Other incentivized quality measures either did not change or worsened. The small improvements were based on very low baseline rates ranging between nine and 12 percent, and the authors commented that these measures might be difficult to improve.

We also reviewed two studies of fair quality. Hittle et al.[75] found that only two measures (improvement in pain interfering with activity and improvement in urinary incontinence), which were both non-incentivized, showed significant differences between P4P and comparison home health agencies across one intervention year. Shen[76] found three years post intervention that P4P

in substance abuse clinics was associated with a reduction in the proportion of clients classified as most severely ill.

Strength of Evidence: Insufficient. There is a lack of evidence regarding the use of P4P in other health settings to say what the impacts might be.

Conclusion

Only a small number of studies investigated P4P's effect on measures of clinical outcomes, and these studies found modest positive results. However, the results were generally insignificant in the highest quality studies. The studies focused on a relatively small number of outcome measures; consequently, it is unknown what P4P's effects might be for other outcome measures especially long-term outcomes. The selection of intermediate outcomes as the focus of the P4P incentive is a function of the program sponsor's ability to observe the outcome within a proximal period of time.

5b. What cost savings attributable to VBP can we expect, and over what time horizon?

Few studies have examined the impact of P4P on costs. Unfortunately, these studies provide very little information on what effects we might expect as the studies were of variable quality and found generally positive results, but the highest-quality studies found modest or statistically insignificant results.

P4P Programs Focused on Physicians or Physician Groups

Four studies evaluated the effect of P4P on costs in the physician or physician group setting. Because these studies are small in number and of relatively low quality, this literature provides little guidance on the potential systematic effect on costs that might be expected as the result of P4P programs.

Two studies that we rated as poor found significant cost savings in P4P programs. Both of these studies were simple pre-post studies with no comparison group or did not include adequate controls for confounding factors.

- Curtin et al.[3]—In a P4P program between Excellus health plan and the Rochester Independent Practice Association, the program resulted in a return on investment of 1.6:1 in the first year and 2.5:1 in the second year based on cost trend estimates related to diabetes care.
- Leitman et al.[39]—A P4P program at Beth Israel Medical Center paid physicians based on their performance on over 20 measure of inpatient quality. The study found that the program led to $7 million in cost savings over a four-year period. These savings may have been driven by a gain-sharing component that was incorporated into the program.

Two studies were of good methodological quality:

- Rosenthal et al.[70]—A P4P program targeted at pregnant members of a union health plan and their prenatal care providers led to lower spending (around $235) in the first year of life over three intervention years.

- Fagan et al.[40]—In a P4P program sponsored by a large managed care plan, no significant differences were observed between P4P and comparison practices in the average total medical cost trends for patients with diabetes over a two-year period.

Based on our environmental scan of P4P programs, we found one estimate of return on investment. A preliminary internal assessment of four of United Healthcare's P4P pilots that were based on a PCMH model showed gross savings on medical costs of 4.0 to 4.5 percent per year for two years. After calculating the additional cost for care coordination and bonuses to the practices, net savings averaged about 2 percent for a 2:1 return on investment (UnitedHealth Group, 2012).

Strength of Evidence: Insufficient. There is a lack of evidence regarding the effect of P4P in physician practices related to costs.

P4P Programs Focused on Hospital Groups

Two studies examined the effect of P4P on costs in the hospital setting. Both studies were based on the CMS HQID program and of good methodological quality:

- Ryan[71]—The change in risk-adjusted costs was not significantly different between the P4P and comparison hospitals using seven years of data.
- Kruse et al.[77]—There was no significant effect of P4P on hospital revenues, costs, and margins or Medicare payments (index hospitalization and one year after admission) for AMI patients using three years of data.

Strength of Evidence: Insufficient. There is a lack of evidence regarding the effect of P4P hospitals related to costs.

Conclusion

The studies with the strongest designs report that there is little to no effect on costs. However, it is reasonable to assume that any substantial reductions in costs cannot be observed in the short time period of the studies, especially for physician- or physician-group-based evaluations, which often focused on chronic diseases, for which the cost implications of better disease management could take years to observe; however, a longer time period for observing outcomes presents the opportunity for other influences to effect the outcomes, making it more difficult to isolate P4P effects. Also, most P4P programs focused on reducing the underuse rather than the overuse of health services, and increased costs are associated with provision of these services.

One could potentially expect to observe short-term improvements in in-hospital costs, but such cost reductions were not observed in two studies of relatively good methodological quality. These studies do not provide sufficient evidence on what the effect of P4P is on costs in the inpatient setting. Because these studies focus on a single program (CMS HQID), it is difficult to generalize to other programs. Additionally, changing performance on a different set of measures might lead to different conclusions about effects on costs. Further evidence is likely to change the estimates and our confidence in those estimates. Impact studies that focus on how and under

what circumstances P4P could contribute to reductions in costs would be a valuable contribution to the literature. This information will likely require evaluations that have even more extended observation periods than what is presently available, particularly in the physician and physician group setting.

Table 3.4. Evidence on Effectiveness of Physician and Physician Group Pay-for-Performance Programs

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Amundson et al., 2003[30]	Health Partners P4P focused on tobacco Ask and Advice rates from 1996 to 1999	Longitudinal study of participants	Bonus pool	**Process:** Documentation and discussion of tobacco use	**Process:** Mean ask rate increased from 49% to 73% Advise rate increased from 32% to 53%	**Poor:** Regional population, no modeling to control for confounders
An et al., 2008[49]	Collaborative project between Fairview Physician Associates and multiple Minnesota health plans to encourage referrals to health plan sponsored quit line from 2005 to 2006	RCT of usual care vs. P4P for quit line referrals	Clinic receives $5,000 for 50 quit line referrals	**Process:** Rates of referral; contact and enrollment after referral; and project costs	**Process:** 11.4% of smokers were referred in P4P group compared with 4.2% in the control group (p=0.001)	**Fair**
Armour et al., 2004[32]	Large managed care health plan operating in the southeastern United States implemented a year-end bonus program that was designed, in part, to improve colorectal cancer screening use among an individual practice association's PCPs from a 10-month period across 2001–2002	Pre-post study of P4P cohort	Bonus payment	**Process:** Colorectal cancer screening	**Process:** From 2000 to 2001, colorectal cancer screening use increased from 23.4% to 26.4% (p< 0.01).	**Poor:** Short study period, cross-sectional with limited controls

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Bardach et al., 2013[147]	P4P experiment between April 2009 and March 2010 among small primary care practices (<10 physicians) in New York City. In addition to financial incentives, clinics were provided with EHR software with decision-support and patient registry functions and QI specialists that offered technical assistance.	Cluster-RCT, 84 small primary care practices. Intervention received incentives and quarterly performance reports, while control received only performance reports. One-year evaluation.	Incentive paid to the clinic/practice. Incentive paid for every instance of patient meeting the quality criteria. Higher incentive payments given for patients who were sicker, had Medicaid insurance or were uninsured. Bonuses were a maximum of $200/patient and $100,000/clinic Range of payments was to clinics was $600–$100,000 (median $9,900).	**Process:** Aspirin or anthrombotic prescription Smoking cessation **Outcomes:** Blood pressure control Cholesterol control	**Process:** Adjusted change in performance significantly higher in the intervention group than controls for aspirin or antithrombotic prescription by 6.0% (p=0.001)for patients with ischemic vascular disease or diabetes **Outcomes:** Adjusted change in blood pressure control significantly higher in the intervention group than control by • 5.5% (p=0.01) among patients with only hypertension • 7.8% among patients with hypertension and diabetes • 7.8% (p=0.01) for patients with hypertension, diabetes and ischemic vascular disease No difference in cholesterol control (p=0.22) Changes were higher for uninsured or Medicaid patients in intervention clinics compared with controls, except for cholesterol control.	**Good:** Randomized study design, although short study duration. Findings may not generalizable to larger practices or those without EHRs or QI assistance.

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Beaulieu and Horrigan 2005[41]	In 2001, a managed care organization in upstate New York designed and implemented a pilot program to financially reward doctors for the quality of care delivered to diabetic patients across an 8-month period.	Pre-post with comparison group	Incentive payment equivalent to a 12% increase in PMPM reimbursement if performance goals are met	**Process:** 6 measures of diabetes care quality **Outcome:** 3 diabetes outcome measure	**Process:** Physicians and patients achieved significant improvement on five out of six process measures. **Outcome:** Physicians and patients achieved significant improvement on two out of three outcome measures (HbA1c control and LDL control).	**Poor:** Small number of study participants (n= 17 physicians). Physicians self-selected; one small region, short duration, physicians not matched at baseline. Comparison patients had higher baseline performance on all measures
Chen et al., 2010a[50]	P4P program initiated by preferred provider organization (PPO) in Hawaii from 1998 to 2007	Compared pre-post changes of intervention group to comparison group in a different state	Additional 1.5–7.5% of base salary to perform processes of care	**Process:** ACE inhibitor use among CHF patients, mammography, cervical cancer screening, colorectal cancer screening, HbA1c testing for diabetes, the varicella vaccine, and the measles, mumps, rubella (MMR) vaccine	**Process:** P4P group had significantly greater increases in quality scores than the comparison group for cervical cancer screening and HbA1c testing. P4P group had significantly greater increases than the non-P4P group in quality scores for mammography and varicella for the 2nd to 3rd year. P4P group improved less than the non-P4P group for colorectal cancer screening every year, except from the 3rd to the 4th year	**Fair**

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Chen et al., 2010b[48]	PPO in Hawaii provided incentives to physician to improve quality and reduce hospitalizations from 1999 to 2006	Longitudinal study comparing participating practices with nonparticipating practices	1.5–7.5% of base salary to perform processes of care	**Process:** Diabetes processes of care **Outcome:** Hospitalizations	**Process:** Improved diabetes quality care compared with non-P4P participating physicians among patients who saw p4P providers throughout entire study period (OR=1.20; 95% CI, 1.05–1.37, p<0.01). Reduction in hospitalization for patients who saw p4P providers throughout entire study period	**Fair**
Chen et al., 2011[149]	Health plan in Hawaii incentivizes participating physicians additional payments to improve 2 cardiovascular disease quality measures from 2000 to 2006	Longitudinal multivariate regression models comparing participants to nonparticipants	Bonus of 3.5% of professional fees	**Process:** LDL testing, statin prescribing	**Process:** P4P group improved (32%–70%) compared with non-P4P group (40%–61%) on quality composite	**Fair**
Chien et al., 2010[22]	New York Medicaid nonprofit plan implemented a P4P program that incentivized immunization delivery to 2-year-olds from 2003 to 2007	Difference-in-differences comparing participants and nonparticipants pre-post	$200 bonus payment for each fully immunized 2-year-old	**Process:** 2-year old immunizations	**Process:** Immunization rates within Hudson Health Plan rose at a significantly, albeit modestly, higher rate than the robust secular trend noted among comparison health plans.	**Good:** Regional but multiple years of observation and strong difference and difference design

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Chien et al., 2012b[69]	New York Medicaid nonprofit plan implemented a P4P program that incentivized improvements in diabetes care and outcomes in 2003–2007	Difference-in-differences comparing participants and nonparticipants pre-post	$100–$300 bonus payments for each patient completing all the missing care processes	**Process:** Diabetes quality measures (HbA1c testing, lipid testing, dilated eye exams, lipid control) **Outcome:** Diabetes outcome measures (e.g., BP and HbA1c and LDL levels)	**Process:** Between pre- and post-intervention periods, changes on available diabetes measures were not statistically significant **Outcome:** Changes in diabetes outcome measures were not statistically significant when compared with non-Hudson plans	**Good:** Regional but multiple years of observation and strong difference and difference design
Chung et al., 2003[26]	Voluntary P4P program implemented by a health plan in Hawaii from 1997 to 2000.	Time trend of participants	3.5% above base fees	**Process:** Use of ACE inhibitors or angiotensin receptor blockers in CHF, measurement of HbA1c in diabetes, and rates of childhood immunizations	**Process:** ACE inhibitor rate increased from 40.8 to 64.2% for CHF patients ($p<0.001$) HbA1c testing increased from 51.5 to 79.6% ($P<0.0001$) MMR immunization rates varied and no consistent tend could be identified	**Poor:** No contemporaneous control group, case study only
Chung et al., 2010a[103]	RCT of the effects of the frequency of a P4P bonus on performance in Palo Alto Medical Foundation over the course of a 1-year study period.	RCT	Bonus payment of up to 2% of base salary	**Process:** Six process measures (prescription of asthma controller, cervical cancer screening, chlamydia screening, colon cancer screening, whether the height and weight were measured and recorded, and documentation of tobacco use history) **Outcome:** 3 outcome measures for diabetes control (BP 130/80mmHg, HbA1co7%, and LDLo100 mg/dL)	**Process:** Frequency of bonus payment did not affect process or outcome measures.	**Fair**

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Chung et al., 2010b[33]	P4P program within single clinic in California from 2005 to 2007	Pre-post comparison of participants	Bonus payment of up to 2% of base salary	**Process:** 5 measures related to screening, asthma medication prescribing, and prevention	**Process:** From 2006 to 2007, 8 of 9 incentivized and previously reported measures showed significant improvement (mix of process and outcome measures)	**Poor:** Single practice no comparison group
Coleman et al., 2007[27]	A large federally qualified health center implemented incentives for absolute performance and improvement on process and outcome measures in 2004.	Pre-post comparison of single practice	Reduction in base salary couple with bonus payments for meeting productivity goals	**Process:** Avg. annual # of encounters per diabetic patient, % diabetic patients with any HbA1c test, **Outcome:** % diabetic patients with recommended number of HbA1c tests, % diabetic patients with controlled blood sugar (HbA1c <7, HbA1c<9).	**Process:** From 2003 (pre-P4P) to 2004 (1st year P4P), significant increase (16.2%) in biannual HbA1c testing for diabetic patients (p<0.001) **Outcome:** No significant improvement in blood sugar control (HbA1c< 7 or HbA1c <9) in ACCESS patients or Medicaid patients from NCQA dataset (OLS p=.1639)	**Poor:** Single organization, no comparison group, and relatively short time frame
Collier, 2007[38]	A community health care system implemented a P4P program for 12 hospitalists on a range of structural, process, and utilization measures from 2003 to 2006	Pre-post comparing participants to nonparticipants	Bonus	**Structure:** 24/7 access to care, maintaining at most an 18:1 physician to patient ratio, dictating medical records within 12 hours and providing discharge summaries within 24 hours, attending monthly hospital meetings, and having membership in the Society of Hospitalists **Process:** CMS/Joint Commission process measures	**Structure:** Almost all of the measures were accomplished **Process:** Although the contracted group did not consistently meet all Joint Commission/CMS targets, compliance with most quality indicators improved to a greater extent than a concurrent non-contracted group.	**Poor:** Only a single organization, and analytic methods poorly explained

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Curtin et al., 2006[3]	P4P program that was a 5-year partnership (2000–2004) between Excellus health plan and a Rochester, New York, independent practice association	Pre-post cost analysis focused on return on investment	10% salary withhold returned when goals are met	**Costs:** Costs PMPM Return on investment	**Costs:** Positive return on investment of 1.6:1.0 in 2003 and 2.5:1.0 in 2004	**Poor:** Single entity and "benefit" measured simply as pre-post comparison. Little analytic work to deal with confounding factors.
Cutler et al., 2007[28]	IHA program is a state-wide P4P program providing physician groups with bonuses for meeting patient experience, process, and outcome measure. This study focuses on Mercy Medical Group.	Cross sectional (2004) comparison of participants and nonparticipants	Bonus above base PMPM capitation payment	**Process:** LDL testing and control for patients with diabetes	**Process:** Higher proportion of patients in P4P group who attained LDL-C goal (<130 mg per dL) those in the routine care (78.2% vs. 55.7%, p<.001). Higher rate of achieving a LDL-C <100 mg per dL than those in the routine care group (46.7% vs. 35.2%, p =.004)	**Poor:** Short study period, cross-sectional, no controls for confounding factors.
Fagan et al., 2010[40]	Intervention by national managed care organization to provide P4P bonus payments to 9 PCP practices for meeting quality of care measures	Longitudinal (2004–2006) study in which pre- and post-data from intervention compared with comparison practices	Bonus payment up to 20% of the capitation fee for Medicare managed care organization patients	**Process:** 5 incentivized quality measures (influenza vaccine, HbA1c testing, eye exam, LDL screening, and nephropathy screening), 2 non-incentivized measures (avoiding short-acting antihypertensive and prescribing an ACE/ angiotensin receptor blocker medication for diabetics with renal insufficiency) **Costs:** Emergency department utilization, and total paid costs	**Process:** Quality of care generally improved for both groups during the study period. Only slight differences were seen between the intervention and comparison group trends and changes in trends over time. **Costs:** No significant differences were observed in the average total medical cost trends per member per month (p=.42) between P4P and non-P4P members with diabetes from baseline to follow-up	**Good:** Relatively large region, difference-difference design to control for time invariant confounders.

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Fairbrother et al., 2001[23]	RCT of 57 inner-city physicians randomized to a P4P bonus, enhanced-FFS, or control group in 1997–1998	RCT	$1,000–$7,500 bonus depending on improvement level	**Process:** Up-to-date immunization coverage	**Process:** Both the bonus and the enhanced FFS groups improved significantly in documented up-to-date immunization status (Bonus: 49.7 to 55,6%, p<0.05; Enhanced FFS: 50.8 to 58.2%, p<0.01) compared with the control group. Steady increases, but no significant difference in number of well child visits. Improvement was due primarily to improved documentation rather than actual vaccines given. Missed opportunities (when vaccines were due but not given) did not change.	**Fair**
Felt-Lisk et al., 2007[44]	5 Medicaid health plans that implemented P4P programs from 2002 to 2005	Pre-post changes in participants with a limited comparison to national trends	Bonus payments based on the number of patients receiving well-baby visits	**Process:** % of plan members with 6 or more well-baby visits by age 15 months	**Process:** From pre-implementation (2002 to 2003) to post implementation (2004 to 2005), 2-year average HEDIS scores improved 7.5–27 percentage points. Large effects not seen in 4 of 5 plans.	**Fair**
Gavagan, et al., 2010[51]	Rewarding Results Collaborative Demonstration: Physicians at 6 of 11 clinics were given incentives for achieving group targets in preventive care.	Longitudinal analysis with comparison group	$4,000–$12,000 bonus payment depending on performance	**Process:** Preventive care (cervical cancer screening, mammography, pediatric immunization)	**Process:** Found no evidence for a clinically significant effect of financial incentives on performance of preventive care	**Fair**

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Gilmore et al., 2007[25]	P4P program providing bonuses to individual physicians for absolute performance on patient experience, structural, quality and practice pattern measure from 1998 to 2003	Compared changes over time between participating physicians and nonparticipating physicians	Bonus of 1%–5% of base professional fees	**Process:** 11 process measures related to screening, care for diabetes, hypertension, asthma, CHF, and high cholesterol, prevention	**Process**: Positive association between having seen only program-participating providers and receiving recommended care for all 6 years recommended care for all 6 years (OR: 1.09, 95%: 1.072–1.10).	**Fair**
Greene et al., 2004[35]	Large, multifaceted QI intervention consisting of physician education, profiling, and a financial incentive, to improve treatment quality for acute sinusitis in Rochester from 1999 to 2001	Pre-post no comparison group	15% payment withhold returned based on performance	**Process:** Overall exceptions per 1,000 episodes, acute sinusitis care pathway exceptions per 1,000 episodes, services per 1,000 episodes of acute sinusitis	**Process:** A statistical process control chart showed a shift toward recommended treatment patterns after our intervention.	**Poor:** No comparison group and no apparent controls for confounding factors.
Hung and Green 2012[31]	AHRQ health promotion initiative offering incentives to PCPs to improve on smoking cessation measures	Cross-sectional comparison of participants and nonparticipants	Unclear	**Process:** Smoking cessation counseling, linking patients to smoking cessation services in community	**Process:** Practices that were involved with P4P had greater odds of offering recommended cessation counseling (OR= 27.6, p <0.01)	**Poor:** Single year, small sample size, and limited controls for confounding factors.

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Larsen et al., 2003[29]	Health care system implemented a multi-faceted diabetes care program, which included financial incentives for individual physicians for diabetes QI from 1998 to 2002	Longitudinal analysis no comparison group	Bonus of 0.5% to 1% of total physician compensation	**Process:** Rates of testing of HbA1c and LDL, rate of annual eye exams, **Outcome:** LDL and HbA1c values	**Process:** HbA1c test increased from 78.5% in 1998 to 90.5% in 2002. LDL cholesterol screening test within the prior 2 years increased from 65.9% in 1998 to 91.7% in 2002. Annual eye exam increased from 52% in 1998 to 62% in 2002. **Outcome:** % with HbA1c less than 7.0 increased from 33.5% in 1998 to 52.8% in 2002. Average HbA1c decreased from 8.1 in 1998 to 7.3 in 2002. % with HbA1c greater than 9.5 decreased from 34.6% in 1998 to 21.4% in 2002. % with LDL cholesterol was less than 130 mg/dL increased from 39.9% in 1998 to 69.8% in 2002.	**Poor:** Single system, no comparison group, no controls for confounders.
Leitman et al., 2010[39]	Beth Israel Medical Center implemented a P4P and shared savings program for individual physicians using patient experience, patient safety, process, outcome, and efficiency measures between 2006 and 2009.	Pre-post analysis comparing participating and nonparticipating physicians	Gainshare	**Cost:** Cost-savings, average LOS, **Process:** Quality measures for AMI, CHF, pneumonia **Outcome:** 30-day mortality or readmission	**Cost:** $7 million savings **Process:** Change in quality measures not statistically significant **Outcomes**: No measurable change in 30-day mortality or readmission	**Poor:** Single system, compared participating physicians with nonparticipating physicians, with unclear controls for confounding factors.

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Lester et al., 2010[46]	35 medical facilities participating in a P4P program through Kaiser Permanente Northern California from 1997 to 2007.	Longitudinal analysis of participants including removal of incentives	Bonus	**Process:** Screening for diabetic retinopathy, cervical cancer **Outcome:** Control of hypertension (systolic blood pressure <140 mm Hg), Glycemic control (HbA1c <8%)	**Process**: Removing incentives for diabetic retinopathy screening declined on average by approx. 3% per year (mean change 3.1%, 95% CI, 2.4% to 3.8%) and cervical cancer screening by an average of approx. 2% per year (mean 1.6%, 95% CI, 1.1% to 2.1%) **Outcome:** Hypertensive adults whose systolic BP was less than 140 mm Hg increased (58.3%to 78.2%). Glycemic control was incentivize and performance improved from 47% to 69.8%	**Poor:** Pre-post only within a single system.

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Levin-Scherz et al., 2006[45]	Large, heterogeneous integrated delivery network that incorporated physician quality, efficiency, and structural metrics into P4P contract	Longitudinal analysis (2001–2003) comparing to state and national trends	Contracts included some element of withhold, often approximately 10% of hospital and/or physician fees. Some included an opportunity for bonus payments beyond the agreed-upon fee schedule. Withholds were returned or bonuses earned depending on regional service organization and Partners Community HealthCare, Inc.(PCHI) network performance compared with previously agreed targets	**Process:** Performance on adult diabetes and pediatric asthma HEDIS measures	**Process:** HbA1c : Participants improved significantly greater than the statewide improvement rate on (7.0 vs. 4.9 percentage points, p < .05). Diabetic eye exams: participants performance improved, while statewide performance declined slightly (18.7 vs. –0.8 percentage points, p <0 .05). Diabetic LDL screening: Participants' performance improved by almost twice as much as the state average (13.2 vs. 7.4, p < .05). Nephropathy screening: Participant rates improved over twice as much as statewide improvement (15.2 vs. 12.9 percentage points, p<0.05). All four diabetes measures: PCHI's 1st P4P plan achieved significant improvements on all 4 diabetes measures compared with national trends (p<0.05). Pediatric asthma controller: Performance improved more than the state average on every measure except pediatric asthma controller use (1.7 vs. 3.9 percentage points, p >0.05).	**Fair**

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Mandel and Kotagal 2007[36]	54 pediatric practices in the greater Cincinnati area were involved in a P4P program that rewarded practices for participating in the collaborative, achieving network- and practice-level performance thresholds, and building improvement capability related to asthma from 2003 to 2006.	Longitudinal analysis (interrupted time series) with no comparison group	% of base pay based on reporting, network performance, and practice performance	**Process:** Medication control, flu shots, and written self-management plans	**Process**: % of the network asthma population receiving "perfect care" increased from 4% to 88%. %of the network asthma population receiving the influenza vaccine increased from 22% to 41%,	**Poor:** Analytic methods insufficiently explained to make strong determination.
Mullen et al., 2010[42]	PacifiCare implemented a QI program in California in conjunction with the IHA P4P program. Study analyzed effects of implementing both programs on incentivized and non-incentivized measures from 2001 to 2005.	Difference-in-differences	Bonus payment of $500–$5,000 based on performance	**Process:** Measures related to screening, diabetes, and prevention	**Process:** Fail to find evidence that initiative either resulted in major improvement in quality or notable disruption in care	**Good:** Regional intervention but strong design with difference-in-differences approach and multiple years of data.
Pearson et al., 2008[5]	P4P programs introduced into physician group contracts from 2001–2003 by 5 major commercial health plans in Massachusetts	Pre-post analysis with comparison group	Combination of bonuses and withholds ranging from $200 to a high of approximately $2,500 per PCP	**Process:** Measures related to process measures related to screening, diabetes, and prevention	**Process:** Not associated with greater improvement in quality compared with a rising secular trend	**Fair**

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Petersen et al., 2013[148]	RCT of P4P incentives among Virginia primary care practices for care (n=83 physicians and 42 non-physicians in 12 study sites) provided to hypertensive patients. Sites were randomized into 4 groups: (1) individual clinician-level incentives, (2) practice-level incentives, (3) combined-level incentives, and (4) no incentives. Participants were provided with educational webinars regarding treatment guidelines, and customized audit and feedback reports for 16 months starting in April 2008.	RCT with time trended analysis	Bonus payments Mean payment of $4,270 in combined group, $2,672 in individual group, and $1,648 in practice group	**Process:** Use of recommended antihypertensive medications or any medication management (start a medication, add a medication, or dose adjustment) **Outcomes:** Blood pressure control or appropriate response to uncontrolled blood pressure	**Process:** While guideline-recommended medication increased significantly during 16-month period, there was no significant change compared with controls.' Difference in proportion of patients receiving any medication adjustment among the individual-level physician group compared with the control group was 15.36% (p=0.05) **Outcomes:** Adjusted absolute difference of 8.36% difference in proportion of patients achieving BP control or receiving appropriate response between individual incentive group and controls (p=.005) Follow-up for 12 months after the end of the incentive found that performance gains were not sustained and declined substantially, though not back to pre-intervention levels	**Good:** RCT with strong post hoc analysis to validate results. 16-month intervention period; small number of clinic sites.

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Pourat et al., 2005[34]	Studies financial incentives and sexually transmitted disease services in in a cross-sectional sample of PCPs contracted with Medicaid managed care organizations in 2002 in 8 California counties	Cross-sectional comparison using regression	Presence of unspecified financial incentives from physician surveys	**Process:** Five measures of sexually transmitted disease	**Process:** Physicians reimbursed with capitation and a financial incentive for management of utilization (odds ratio [OR] = 1.63) or salary and a financial incentive for management of utilization (OR = 2.63) were more likely than those reimbursed under other methods to prescribe chlamydia drugs for the partner. PCPs least often reported they annually screened females aged 15–19 years for chlamydia (OR = 0.63) if reimbursed under salary and a financial incentive for productivity, or screened females aged 20–25 years (OR = 0.43) if reimbursed under salary and a financial incentive for financial performance	**Poor:** Simple cross-sectional associations.
Rosenthal et al., 2005[10]	PacifiCare implemented a P4P program in California, incentivizing patient experience and process measure from 2001 to 2004.	Difference-in-differences comparing participants in California to nonparticipants in the Pacific Northwest	$0.23 per member per month for each performance target that was met or exceeded.	**Process:** Cervical cancer screening, mammography, and HbA1c testing	**Process:** Significant improvement in cervical cancer screening relative to the control group (3.6%). No significant improvement on mammography (p=0.13) and hemoglobin A1c testing (p=0.50).	**Good:** Regional intervention but strong design with difference-in-differences approach and multiple years of data

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Rosenthal, 2008[52]	Bridges to Excellence was first implemented in Massachusetts in 2003, with 2 major physician reward components: the Physician Office Link and the Diabetes Care Link.	Cross sectional comparison of non-recognized physicians in Massachusetts.	Up to $50 for each patient covered by a participating employer	**Process:** Process measures related to diabetes and preventive care. **Utilization:** Patient resource use, number of episodes per patient and the total resource use per episode	**Process:** In one cohort, better performance on measures of cervical cancer screening, mammography, and glycolated hemoglobin testing. In the other cohort, significantly better performance on all 4 diabetes process measures of quality, with the largest differences observed in microalbumin screening (17.7%). **Utilization:** Among recognized practices, significantly greater % of their resource use accounted for by evaluation and management services (3.4%), and a smaller % accounted for by facility (−1.6%), inpatient ancillary (−0.1%), and non-management outpatient services (−1.0%). Recognized physicians had significantly fewer episodes per patient (0.13) and lower resource use per episode ($130).	**Fair**
Rosenthal et al., 2009[70]	Culinary Health Fund, a union-sponsored health plan, offered members and providers financial incentives to seek prenatal care.	Panel data analysis of outcomes and spending for participants and nonparticipants using instrumental variables to account for selection bias	$100 to both the pregnant member and the member's network obstetrician or midwife	**Cost/utilization:** NICU admissions, spending in the first year of life **Outcomes:** Low birth weight	**Cost/Utilization:** Lowered odds of neonatal intensive care unit admission (0.45; 95% CI, 0.23 – 0.88) Lowered spending in the first year of life (estimated elasticity of –0.07; 95% CI, –0.12 to –0.01) **Outcome:** No reduction in low birth weight (0.53; 95% CI, 0.23–1.18)	**Good:** Longitudinal study with strong design, including instrumental variables to account for confounding factors.

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Roski et al., 2003[47]	40 clinics of a large multispecialty medical group practice were randomly allocated to receive performance incentives related to smoking cessation from 1999 to 2000.	RCT focused on smoking cessation, provider adherence to accepted guidelines and associated patient outcomes. 40 clinics of a large multispecialty medical group practice were randomly allocated to control, incentive, and registry groups.	Clinics that met both goals with one to seven providers could receive a $5,000 award, and clinics with eight or more providers were eligible for a $10,000 bonus. Clinics who reached or exceeded only one of the two performance goals were eligible for half the amount.	**Process**: Referral to and use of counseling program **Outcomes:** Quit rate	**Process:** Patients visiting registry clinics accessed counseling programs statistically significantly more often (P 0.001) than patients receiving care in the control condition **Outcomes:** Quitting rate (7-d sustained abstinence, not-incentivized) was 22.4% for the P4P group, 21.7% for the incentive registry group, and 19.2% for the control group	**Fair**
Serumaga, 2011[24]	UK National Health Service Quality and Outcomes Framework	Interrupted time series analysis (2000–2007)	PCPs can receive up to 25% of base salary	**Process:** Rates of blood pressure monitoring **Outcomes:** Blood pressure over time, blood pressure control, treatment intensity, hypertension related outcomes, all-cause mortality	**Process:** After accounting for secular trends, no changes in blood pressure monitoring (level change 0.85, 95% confidence interval −3.04 to 4.74, P=0.669 and trend change −0.01, −0.24 to 0.21, P=0.615), control (−1.19, −2.06 to 1.09, P=0.109 and −0.01, −0.06 to 0.03, P=0.569), or treatment intensity (0.67, −1.27 to 2.81, P=0.412 and 0.02, −0.23 to 0.19, P=0.706) were attributable to P4P. **Outcomes:** P4P had no effect on the cumulative incidence of stroke, myocardial infarction, renal failure, CHF, or all-cause mortality in both treatment-experienced and newly treated subgroups.	**Fair**

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Unutzer et al., 2012[37]	The state of Washington implemented a population-focused, integrated care program for safety net patients in 29 community health clinics related to depression from 2008 to 2010.	Survival analyses, which examined the time to improvement in depression before and after implementation of the P4P program.	Annual program funding to participating clinics was contingent on meeting several quality indicators	**Process:** Timely follow-up of patients in the program, psychiatric consultation for patients who do not show clinical improvement, and regular tracking of psychotropic medications **Outcome:** Treatment response	**Process**: After implementation of the P4P incentive program, participants were more likely to experience timely follow-up, and the time to depression improvement was significantly reduced **Outcomes:** The hazard ratio for achieving treatment response was 1.73 (95% confidence interval = 1.39, 2.14) after the P4P program implementation compared with preprogram implementation.	**Poor:** Simple pre-post with no comparison group.
Young et al., 2007[8]	PCPs in Rochester, New York, received withheld bonuses for performance on process and patient experience measures. Focused on diabetes measures.	Pre-post with no comparison group	5% physician fees withheld to fund incentive pools and returned based on performance	**Process:** 5 diabetes measures: 2 Hemoglobin A1c tests, 1 LDL screening, 1 urinalysis/microalbumin, 1 flu vaccination, and 1 eye exam	**Process:** Post-P4P implementation, statistically significant increases for all measures were observed, with largest increases for LDL screening and eye exams. No significant interaction term for every measure, indicating that there was no difference between the post- and pre-intervention trends.	**Poor:** Regional population, simple pre-post, no controls for confounding factors.
Young et al., 2010[150]	P4P programs in 3 safety net settings in Chicago, offering incentives to physician groups for performance on process-of-care measures	Two case studies	Bonus of up to $4,000 based on performance	**Process:** Program A: annual retinal eye exam, annual HbA1c testing for diabetics, prescription of controller medications for patients with asthma, and 6 well-child visits. Program B: Annual HbA1c test, annual LDL check, and annual foot exam.	**Process:** No evidence that P4P led to substantial improvements in quality.	**Poor:** Limited to two case studies.

Table 3.5. Evidence on Effectiveness of Hospital Pay-for-Performance Programs

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Atkinson et al., 2010[154]	Case study of Long Island Health Network P4P program, implemented in 2004 and operated by 10 clinically integrated hospitals	Longitudinal analysis (2004–2008) of single integrated system	Part of annual update at risk. Amount at risk unspecified	**Process:** 23 core Hospital Compare measures **Utilization:** Case mix–adjusted LOS	**Process:** Overall composite measure of quality has shown a steady increase over time from 78 in the first quarter of 2004 to 93.3 in the first quarter of 2008 **Utilization:** Case mix-adjusted average LOS has decrease of about 0.25 days from 2003 to 2008	**Poor:** Case study within a single organization, no comparison group, no statistical testing
Berthiaume et al., 2004[156]	Hospital Quality and Service Recognition program: Implemented by the Hawaii Medical Services Association, focused on GWTG-CAD	Single year cross section from 2002	Bonus payments provided based on point system consistent with GWTG-CAD program	Number of hospitals receiving incentives	**Process:** 4 of 13 hospitals attained 85% adherence to the GWTG-CAD performance measures	**Poor:** Small sample size, no comparison group, no statistical testing, results included only the proportion of hospital meetings goals and receiving incentives
Berthiaume et al., 2006[155]	Hospital Quality and Service Recognition program: Implemented by the Hawaii Medical Services Association, with 17 hospitals focused on GWTG-CAD	Longitudinal analysis (2001–2004) of participants	Bonus payments provided based on point system consistent with GWTG-CAD program	**Outcomes:** Surgical/OB LOS and complications, patient experience	**Outcomes:** Significant reduction in Surgical LOS, no change in OB LOS No statistically significant change in complications No statistical significant change in patient experience reported	**Poor:** Small sample size, no comparison group

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Calikoglu et al., 2012[57]	Quality-Based Reimbursement Program and the Hospital-Acquired Conditions Program sponsored by the State of Maryland studied from 2009 to 2011	Longitudinal analysis comparing MD hospital trend with national trend	Rewards for highest performers and penalties for lowest performers. Reallocation is the % of total inpatient revenue that the hospital was penalized or rewarded by, based on its performance score. The maximum penalty for the quality-based reimbursement program is set at 0.5%, and the distribution of penalties and rewards is determined based on a linear scale.	**Safety:** 3M's 64 preventable conditions list **Process:** 19 core CMS and Joint Commission process measures in 4 care domains: heart attack, CHF, pneumonia, and surgical infection prevention.	**Safety:** Preventable conditions declined, especially infection-related conditions (All included: –18.59%, infection-related –27.83%, all other –14.33% p<0.001 **Process:** Only measure that improved faster was influenza vaccination for pneumonia patients (+20.5% in MD vs. +15.1%).	**Fair**

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Glickman et al., 2007[53]	CMS HQID	Longitudinal analysis (2003–2006) comparing change in participants to nonparticipants	HQID methodology (see page 48 for details)	**Process:** *CMS measures:* aspirin at arrival, aspirin at discharge, angiotensin-converting enzyme inhibitor or angiotensin receptor blocker for left ventricular systolic dysfunction, Smoking cessation counseling for active or recent smokers, Beta Blocker at arrival, Beta Blocker at discharge *Non-CMS measures:* Glycoprotein IIb/IIIa inhibitor use, clopidogrel at discharge, any heparin use, lipid-lowering medication, dietary modification counseling, referral for cardiac rehabilitation, electrocardiogram within 10 minutes, cardiac catheterization within 48 hours **Outcomes:** In-hospital death	**Process:** Slightly higher rate of improvement for 2 of 6 targeted incentivized therapies at P4P vs. control hospitals for aspirin at discharge (OR 1.31 vs. 1.17, p=.04), smoking cessation counseling (OR 1.50 vs. 1.28, p=.05). No significant difference in a composite measure of the 6 incentivized measures between groups. **Outcomes:** No evidence that in-hospital mortality improvements were incrementally greater at P4P hospitals (change in odds of in-hospital death per half-year period, 0.91 vs. 0.97, p=.21).	**Good:** Solid design with a comparison group to account for fixed difference in outcomes across practices, adjusted for patient risk in mortality models

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Grossbart, 2006[153]	CMS HQID	Difference – in-differences from 2003–2004 comparing participating hospitals within Catholic Healthcare partners to those that did not participate	HQID methodology (see page 48 for details)	**Process:** Composite quality scores in 3 clinical areas: AMI, CHF, and pneumonia. Number of opportunities and % improvement for each measure of AMI, CHF, and pneumonia	**Process:** Participating hospitals improved their composite scores by 9.3% versus 6.7% for nonparticipating hospitals ($p < .001$). For CHF, improvement from baseline to the 1st year for participating hospitals was 19.2% versus 10.9% for nonparticipating hospitals in CHF ($p < .001$). In the area of AMI, the improvement from baseline to the 1st year for participating hospitals was 3.1% versus 2.9% for nonparticipating hospitals, although this was not significant ($p = .730$). Among pneumonia patients, nonparticipating hospitals slightly outpaced the pay-for-performance cohort (7.9% vs. 7.2%), although again, the difference was not significant ($p = .395$).	**Fair**

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Herrin et al., 2008[60]	Health care system in Texas implemented a P4P program that distributed bonuses to director/clinical managers and chief executive officers for patient experience, process, and efficiency measure.	Longitudinal analysis (2002–2005) with comparison hospitals in Texas	Portion of salary at risk based on performance, ranging from 10% for clinical managers to 60% for the chief executive officer.	**Process:** Quality index based on 13 core Joint Commission measures related to AMI, pneumonia, CHF, and surgical site prevention **Outcomes:** Mortality	**Process:** On seven measures, Baylor Healthcare System hospitals improved compliance more rapidly. For three of the core measures, BHCS hospitals increased compliance significantly faster: beta-blockers at admission (p = .04), beta blockers at discharge (p = .007), and antibiotics within 4 hours (p = .014). In contrast, for the three non-exposed measures, BHCS hospitals had average changes that were smaller or that were even more negative, though not significantly so, than other hospitals reporting to the Joint Commission. **Outcome:** No significant difference in mortality rate.	**Fair**
Jha et al., 2012[73]	CMS HQID	Longitudinal analysis (2003–2009) with comparison group	HQID methodology (see page 48 for details)	**Outcome:** 30-day mortality among patients who had AMI, CHF, pneumonia or who underwent CABG in HQID and non-HQID hospitals	**Outcome:** At baseline, the composite 30-day mortality was similar for HQID and non-HQID hospitals. The rates in mortality per quarter decreased at the HQID and non-HQID hospitals were similar (0.04% and 0.04%, difference, −0.01 percentage points; 95% CI, −0.02 to 0.01). After 6 years, mortality remained similar in HQID and non-HQID hospitals (11.82% and 11.74%; difference, 0.08 percentage points; 95% CI, −0.30 to 0.46). No evidence that HQID led to a decrease in 30-day mortality.	**Fair**

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Kruse et al., 2012[77]	CMS HQID	Difference-in-differences using data from 2002 to 2005	HQID methodology (see page 48 for details)	**Costs:** Hospital revenues, costs, and margins or Medicare payments (index hospitalization and 1 year after admission) for AMI patients	**Costs:** No significant effect of P4P on hospital revenues, costs, and margins or Medicare payments (index hospitalization and 1 year after admission) for AMI patients.	**Good:** Utilized a difference-in-differences design with a strong empirical framework to also account for time-variant hospital characteristics
Lindenauer et al., 2007[59]	CMS HQID	Longitudinal analysis (2003–2006) using an exact match approach to match HQID hospitals with controls	HQID methodology (see page 48 for details)	**Process:** 10 individual process measures of AMI, CHF, and pneumonia and composite scores for AMI, CHF, pneumonia, and all combined	**Process:** Pay-for-performance hospitals showed significantly greater improvement than did control hospitals in 7 of the 10 individual measures. Pay-for-performance hospitals also achieved greater improvement in all the composite process measures, with differences ranging from 4.1% for pneumonia (P<0.001) to 5.2% for CHF (P<0.001).	**Good:** Large national sample with a solid matching methodology to account for potential confounders.
Nahra et al., 2006[157]	Blue Cross Blue Shield of Michigan implemented a hospital incentive system for heart-related care involving 85 hospitals.	Pre-post comparison among participating hospitals	% add-on to hospitals' inpatient DRG reimbursements from Blue Cross Blue Shield of Michigan. Maximum possible add-on for heart related care has increased from 1.2% of a hospital's BCBSM inpatient DRG reimbursements in 2000–2002 to 2% of a hospital's Blue Cross Blue Shield of Michigan inpatient DRG reimbursements in 2003	**Process:** Aspirin at discharge; AMI patients receiving beta blocker at discharge; CHF patients receiving ACE inhibitor prescriptions at discharge. **Outcome:** Quality-adjusted life years	**Process:** Aspirin at discharge patients from 87% to 95%, Beta blockers from 81% to 93%, and ACE inhibitors from 70% to 80%. **Outcome:** Improvement in quality-adjusted life years between 733.3 and 1,701.2	**Poor:** Limited to a single region, no comparison group, no controls included in calculation of "benefit"

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Nicholas et al., 2011[54]	CMS HQID	Longitudinal analysis (2003–2005) with comparison group	HQID methodology (see page 48 for details)	**Process:** CMS core measures	**Process:** P4P hospitals did not preferentially increase efforts for easy tasks in patients with CHF or pneumonia, but they did exhibit modestly greater effort on easy tasks for heart attack admissions.	**Good:** Multiple years of a large national sample, strong analytic design using fixed and random effects and hospital characteristics to control for potential confounders
Ryan et al., **2009**[78]	CMS HQID	Difference-in-differences using multiple years of data (2000–2006)	HQID methodology (see page 48 for details)	**Costs:** Risk-adjusted 60-day cost for AMI, CHF, pneumonia, or CABG **Outcomes:** Risk-adjusted 30-day mortality for AMI, CHF, pneumonia, or CABG	**Costs:** No evidence that the HQID had a significant effect on risk-adjusted 60-day cost **Outcomes:** No evidence that the HQID had a significant effect on risk-adjusted 30-day mortality	**Good:** Multiple years of a large national sample, strong analytic design using fixed and random effects and hospital characteristics to control for potential confounders
Ryan and Blustein 2011[55]	MassHealth	Longitudinal analysis (2004–2009) with comparison group	Hospitals were eligible to receive three types of rewards: "Attainment Award," given to hospitals with composite scores exceeding the median from HQID hospitals 2 years prior; and "Improvement Award," given to hospitals scoring above the median of HQID hospitals in the current year and also ranking within the top 20% in terms of QI among HQID hospitals.	**Process:** CMS core measures for pneumonia and surgical site infections	**Process:** Estimates from preferred specification, found small and non-significant program effects for pneumonia (–0.67 percentage points, p>0.10) and SIP (–0.12 percentage points, p>0.10)	**Good:** Multiple years of a large national sample, strong analytic design using fixed effects and hospital-specific time trends to control for potential confounders

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Ryan et al., 2012a[90]	CMS HQID	Matched difference-in-differences using multiple years of data (2004–2009)	HQID methodology (see page 48 for details)	**Process:** Composite process quality scores for AMI, CHF, and pneumonia	**Process:** In every case, HQID hospitals improved their quality more than matched comparison hospitals in phase I HQID hospitals experienced a weakening of QI relative to matched comparison hospitals in phase II. In both phases, average adjusted annual QI was greater for demonstration hospitals than for matched comparison hospitals for each diagnosis. Overall difference-in-differences estimates indicated that HQID hospitals improved less in phase II than phase I, compared with comparison hospitals, the difference was significant for HF and pneumonia, but not AMI.	**Good:** Large national sample, used match comparison group, and differences-in differences to account for other time invariant differences between hospitals

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Sutton et al., 2012[72]	P4P program implemented in 24 hospitals in the northwest UK	The triple-difference (2007–2010) analysis captured the effect of the program on mortality for the conditions included in the program in the northwest region in addition to changes over time in overall mortality in the northwest region and differences in mortality between the conditions included and not included in the program between the northwest region and the rest of England	HQID methodology (see page 48 for details)	**Outcome:** Changes in mortality	**Outcome:** Risk-adjusted, absolute mortality for the conditions included in the pay-for-performance program decreased significantly. Absolute reduction of 1.3 percentage points (95% confidence interval [CI], 0.4 to 2.1; P = 0.006) Relative reduction of 6%, equivalent to 890 fewer deaths (95% CI, 260 to 1500) during the 18-month period. The largest reduction, for pneumonia, was significant (1.9 percentage points; 95% CI, 0.9 to 3.0; P<0.001), No significant reductions for acute myocardial infarction (0.6 percentage points; 95% CI, −0.4 to 1.7; P = 0.23) and CHF (0.6 percentage points; 95% CI, −0.6 to 1.8; P = 0.30).	**Good:** Very strong analytic approach with multiple sensitivity checks

Reference	Program Description	Study Design	Incentive Structure	Measures Examined	Findings	Assessment of Methodological Quality
Werner et al., 2011[56]	CMS HQID	Longitudinal analysis (2004–2008) with matched comparison group	HQID methodology (see page 48 for details)	**Process:** CMS core measures for AMI, pneumonia, and CHF and calculated the composite scores for pneumonia and CHF	**Process:** Performance of the hospitals in the project initially improved more than the performance of the control group: More than half of the pay-for performance hospitals achieved high performance scores, compared with less than a third of the control hospitals. However, after five years, the two groups' scores were virtually identical.	**Good:** National sample of intervention practices over time matched to large number of comparison practices using a number of key variables

Table 3.6. Evidence on Effectiveness of Pay-for-Performance Programs in Other Settings

Reference	Program description	Study design	Incentive structure	Measures examined	Findings	Assessment of Methodological Quality
Hittle et al., 2011[75]	Medicare implemented the Home Health Agency P4P demonstration and incentivized improvements in patient outcomes and cost-savings to Medicare	RCT from 2007 to 2008 comparing treatment, control, and nonparticipants	Program cost savings were distributed to the highest-performing agencies and the most improved	**Outcome:** 21 measures of activities of daily living; 7 incentivized, 14 not incentivized	**Outcome:** Only 2 measures (improvement in pain interfering with activity and improvement in urinary incontinence), which were both non-incentivized, showed significant differences btw treatment and control participating home health agencies. **Utilization:** No significant difference in change between treatment and control hospitalization or emergent care	**Fair**
Shen, 2003[76]	Maine Office of Substance Abuse incentivized nonprofit providers to care for high-priority substance abuse clients	Office of Substance Abuse clients were compared before and after the intervention to Medicaid patients	Annual payment update dependent on previous performance	**Outcomes:** The proportion of outpatient clients classified as being the most severely ill	**Outcome:** Performance-based contracting had a significantly negative marginal effect on the probability of Office of Substance Abuse clients being most severe	**Fair**
Shepard et al., 2006[61]	Addiction services company offered incentives to 11 substance abuse counselors providing outpatient aftercare treatment	RCT from 1994 to 1996	Counselor could earn a bonus of $100 for each client who completed at least five treatment sessions	**Process:** Number of treatment sessions	**Process:** 59% of patients in treatment group completed at least five sessions, whereas 33% in comparison group completed the same	**Fair**

Reference	Program description	Study design	Incentive structure	Measures examined	Findings	Assessment of Methodological Quality
Werner, 2013[74]	Medicaid's nursing home P4P from 2001 to 2009	Difference-in-differences	Point system translating into a per-diem add-on	Resident-level indicator of clinical outcomes (e.g., falls, pressure sores, catheter insertion, and restraints) and facility-level regulatory deficiencies (total number of deficiencies in a given year and the number of immediate jeopardy deficiencies).	**Outcome:** Three clinical quality measures (the % of residents being physically restrained, in moderate to severe pain, and developed pressure sores) improved, other targeted quality measures either did not change or worsened. Two structural measures (total number of deficiencies and nurse staffing) worsened slightly under P4P	**Good:** Multiple years with difference-in-differences design

6. Does performance on unmeasured aspects of quality of care suffer when providers focus on improving performance on what is being measured ("teaching to the test)"? Conversely, are there "spillover effects" whereby quality improvement efforts improve care more broadly?

We found 21 articles (Table 3.7) that examined effects on unmeasured areas, meaning there was some assessment of possible unintended or spillover effects. The types of effects assessed included gaming the data used to generate scores, focusing only on improving areas that are measured and incentivized by the P4P program and ignoring clinically important areas that are not, avoiding sicker or more challenging patients when providing care, providing care that is not clinically recommended, and examining non-incentivized areas of performance to assess whether changes providers make more broadly affect care delivery.

Overall, the studies show small to no unintended effects.

Unintended Effects

A study by Beaulieu and Horrigan[41] did not find that physicians reallocated effort away from preventive screening (colorectal cancer and mammography screening were not incentivized measures) toward diabetes care (which was incentivized). One of the stronger studies we reviewed by Glickman et al.[53] compared hospitals in the Premier HQID to non-incentivized hospitals in the CRUSADE (i.e., Can Rapid risk stratification of Unstable angina patients Suppress ADverse outcomes with Early implementation of the American College of Cardiology/American Hospital Association guidelines) project and did not find any negative effects on other aspects of clinical care given simultaneous hospital participation in a QI registry. There was no difference found in the composite measures of AMI treatments, and rates of improvement did not differ, except prescribing of lipid-lowering medication at discharge, which was significantly higher at P4P hospitals (OR=1.23 vs. 1.13, p=.02). The absence of observed negative effects may in part be due to the fact that many of the P4P interventions studied were either small in scale or did not put substantial amounts of revenue at risk (which may occur under newer models of VBP).

Healy and Cromwell[86] evaluated the impact of CMS's policy related to nonpayment for selected preventable HACs in three states and found some evidence of gaming of data across payers. They found that undercoding had taken place by moving HACs to the secondary diagnosis code fields nine and above, which were not captured by the measure specifications. The amount of undercoding found varied by type of HAC, with the highest occurring for falls and trauma. Hospitals also undercoded HACs for hospital-acquired stage III or IV pressure ulcers, catheter-associated urinary tract infection, and vascular-catheter-associated infection. The authors also saw a greater use of all eight primary diagnoses fields used to compute the HAC score among Medicaid patients, which they surmised was a result of these patients likely being sicker. Two more recent retrospective studies conducted in the Veteran's Health Administration found evidence of overtreatment of patients with blood pressure and diabetes, which the authors

of the study observe be a function of using target-based performance measures (e.g., percentage of all diabetic patients with HbA1c level <8). The first study found potential overtreatment of ~8 percent for high blood pressure management,[80] and the second study found potential overtreatment of ~13 percent for lipid management with high dose statins.[82]

Spillover Effects

In a small number of cases, there was evidence of improvement on non-incentivized measures within the same conditions that were the target of the incentives. Several of the studies suffered from methodological problems in their design that make it difficult to assess any improvements or declines—specifically, not controlling for secular changes or trends that could explain any of the observed differences.

A study by Mullen et al.[42] attempted to measure potential spillover effects on unpaid measures (diabetic eye exams, ACE inhibitor for seniors with CHF, appropriate use of antibiotics, management of cholesterol-lowering drugs, chlamydia screening, and asthma-related emergency room visits). Although there was a slight decline in performance on a few of the measures, the authors of this study concluded that the non-incentivized measures do not give a clear picture of response patterns to P4P, either positive spillovers or disruption in care.

Healy and Cromwell[86] also found limited evidence of positive spillover effects of the CMS HAC–Present on Admission program on payers other than Medicare for two of the three conditions evaluated. However, they cautioned that the results could be interpreted as showing no impact of the Medicare HAC–Present on Admission program on the three studied HACs. In the Maryland HAC study by Calikoglu et al.,[57] the state of Maryland instituted audit procedures to prevent coding problems and did not report coding irregularities (98 percent were found to be coded correctly). Among the complications that were not part of Maryland's nonpayment policy for HACs, there was an increase, though this could have resulted from improved documentation of these conditions or actual increases in complications. Therefore, one cannot conclude from this study that the incentive policy led to worse performance on those things that were not measured.

The Hittle et al.[75] study of use of P4P in the home health agency setting found that those sites exposed to P4P performed slightly better, although not statistically significantly different than the control group on the non-incentivized measures (improvement in pain interfering with activity and improvement in urinary incontinence).

A study of the UK P4P experiment[84] showed that performance for incentivized indicators for three conditions was substantially higher at all three time points (1998, 2003 pre-P4P, and 2005 post-P4P) than for indicators without incentives. However, the rate of improvement did not differ between 2003 and 2005 for clinical indicators with and without financial incentives. Although this study does not provide insights on the effects of financial incentives on care provided for conditions that were not incentivized, the evaluators hypothesized that there might have been a

spillover effect between incentivized and non-incentivized indicators focused on the same conditions.

While not included in our evidence table, a study of 79 physician organizations in Massachusetts by Mehrotra et al.[158] found that when queried about possible unintended consequences or adverse effects, providers did not note these concerns.

Strength of Evidence: Low. At this stage, undesired effects look minimal to nonexistent, though many of the studies are not sufficiently strong to assess these effects. There are few studies that examine spillover effects to provide evidence of the effects. As P4P program designs change and incentives to engage in undesired ways increase as more money is at risk, it will be important to continue to monitor for unintended consequences.

Table 3.7. Pay-for-Performance's Effect on Unmeasured Areas—Unintended and Spillover Effects

Reference	Program Description	Unintended Consequences	Improvements in Areas Not Incentivized by Program (Spillover Effects)	Assessment of Methodological Quality
An et al., 2008[49]	RCT of usual care vs. P4P for smoking quit line referrals in 25 usual care clinics with 24 P4P clinics. 10 month study period from 2005–2006.	No evidence of unintended consequences. Referral rates of contact and subsequent enrollment in quit services did not differ between usual care and P4P sites.	Not reported	**Poor:** Small intervention, short time period. Strength is randomization of clinic sites.
Beard et al., 2013[80]	Retrospective cohort study assessing measures within the VAs for appropriate care and overtreatment of lipid management among a cohort of patients with diabetes. 1-year study period from 2010–2011.	13.7% received potential overtreatment: high-dose statins for patients with no diagnosis of ischemic heart disease either during or before the measurement period.	Not reported	**Fair :** Data did not capture care provided outside of the VA. Strength is large nationally representative sample.
Beaulieu and Horrigan 2005[41]	Independent Health managed care plan in New York state physician P4P program (n=17 physicians). Focus on diabetes process and outcome measures. 8-month study period from 2001 to 2002.		Assessed performance on two non-incentivized measures for mammogram and colorectal screening. 10 physicians improved, 7 remained unchanged. Authors concluded that physicians did not reallocate effort away from preventive screening toward diabetes care.	**Poor:** Small number of study participants (n= 17 physicians). Physicians self-selected; one small region, short duration, physicians not matched at baseline. Comparison patients had higher baseline performance on all measures

Reference	Program Description	Unintended Consequences	Improvements in Areas Not Incentivized by Program (Spillover Effects)	Assessment of Methodological Quality
Healy and Cromwell 2012[86]	CMS identified 8 conditions for which it would no longer pay a higher DRG rate if the conditions occurred in the inpatient setting and were not present on admission. 3-year evaluation from 2008 to 2010.	Across all payers, counting all secondary diagnosis codes had the greatest positive effect in raising HAC rates for Medicare and Medicaid beneficiaries. Evidence of undercoding HACs for trauma and falls, deep vein thrombosis/PE following certain orthopedic procedures, stage III or IV pressure ulcer, catheter-associated urinary tract infection, and vascular-catheter-associated infection. Highest undercoding rates found for trauma and falls and deep vein thrombosis/PE after orthopedic procedures. No consistent pattern in coding could be found across hospital characteristics across the HACs.	Assessed rates of decline in HACs among non-Medicare payers as a result of the Medicare HAC-Present on Admission nonpayment. No consistent pattern in the reporting of the rates of HACs across 3 years or by type of payer or by state.	**Fair:** Examined variation across 4 states in reported rates and differences in coding.
Calikoglu et al., 2012[57]	Two P4P programs implemented in 2008 by the state of Maryland, one focused on process measures and one on HACs. (2007–2010)	No evidence of unintended consequences. Audits to guard improper coding found 98% of hospitals were coding correctly present on admission	Not reported	**Poor:** Measured change compared with base period for HACs. No accounting for secular effects and anticipatory behavior related to implementation of CMS non-payment policy going into effect in 2012. Regional effort in an all payer state. No controls for confounders. No comparison group or trends prior to implementation of program.

Reference	Program Description	Unintended Consequences	Improvements in Areas Not Incentivized by Program (Spillover Effects)	Assessment of Methodological Quality
Campbell and Marchildon, 2007[84]	UK P4P contract for family practitioners started in 2004. Study assesses longitudinal change at three time points 1998, 2003 and 2005 after introduction of P4P in 2004	Not reported	Performance on indicators with incentives for three conditions examined was substantially higher at all three time points than for those without incentives. The rate of improvement between 2003 and 2005 for clinical indicators for which financial incentives were provided, as compared with those for which they were not, did not differ significantly from the rate predicted based on the trend between 1998 and 2003. There may have been a halo effect between incentivized and non-incentivized indicators focused on the same conditions. The finding of no significant difference in the rate of improvement between clinical indicators for which financial incentives were provided and those for which they were not provided suggests that the P4P program may not necessarily have been responsible for the acceleration in improvement found between 2003 and 2005.	**Fair:** Absence of a control group as P4P was implemented nationally. Small sample size to assess spillover effects. Results may not be generalizable to the US. UK program had EHRs in all clinical practices with prompts for clinical measures, national health insurance, substantial incentives, and a history of significant investments in QI efforts that started measures on upward trajectory prior to P4P
Campell et al., 2009[159]	UK P4P contract (Quality Outcomes Framework) for PCPs started in 2004. 136 performance indicators Interrupted time series analysis examined longitudinal change for 42 practices at four time points before and after implementation of P4P (1998 pre-P4P, 2003 pre-P4P,2005 post-P4P, and 2007 post-P4P)	Study found a ceiling effect for primary care practices (2005: practices achieved 96.9% of available clinical quality payment points; 2007: practices achieved 97.8% of available clinical quality points). Continuity of care declined after implementation of P4P in 2005.	Not reported	**Fair:** Absence of a control group as P4P was implemented nationally. Small sample size to assess spillover effects. Results may not be generalizable to the US. UK program had EHR in all clinical practices with prompts for clinical measures, national health insurance, substantial incentives, and a history of significant investments in QI efforts that started measures on upward trajectory prior to P4P

Reference	Program Description	Unintended Consequences	Improvements in Areas Not Incentivized by Program (Spillover Effects)	Assessment of Methodological Quality
Chung et al., 2010[103]	Palo Alto Medical Clinic physician P4P program (primary care). 9 incentivized clinical outcome and process measures during study period from 2005 to 2007.	Not reported	Accelerated improvement for 1 of 5 non-incentivized measures (BP control for hypertensive patients) from 65% to 72% (p=0.01)	**Poor:** Compares 2006- 2007 performance against 2005–2006 (pre-post) in same organization. Not match providers or patients within providers. One organization with unique characteristics (EHR, low patient turnover, high patient socioeconomic status (SES), history of physician feedback on performance); overlap of measures with the statewide IHA P4P program
Collier, 2007[38]	A community health care system implemented a P4P program for 12 hospitalists regarding standards on access, timeliness of medical record dictation, and participation in monthly hospitalist meetings, quality measures, and self-directed learning. (pre-P4P 2003–2004 vs. post-P4P 2005–2006)	Not applicable	Average LOS for patients (not incentivized) decreased more for patients of P4P hospitalists from 2005 to 2006 (5.22 to 4.84 days, excluding outliers,) than non-P4P hospitalists (4.89 to 4.87 days, excluding outliers).	**Poor:** Does not account for secular improvement trends in Joint Commission/CMS measures and declines in LOS. Concurrent non-contracted group and non-hospitalists (not matched). Only a single organization and analytic methods poorly explained. Unclear if results generalize.

Reference	Program Description	Unintended Consequences	Improvements in Areas Not Incentivized by Program (Spillover Effects)	Assessment of Methodological Quality
Drake et al., 2007[160]	CMS HQID incentivized hospital performance on 5 clinical conditions. Evaluated 130 top-performing hospitals on the pneumonic antibiotic timing measure in the 1st year of the HQID (2003–2004) and changes in antibiotic prescription rates for other clinical conditions.	Increased rate of meeting the pneumonia antibiotic timing measure was correlated with an increase in inappropriate pneumonia antibiotic use among patients with CHF, asthma, and chronic obstructive pulmonary disease. There was insufficient data to assess antibiotic use rates for pulmonary embolism, pulmonary edema and respiratory failure, and bronchiolitis and respiratory syncytial virus.	Not reported	**Poor:** No multivariate analysis, simply demonstrated that better performance on antibiotic timing was correlated with inappropriate prescribing in some circumstances
Fagan et al., 2010[40]	Longitudinal study analyzing claims files of 20,943 adults aged ≥65 with diabetes receiving care from 9 primary care practices in Alabama, Tennessee, and Texas. Evaluated performance on 5 incentivized measures, 2 non-incentivized measures, and 2 resource-use measures was evaluated (1,587 intervention patients and 19,356 patients in comparison practices). (2004–2007)	Not applicable	No evidence of spillover effect of P4P on non-incentivized measures (short-acting antihypertensive medication (OR=1.11 95% CI (.58, 2.13)) or prescribing an ACE for those with renal insufficiency (OR=0.76 95% CI (0.54, 1.06)).	**Good:** Quasi-experimental longitudinal study (pre-post data). Relatively large region, difference-difference (like) design to control for time invariant confounders

Reference	Program Description	Unintended Consequences	Improvements in Areas Not Incentivized by Program (Spillover Effects)	Assessment of Methodological Quality
Glickman et al., 2007[53]	Patients with non-ST-segment elevation myocardial infarction enrolled in CRUSADE exposed to CMS HQID demonstration Evaluation program from 2003–2006.	No deleterious effect on other aspects of clinical care given simultaneous hospital participation in a QI registry not involving financial incentives.	For composite measures of AMI treatments not subject to incentives, rates of improvement were not significantly different between P4P hospitals and controls (P4P hospital composite OR =1.09 vs. 1.08 for controls, p=.49), except lipid lowering medication, which was significantly higher at P4P hospitals (OR=1.23 vs. 1.13, p=.02)	**Good:** Observational, patient-level analysis. Large sample, multiple years of data. Solid design with a comparison group to account for fixed difference in outcomes across practices, adjusted for patient risk in mortality models
Herrin et al., 2008[60]	Baylor Health Care System in Texas implemented a P4P program in 2001 at 5 hospitals. Bonuses to director/clinical managers and chief executive officers for patient experience, process, and efficiency measures. Study period from 2001–2005.	Not reported	No evidence of spillover effects. Compared 3 measures not exposed to P4P (percutaneous coronary intervention within 120 minutes, thrombolytic therapy within 30 minutes for AMI, and discharge instructions for CHF). P4P hospitals had smaller average increases or larger average decreases than comparison hospitals, but differences were not significant. No significant difference in mortality rate.	**Fair:** Weak study design (pre-post), though some attempt to control for confounds. Comparison hospitals may differ substantially from 5 exposed to this intervention. Does not control for selection effects in measures reported to Joint Commission (which were voluntary)
Hittle et al., 2011[75]	Medicare Home Health Agency P4P demo. Incentivized improvements in outcomes and cost-savings to Medicare. Evaluation of demo from 2007–2008.	Not reported	Among the non-incentivized measures, treatment sites performed slightly better (though not significant differences) than the control group. Two non-incentivized measures (improvement in pain interfering with activity and improvement in urinary incontinence) showed significant differences, with treatment group outperforming controls.	**Fair**

Reference	Program Description	Unintended Consequences	Improvements in Areas Not Incentivized by Program (Spillover Effects)	Assessment of Methodological Quality
Jha et al., 2012[73]	CMS HQID incentivized hospital performance on 5 clinical conditions. Study examined association between performance on incentivized measures and inpatient mortality for AMI, pneumonia, and CHF. Program evaluation from 2003–2009.	Not reported	No difference in trends in mortality rates between HQID and non-HQID hospitals (p=0.36) for outcomes that were not linked to incentives (CHF, and pneumonia)	**Fair**
Kerr et al., 2012[82]	Retrospective cohort study assessing measures within the VA for appropriate care and overtreatment of high blood pressure among a cohort of patients with diabetes. 1-year study period from 2009 to 2010.	~8% had potential overtreatment. Patients with potential overtreatment were found to be older, male, have ischemic heart disease, and have lower mean index BP. Among patients older than 76 with diabetes, ~12% were potentially over treated.	Not reported	**Fair:** Retrospective cohort design shows that overtreatment are approaching rates of under treatment solely in the VA. Strength of the study is a very large sample of clinics and patients.

Reference	Program Description	Unintended Consequences	Improvements in Areas Not Incentivized by Program (Spillover Effects)	Assessment of Methodological Quality
McDonald and Roland 2009[161]	Comparison of providers exposed to UK Quality and Outcomes Framework P4P program and medical groups in California exposed to IHA P4P program. Qualitative interviews with 40 physicians to assess physician perspective on unintended consequences of P4P programs.	UK physicians reported P4P changed the nature of the office visit (due to large number of performance measures (n=80) and heavy reliance on EHRs to prompt delivery of services), while California physicians expressed resentment about P4P and less motivation to act on incentives. California physicians were less aware of targets and witnessed less change in the nature of office visits. California physicians reported frustration with the inability to exclude patients from performance calculations, with some reporting undesirable behaviors such as dropping non-compliant patients. California physicians in the medical group with the largest incentives reported accusing patients of damaging their performance rating or lying to patients about the financial consequences of their refusing to comply. Most California physicians expressed concern that performance targets diminished clinical autonomy, while English physicians did not feel the same.	Not reported	**Poor:** Difficult to generalize more broadly to other US P4P programs. California physician sample drawn from 4 organizations that ranged in size from 600 to 3,000 physicians, with various percentages of payment linked to P4P. The 4 U.S. groups may not be representative of the broader experience in the IHA program or nationally. All physicians in UK sample use EHR with prompts for quality indicators, while only 7 of the physicians in U.S. sample used EHR

Reference	Program Description	Unintended Consequences	Improvements in Areas Not Incentivized by Program (Spillover Effects)	Assessment of Methodological Quality
Mullen et al., 2010[42]	PacifiCare implemented a QI program in California in conjunction with the IHA P4P program. Study analyzed effects of implementing both programs on incentivized and non-incentivized measures. (2001–2005).	No evidence of disruptions in care	Unclear effects on non-incentivized measures No real gains associated with diabetic eye exam rates, despite other diabetic measures being rewarded by QI program and IHA. No changes found for non-incentivized heart-related measures relative to control group. Non-incentivized appropriate antibiotic use declined slightly. Despite the presence of 2 other incentivized measures for women's health (breast cancer screening and cervical cancer screening), the non-incentivized Chlamydia screening rates decreased by ~2–5% points relative to its time trend and the Northwest control group.	**Good:** Regional intervention but strong design with difference-in-differences approach and multiple years of data
Nicholas et al., 2011[54]	Examined whether hospitals increase efforts on easy tasks relative to difficult tasks to improve scores under P4P, using the HQID demonstration data. Measures were classified as easy or difficult to improve based on whether they introduce additional per-patient costs and compared process compliance on easy and difficult tasks at hospitals eligible for HQID bonuses relative to hospitals engaged in public reporting. Study period from 2003to 2005.	Study found little evidence that hospitals changed allocation of efforts across tasks to maximize performance scores at lowest cost. P4P hospitals did not preferentially increase efforts for easy tasks in patients with CHF or pneumonia, but they did exhibit modestly greater effort on easy tasks for heart attack admissions.	Not reported	**Good:** Multiple years of a large national sample, strong analytic design using fixed and random effects and hospital characteristics to control for potential confounders

Reference	Program Description	Unintended Consequences	Improvements in Areas Not Incentivized by Program (Spillover Effects)	Assessment of Methodological Quality
Shen, 2003[76]	Maine Office of Substance Abuse incentivized nonprofit providers to care for high-priority substance abuse clients through performance-based contracting. Study period from 2001 to 2005.	Found selection effects, with the most severely ill group significantly declining in treatment under the performance-based contract by 7% (P≤ 0.001), compared with 2% among the Medicaid comparison groups.	Not reported	**Poor:** Simple pre-post, small region
Young et al., 2010[150]	Analyzed P4P programs in 3 safety net settings in Chicago, offering incentives to physician groups for performance on process-of-care measures. Study period from 2005 to 2007.	No evidence that P4P compromised quality on unmeasured areas. Survey responses indicated that participating physicians did not have strong concerns about unintended consequences.	Performance on non-incentivized measures (adolescent well-child visits, LDL screening, and nephropathy) increased during study period.	**Poor:** Limited to two case studies

7. If a provider/institution performs highly on all the VBP metrics but has average performance on everything that is not measured, which proportion of total potential improvement in health will be achieved? (In other words, if we imagine that a high-performing health system produces "X" amount more quality-adjusted life years than an average-performing system, what fraction of that X would be produced by a health system that was higher-performing on metrics commonly included in VBP programs currently, but was average-performing in unmeasured areas?)

We identified no studies that directly addressed this. Many VBP programs focus on process measures and intermediate outcomes. As discussed in in question 4, improvements in process measures are weakly associated with improvement in outcome measures. Furthermore, performance on process measures typically explains very small amounts of variation in outcome measures—frequently less than 10 percent. The extent to which these associations represent causal relationships is unclear, as few studies are adequately designed to assess this. It is possible that there are typically unmeasured provider characteristics that influence both process and outcome measures, resulting in biased results due to omitted variable bias. The studies that utilized methods to assess causality show very few associations between improvements in process measures and improvements in outcomes.

Strength of Evidence: Insufficient. We found no information in the published literature that directly addressed this question. Please refer to Chapter Six for the summary of the TEP's discussion.

8. How likely is it that improvements in our ability to measure what is important will change enough over the next five to ten years to significantly affect the answer to (7)?

Strength of Evidence: Insufficient. We found no information in the published literature that discussed this. Please refer to Chapter Six for the summary of the TEP's discussion.

9. Are there unexpected effects of VBP programs, including impacts on racial/ethnic and socioeconomic disparities, and access to care?

Many commentaries and P4P studies have commented about possible unintended effects, especially for low-SES patients; however, the empirical evidence on the effects of P4P on disparities is limited. Our review of the P4P literature found five studies that attempted to distill empirically the positive and negative effects of incentive programs on disparities (Table 3.8). The limited evidence that exists shows that, to date, there have been few effects either worsening or reducing disparities. This may be a function of the small size of incentives that have been used in the United States. We included one study in our review from the large P4P experiment in the UK, although the findings may not generalize to the United States due to substantial differences in the delivery system.

The Ryan study,[89] which had a strong design, found no negative access effects related to avoiding treating minority patients after introduction of the Premier HQID. The Jha et al. study[88] found that within the HQID there was a closing of the disparities gap, as measured by the DSH index, between hospitals with low and high DSH indices. A more recent study by Ryan et al.[58] found that changes to the HQID incentive structure resulted in a redistribution of available incentive payments between Phase I and II of the program, with a greater proportion going to hospitals with greater socioeconomic disadvantage (as measured by the DSH index). This effect was a function of changes in the structure of the incentive and not due to lower-performing hospitals actually improving more.[90] This study found that disparities had neither worsened nor reduced. A study by Doral et al. from the UK[91] found a lessening of the disparities gap in performance among primary care practices. These authors caution that PCPs under this incentive scheme could engage in "exception reporting," excluding patients from the quality measure calculation, which would lessen incentives to selectively go after better-risk/healthier patients. Exception reporting was also a feature of the HQID demonstration, and Ryan noted that this design feature might have prevented hospitals from reducing access to more challenging patient populations.

Other studies that explore the issue of disparities include a simulation study by Werner et al.[162] and a qualitative study of hospital executives by Weinick et al.[163] In the Werner study, researchers used data from 2004–2006 Hospital Compare (pay-for-reporting) to assess the potential effects of P4P on safety net hospitals by simulating difference in the predicted change in performance at hospitals with high and low percentages of Medicaid patients (10 measures). They also estimated payments the hospitals would have received had they been exposed to the same incentive rules in the HQID program. They found small projected differences in performance and incentives. However, the authors caution that safety net hospitals may suffer from relative comparisons under pay-for-reporting or P4P. They assert that this may exacerbate disparities unless design elements work to mitigate the effects (such as paying for improvement).

The Weinick et al. study[163] found that hospital executives expressed concerns that P4P programs may draw resources away from aspects of care important to minorities (e.g., patient education programs and interpreter services), could exacerbate existing resource constraints in safety net hospitals, and might encourage insurers to selectively go after better-risk/healthier patients. There was a desire to understand how best to address disparities and to consider alternative approaches in P4P design, including using incentives to improve access to minority patients and to target elements of care that are important to minorities (cultural competence, communication skills).

Strength of Evidence: Low. The few empirical studies that have been conducted have either no effects or ambiguous effects. Only one relatively weak study found positive effects in lessening gaps in performance. It is possible that additional research will change the estimate or confidence in the estimate of the effect as a function of alternative P4P program designs.

Table 3.8. Unexpected Effects on Access and Disparities of Pay-for-Performance Programs

Reference	Program Description	# of Providers or Patients Studied	Effect on Access to Care	Effect on Disparities	Assessment of Methodological Quality
Chien et al., 2010[22]	Hudson Health Plan (Medicaid) implemented a P4P program that incentivized immunization delivery to 2-year-olds according to the recommended series. $200 bonus/child (15–25% above base reimbursement) (2003–2007)	115 Hudson primary care practices; 16 comparison health plans	Not reported	No exacerbation in preexisting disparities. Racial/ethnic disparities fluctuated, but remained essentially unchanged.	**Good:** Regional but multiple years of observation. Case comparison and strong difference and difference design

Reference	Program Description	# of Providers or Patients Studied	Effect on Access to Care	Effect on Disparities	Assessment of Methodological Quality
Doran et al., 2008[91]	UK National Health Service Quality and Outcomes Framework P4P program. Bonus payments to PCPs achieving threshold quality targets for various clinical and patient experience quality measures. (2004–2007).	7367 general primary care practices	Not reported	Primary practices in the more deprived quintile improved at the fastest rates (increase by 7.6% compared with the least deprived quintile, 4.4% increase). Gap in median achievement between highest and lowest deprivation quintiles narrowed from 4.0% (year 1) to 1.5% (year 2) to 0.8% (year 3). The variation in achievement decreased at faster rate for practices in most deprived areas. Patterns were consistent across all 48 indicators. By year 3, the SES gradient had almost disappeared, though the poorest-performing practices remained concentrated in most deprived areas.	**Good:** Compared a large number of practices before and after intervention. Concern about generalizability from UK to the United States due to different characteristics of delivery system (national health insurance with universal access, national health IT system). Only practices with stable populations and complete data collection were included; only fairly unchanged indicators could be analyzed; analyses at the practice not patient level (comorbidity will have led to some patients being counted twice) deprivation was summarized at the level of super-output areas.

Reference	Program Description	# of Providers or Patients Studied	Effect on Access to Care	Effect on Disparities	Assessment of Methodological Quality
Jha et al., 2010[88]	CMS Premier HQID Incentivized hospital performance on 5 clinical conditions. Evaluation examined association between the DSH index and changes in performance for AMI, CHF, and pneumonia. (2003 4th quarter) and July 2006–June 2007)"	251 of 255 HQID hospitals compared with a national sample of 3017 hospitals	Not reported	By 2007, after 3 years of incentives, the DSH index was no longer associated with terminal performance for the three conditions; for non-incentivized hospitals (national sample), a higher DSH index was associated with lower terminal performance for the three conditions. Hospitals with more poor patients caught up to hospitals with fewer poor patients in the incentivized sample of hospital; this did not occur for the national sample comparison group At baseline, among HQID hospitals, a 10-point increase in DSH was associated with a –0.8% (95% CI, –1.3%, –0.3%) lower performance on AMI, and –1.1% (95% CI, –1.7%, –0.5%) lower performance on pneumonia. Non-incentivized hospitals performance was also negatively associated with the DSH index for all 3 measures as baseline. For HQID hospitals, a 10-point increase in the DSH index was associated with a 0.1% lower terminal performance on AMI (p=0.23), a 0.07% higher terminal performance on pneumonia (p=0.72), and no significant difference in terminal performance on CHF (p=0.81). A higher DSH index was still associated with lower terminal performance in the national sample for each of the 3 conditions. In 2007, the interaction term btw the DSH and change in performance for HQID and non-HQID hospitals was significant and negative for AMI (–0.6, p=0.045) and pneumonia (–0.2, p=0.009), but not for CHF (p=0.65). The interaction term btw the DSH and terminal performance for HQID and non-HQID hospitals was statistically significant for pneumonia (–0.8, p<0.001), borderline significant for AMI (–0.4, p=0.064), and not significant for CHF (p=0.174).	**Poor:** Two separate pre-post analyses with different data sets (HQA data for national sample and HQID data for P4P hospitals). Limited adjustments for hospital characteristics. Did not adjust for difference in patient characteristics or match hospitals at baseline. Possible selection effects with HQID hospitals; may differ in ways that are not observed. Results are not generalizable to other hospitals.

Reference	Program Description	# of Providers or Patients Studied	Effect on Access to Care	Effect on Disparities	Assessment of Methodological Quality
Ryan, 2010[89]	CMS Premier HQID P4P program that incentivized hospital performance on 5 clinical conditions. (2000–2006)	3,981,516 Medicare beneficiaries studied	Little evidence that the HQID P4P reduced access for minority patients. No significant pre-post differences in adjusted admission rates to HQID hospitals for any diagnosis. "Other race" beneficiaries had a significant reduction in adjusted admissions in the post period for AMI, but there was a secular reduction in AMI admissions pre-intervention. There was no evidence that hospitals close to thresholds for quality bonuses were more likely to avoid minority patients.	Reductions in CABG rates for each racial and ethnic cohort between pre and post period reflected substitution of CAGB to percutaneous transluminal coronary angioplasty during that period (change in clinical practice). Marginally significant ($p<0.10$) evidence of a reduction in probability of receiving CABG was found for minority patients and other race beneficiaries. Minimal evidence of minority patient avoidance, which may be due to practice of exception reporting (hospitals were allowed to exclude patients from counting toward quality performance).	**Good:** National sample, pre/post implementation of P4P. Strong estimation procedure including a difference-in-differences and time variant patient characteristics (co-morbidity, admission type) and hospital characteristics. Results may not generalize to non-elderly patients.

Reference	Program Description	# of Providers or Patients Studied	Effect on Access to Care	Effect on Disparities	Assessment of Methodological Quality
Ryan et al., 2012b[58]	CMS Premier HQID P4P program that incentivized hospital performance on 5 clinical conditions, Phases I and II of intervention. (2000–2008). Between Phase I and Phase II, CMS shifted the incentive structure from only providing incentive payments to hospitals in the top 2 deciles of performance to paying hospitals that improved or had high absolute performance.	266 hospitals (250 HQID hospitals and 250 comparison hospitals)		In Phase I, there were substantial gaps for receipt of any incentive payment (hospitals in the highest DSH quartile were 32.8 percentage points less likely (;<0.01) to receive any payments than hospitals in the lowest DSH quartile), total incentive payment (hospitals in highest DSH quartile received $26.84/discharge less than those in the lowest DSH quartile), and incentive payment per discharge across the DSH quartiles. In Phase II, the gap was not significant for the receipt of any incentive payment. Gap was reduced but remained significant for incentive payment per discharge: payments per discharge increased for hospitals in the two highest quartiles of DSH, but decreased for hospitals in the lowest DSH quartile. There were no significant reductions in the gap for total payments. From Phase I to Phase II, the median change in incentive payments per discharge –$2.58 for Quartile 1 (lowest DSH), $0.43 for Quartile 2, $6.99 for Quartile 3, and $14.85 for Quartiles 4 (highest DSH), indicating hospitals serving disadvantaged patients received more incentive payments per discharge. Authors caution that the narrowing of the gap in incentive payments was not the result of lower-performing hospitals improving more in response to Phase 2 incentives; changes in the distribution of payments were likely the result of a change in incentive scheme	**Good:** Large national sample, used match comparison group, and differences-in-differences to account for other time invariant differences between hospitals

10. What are the features of the highest-performing providers/institutions and their adaptations to VBP?

Few studies have explicitly examined the features of high-performing providers. We reviewed 14 studies that commented on characteristics associated with high performance (Table 3.9). High-performing providers (mostly P4P studies of physicians or physician groups) had the following characteristics:

- were larger provider organizations[7, 43, 69]
- had more health information technology infrastructure[93–96]
- had a medical group structure (versus an independent practice association structure)
- served a smaller fraction of low-SES or Medicaid patients[43]
- engaged in external QI initiatives[7]
- engaged in more care management processes[7]
- were historically high performers[10, 27]
- used order sets for treating hip and knee replacement, per performance on HQID measures related to surgery; used clinical pathways for treatment of AMI and hip and knee replacement; had a multidisciplinary team with the goal of improving care for AMI and CHF; and used computerized physician order entry systems[97]
- had nursing staff's support for quality indicators and adequate human resources for initiatives to improve performance[97]
- had a higher ratio of family practitioners to patients (UK study).[164]

Werner et al.[56] found that improvements were largest among hospitals that were eligible for larger bonuses, were well financed, or operated in less competitive markets. Three studies showed that hospitals with lower performance at baseline[58, 59] or with a higher DSH index[88] demonstrated larger improvements.

Strength of Evidence: Low to Insufficient. Few studies have addressed this issue, so the evidence is lacking regarding what characterizes high (or low) performers. Studies have been opportunistic in defining the characteristics based on the variables that were available to them, rather than considering more broadly the set of factors that would characterize providers who perform differentially.

Table 3.9. Factors Associated with Performance on Incentivized Measures

Reference	Program Description and # of Providers Studied	Metric Assessed	Characteristics of High Performers	Characteristics of Low Performers
An et al., 2008[49]	RCT of usual care vs. P4P for quit line referrals from 2005 to 2006. The study compared rates of referral; contact and enrollment after referral; and project costs in 25 usual care clinics with 24 P4P clinics.	% of smokers referred to quit line services: number of unique individuals referred divided by the estimated number of smokers seen in the clinic. Costs: Fixed clinic costs were divided equally across both groups. Development costs: time of physicians and staff of project, Fairview Physicians Associates, and health plan. Implementation costs: information packages to clinics, feedback efforts to intervention clinics, including triage fees, staff time, and incentive payments. Pay rates based on annual salaries for participating staff. Costs were from an insurer's perspective.	No associations between the % of smokers referred and clinic specialty type, number of physicians, and presences of EHR. No difference in mean referral rates observed in highly engaged clinics between P4P vs. control clinics (15.1% vs. 14.1% p=0.85). Differences observed for engaged clinics (10.1% vs. 3%, p=0.001) and less engaged clinics (10.1% vs. 1.1%, p=0.02) for P4P vs. control.	Not applicable
Chien et al., 2012[43]	Cross-sectional study of IHA P4P program. Examined the association between physicians organization located in lower SES areas and performance on P4P measures. 11,718 practice sites within 160 physician organizations (2009).	IHA composite performance score and PO area based SES measure based on Krieger's area based measure.	Largest physician groups had a higher likelihood of being ranked in the top 40% of performance than smallest POs (RR=2.55; 95% CI 1.67–3.90, p<0.001), as did medical groups when compared with independent practice associations (RR=2.93, 95%CI 2.00–4.28, p<0.001).	Significant positive relationship between PO SES and P4P performance (trend test p<0.001). POs in higher SES areas had higher performance scores. Median performance score of POs in the highest SES quintile was almost 20 points higher than POs in the lowest quintile. POs with higher percentages of Medicaid revenue were less likely to be in the highest 2 performance quintiles (RR=0.68, 95% CI 0.50–0.93, p=0.017).

Reference	Program Description and # of Providers Studied	Metric Assessed	Characteristics of High Performers	Characteristics of Low Performers
Coleman et al., 2007[27]	Access Community Health Network, a large system of federally qualified health centers, implemented P4P incentives in 2004 for absolute performance and improvement on large set of process and outcome measures. This study examines effects on HbA1c testing and control. Evaluated 1,166 patients treated by 46 PCPs. (out of 266 who treated diabetic patients in the federally qualified health centers) (2002–2004).	Avg. annual # of encounters per diabetic patient, % diabetic patients with any HbA1c test, % diabetic patients with recommended number of HbA1c tests, % diabetic patients with controlled blood sugar (HbA1c <7, HbA1c<9).	High performers remain at the top of the performance distribution.	Low-performing showed greatest improvement

Reference	Program Description and # of Providers Studied	Metric Assessed	Characteristics of High Performers	Characteristics of Low Performers
Damberg et al., 2010[7]	IHA program is a statewide P4P program in California for physician groups. Bonuses for meeting patient experience, process and outcome measures, and health information technology infrastructure. Study examined relationship between performance on P4P measures and use of care management processes. 180 physician groups.	Effect of care management processes on P4P composite performance measure (clinical processes of care).	The Care Management Process (CMP) index demonstrated significant positive associations with performance on 2 of the composite measures, namely diabetes management and intermediate outcomes. Higher performance in diabetes management (3.2 points higher on a 0–100 performance scale) was associated with substantial investments in CMPs (>5 CMPs on a 0–6 scale); each 1.0-point increase on the CMP index translated into a 1.0-point gain for the intermediate outcomes composite (P <.001). Higher engagement in external QI initiatives was significantly positively associated with the processes-of-care component; a 1.0-point increase on the QI index translated into a 1.4-point gain on the CMP index (P = .02). Among the control variables, medical group organization type was significantly associated with higher performance for 2 of the composite measures (3.0–4.6 points higher for medical groups compared with independent practice associations). Physician organization size was positively associated with higher performance on the processes-of-care composite (1.5 points) (P = .002). The net effect of increasing the number of physicians within a PO from 10 to 100 physicians on the log scale would translate into a 3.5-point gain for the processes-of-care composite, with an effect size of 1.5. We observed no relationship between Medicaid revenue and performance.	None reported
Doran et al., 2008[91]	UK National Health Service P4P program (2004–2007). Bonus payments to PCPs that achieve a threshold proportion of patients meeting quality targets for various clinical and patient experience measures. 7367 general primary care practices.	48 clinical activity indicators.	Characteristic with positive association with achievement was the exclusion rate (a 1% higher rate of exclusions was associated with a 0.35% higher rate of achievement in year 2 and 0.16% higher rate in year 3 (p<0.01)). Other associations that were positive (though modest) were the number of PCPs/10,000, the percentage of female PCPs, the percentage medically educated in the UK. Area deprivation scores were significantly associated with reported achievement, but association was very modest. Prior practice performance was associated with increase in achievement over time (the lower the achievement, the greater the increase in achievement).	Larger practice size, population density, the percentage of PCPs >50 years of age, and percentage of patients >65 of age were negatively associated with achievement (p<0.01).

Reference	Program Description and # of Providers Studied	Metric Assessed	Characteristics of High Performers	Characteristics of Low Performers
Doran et al., 2006[164]	The National Health Service funded $3.2 billion in 2004 to provide bonus payments to PCPs that achieve a threshold proportion of patients meeting quality targets. 8,105 practices with 1 or more family practitioners.	2004–2005 performance on 10 clinical quality indicators.	Achievement was higher in practices with a high ratio of family practitioners to patients. (p<.01) However, the multiple regression model explained only 20% of the variation between practices, and all of these effects were small.	Achievement was also lower in larger practices and in practices with a high proportion of family practitioners who received their medical education outside the United Kingdom or were 50 years of age or older, lower in practices that were on the Primary Medical Services contract. (p<.01)
Jha et al., 2010[88]	CMS Premier HQID incentivized hospital performance on 5 clinical conditions. Examined association between the DSH index and changes in performance for AMI, CHF, and pneumonia. 251 of 255 HQID hospitals compared with a national sample of 3017 hospitals. (2003 (4th quarter) and July 2006–June 2007).	Association between the disproportionate share index and baseline quality performance, changes in performance, and terminal performance for AMI, CHF, and pneumonia.	High DSH index was associated with greater improvements for AMI and pneumonia.	Higher DSH index was associated with lower performance for AMI, CHF, and pneumonia at baseline.
Lindenauer et al., 2007[59]	The HQID incentivized hospital performance on 5 clinical conditions. Study examined performance on 10 AMI, pneumonia, and CHF measures in HQID and control hospitals. 613 hospitals part of a national public reporting initiative, 207 of which participated in HQID.	10 individual process measures of AMI, CHF, and pneumonia and composite scores for AMI, CHF, pneumonia, and all combined were considered in HQID and control hospitals.	Largest improvements among hospitals with the poorest baseline performance for CHF. In HQID hospitals, improvement on the composite of the 10 examined process measures was 16.1% for hospitals in lowest quintile and 1.9% for those in highest quintile at baseline (p<0.001).	Not reported

Reference	Program Description and # of Providers Studied	Metric Assessed	Characteristics of High Performers	Characteristics of Low Performers
Nicholas et al., 2011[54]	The HQID incentivized hospital process measures for 5 clinical conditions. Classified HQID process measures as easy or difficult to improve based on whether they introduce additional per-patient costs and compared process compliance on easy and difficult tasks at hospitals eligible for HQID bonuses relative to hospitals engaged in public reporting. 145 (with sufficient data)/255 completing the 3 year HQID; 1089 control hospitals publicly reporting to Hospital Compare. (2002–2005)	Process-of-care measures. Classified incentivized tasks as easy or difficult to improve by considering additional per-patient costs. Hospitals categorized into quintiles based on performance on process composite score in year 1.	Fail to find statistically significant effects for P4P hospitals at either end of the initial quality distribution relative to hospitals with average scores.	Not reported
Rosenthal et al., 2005[10]	PacifiCare implemented a P4P program in California, incentivizing patient experience and process measures, but did not implement a P4P program in the Pacific Northwest. Medical group performance was compared between those in California and those in the Pacific Northwest. Sample of 167 medical groups contracting with Pacificare in California exposed to a financial incentive and 42 medical groups in the Northwest not exposed to the incentive.	Cervical cancer screening, mammography, and hemoglobin A1c testing. Total potential dollars that could have been distributed in each quarter and the total, average, and max payouts. Number of groups in each quarter that received any bonus and the number that reached at least half of the targets.	75% of the dollars were earned by groups that had achieved the benchmarks prior to the incentive program. Physician groups with baseline performance at or above the target improved the least. Mammography rates of physician groups with baseline performance at or above the target improved by only 0.7%, whereas physician groups more than 10% below the target at baseline improved 6.6% (p=0.07). Groups below but within 10% of the target, and physician groups more than 10% below the target were statistically significant for cervical cancer screening (p=0.03; p=0.02).	Not reported

Reference	Program Description and # of Providers Studied	Metric Assessed	Characteristics of High Performers	Characteristics of Low Performers
Ryan, 2012a[58]	Evaluated the HQID, which incentivized hospital performance on 5 clinical conditions. 266 hospitals (250 HQID hospitals and 250 comparison hospitals).	Composite process quality scores for heart attack, CHF, and pneumonia for the HQID and matched hospitals (250 HQID and 250 non-HQID).	Not reported	The HQID hospitals in the lowest quartile demonstrated more improvement than their matched comparison hospitals in phase I, but not phase II. No evidence that HQID hospitals in the lowest initial quartile had greater improvement in performance in phase II.
Sutton et al., 2012[72]	A hospital P4P program modeled off the US Hospital Quality Incentive Demonstration (same indicators and incentives) was implemented in all 24 National Health Service hospitals with an emergency care department in the Northwest region of England. Only top quartile hospital performers received bonus payments equal to 4% of revenue from national tariff from associated activity. 24 hospitals in northwest region, 132 hospitals in all other English regions.	Patient-level changes in mortality by condition.	Small hospitals and hospitals rated as having “excellent” or “good” quality services by the national regulator before the program showed the largest mortality reductions. (not significant effect).	Not reported

Reference	Program Description and # of Providers Studied	Metric Assessed	Characteristics of High Performers	Characteristics of Low Performers
Vina et al., 2009[97]	The HQID incentivized hospital performance on 5 clinical conditions. Surveyed QI leaders at HQID hospitals in the top 2 or bottom 2 deciles of performance. 84 out of 92 hospitals in top (45) and bottom (39) 2 deciles of performance completed surveys.	Overall Composite Quality Score for year 2 of HQID across all 5 conditions. Hospitals with data on 3+ conditions were categorized by deciles. Only the top and bottom 2 deciles were included for analysis. Conducted phone interviews with hospitals focused on QI interventions, data feedback, leadership, organizational support for QI, and organizational culture.	A greater proportion of top-performing hospitals had a CABG surgery program (p=0.01) and a greater proportion of low performers had a slightly higher percentage of Medicaid patients (p=0.02). More top than bottom performers used order sets for treating hip and knee replacements (91.1% vs. 64.1%, p<0.01). More top than bottom performers reported using clinical pathways for the treatment of AMI (48.9% vs. 15.4%, p<0.01), CHF (44.4% vs. 17.9%, p<0.01), pneumonia (37.8% vs. 12.8%, p<0.01), and hip and knee replacement (55.6% vs. 23.1%, p<0.01). More top than bottom performers had a multidisciplinary team with the goal of improving care for AMI (93.3% vs. 76.9%, p<0.05) and CHF (93.3% vs. 69.2%, p<0.01). More top than bottom performers used computerized physician order entry systems (24.4% vs. 7.9%, p<0.05). No significant difference between top and bottom performers with condition-specific educational programs for physicians and nurses, discussion in general forums, public display of hospital data, % of chief medical officers who had the general role of QI, or % who could identify 1+ physician champions per clinical condition (p>0.05). No significant difference in use of order sets for AMI, CHF, pneumonia, and CABG, but use was relatively high in both groups. Mean levels of agreement to statements on organizational support for QI were generally similar, however mean levels of agreement were higher in top performers on statements about nursing staff's support for quality indicators (mean=1.78 vs. 2.28, p<0.01) and adequate human resources for initiatives to improve quality indicator performance (mean=2.18 vs. 2.82, p< 0.01). More top-performing hospitals leaned toward disagreeing with the statement, "Coordinating quality care across different departments is difficult to do at this hospital" (mean=3.53 vs. 2.87, 5-point Likert Scale, p<0.01). In response to a statement about changes taking place very slowly at their organization, top performers were generally neutral (mean=3.49) and bottom performers tended to agree (mean=2.23), (p<.01). Top performers were more likely to agree with their hospitals' propensity to try new initiatives or policies whereas bottom performers tended to be more neutral (mean=2.84 vs. 3.10, p<0.01). Mean level of disagreement with the statement that their institution tended to blame to individuals when something goes wrong was relatively greater in the top performers than in bottom performers (mean= 4.51 vs. 4.05, p<0.05).	See high performers column.

Reference	Program Description and # of Providers Studied	Metric Assessed	Characteristics of High Performers	Characteristics of Low Performers
Werner et al., 2011[56]	The HQID incentivized hospital performance on 5 clinical conditions. Evaluated performance compared with control group. 260 out of 267 hospitals that joined in FY 2004; 780 control hospitals.	Hospital Compare data on AMI, pneumonia, and CHF and calculated the composite scores for pneumonia and CHF (excluded AMI composite because data missing mortality measure) for HQID and control hospitals. Compared performance btw the 2 groups and the change in distribution over time (cumulative % of hospitals meeting the performance thresholds after P4P implementation. Hospitals were stratified based on proxy calculations of bonuses received using the Medicare revenue for incentivized conditions divided by the total hospital Medicare revenue; effects of market competition using the Herfindahl-Hirschmann Index score of the Hospital Service Area; and the baseline financial status by taking the average total margin of the 4 years pre-P4P implementation.	Improvements were largest among hospitals that were eligible for larger bonuses, were well financed, or operated in less competitive markets.	Not applicable

11. What are the characteristics of the lowest-performing providers/institutions and their behaviors in response to VBP?

Regarding the characteristics of low-performing providers under P4P programs, the following were identified:

- Physician organizations' practice sites were located in lower-SES areas (based on the SES of individuals living within the zip codes of the practice sites).[43]
- Physician organizations with higher percentages of Medicaid patients were less likely to be in highest two quintiles of performance.[43]
- Higher DSH index was associated with lower performance on hospital measures at baseline.[88]
- The UK study[91, 164] found that larger practice size, population density, the percentage of PCPs >50 years of age, higher percentage of PCPs who received medical education outside the UK, and the percentage of patients >65 years of age were negatively associated with achievement.

Strength of Evidence: Insufficient. Few studies have addressed this issue, so we lack a full understanding of what characterizes high (or low) performers. Studies have been opportunistic in defining the characteristics based on the variables that were available to them, rather than considering a prior a set of characteristics that might differentiate providers who are low versus high performers.

12. How much does it cost a provider/institution to improve on the measured performance areas?

12a. Are the incentive levels of VBP programs sufficient to cover the costs of investing in quality improvement?

12b. How do organizations weight these factors related to VBP and decide on quality improvement investments?

Overall, few studies exist that address these questions. A study by Mehrotra et al.[158] that was based on interviews with 79 physician group leaders in Massachusetts found variation in the percentage that reported QI initiatives related to specific measures (from 12 percent reporting QI efforts focused on hypertension control to 61 percent reporting QI efforts for HbA1c measurement). A key finding from this study was that the most common QI investment was the development of an internal registry and feedback system for physicians regarding their performance. This study also queried physician leaders on whether the incentives were large enough to motivate quality improvement. Most reported that incentives of 5 percent or more would be required to increase their emphasis on quality improvement; other drivers of QI investments were the clinical importance of the quality measures, the costs and effectiveness of available QI initiatives, the structure of the physician group, the group's operating margin, and the fraction of revenue from the payer making the incentive payment.

Similarly, a 2009 study of 35 physician organizations participating in the IHA P4P program in California[4] found widespread support (28 of 35) for increasing incentives at the organization level to 5–10 percent of capitation payments, which would increase physicians' attention and provide a positive return on investment to the organizations by defraying setup and compliance costs. It was also noted by both health plans and physician organizations five years into the experiment that modest improvements in performance highlighted the need to assess the opportunity costs of investing in P4P versus other types of strategies to drive quality improvement.

In a study by Pham et al.,[165] the authors conducted interviews in 2004–2005 with quality officers in 36 hospitals to understand the impact of hospital reporting programs (e.g., Hospital Quality Initiative, Leapfrog, Joint Commission) on quality improvement (i.e., budgets, setting of priorities, staffing levels). The hospitals reported that they had increased resources for quality measurement and improvement and, per Pham, "believed that reporting increases hospital costs, for both compliance and processes to improve performance." Half of the hospitals interviewed in this study reported adding up to 12 full-time equivalent staff dedicated to quality improvement and reporting. This study also found that for less financially healthy hospitals, reporting and quality improvement was a significant cost burden. More broadly, it was generally hard for the 36 hospitals to assess the net cost burden, as they could not easily measure the impact of improved outcomes on their finances, and the costs are spread over many hospital cost centers. Burdens were especially heavy for data collection, underscoring the inadequacy of health information technology systems at the time this study was conducted. Hospitals indicated that they tend to invest in programs (i.e., set priorities) based on the availability of evidence-based interventions that are available from such organizations as the Institute for Healthcare Improvement and state quality improvement organizations. This helped hospitals minimize the resources they had to deploy searching for evidence-based interventions.

A qualitative study of P4P to assess hospital executives' perspectives on disparities and P4P[163] reported that hospital executives were concerned about the resources required to respond to incentives, in particular good health IT systems to report on quality measures, which most did not have at that stage.

Strength of Evidence: Not applicable, descriptive only.

Improving the Performance of Value-Based Purchasing Programs

13. What are the critical gaps in knowledge about VBP, and how can these gaps be addressed?

Because rigorous evaluation methods were often not used in studies assessing the effects of P4P, the findings in the literature do not provide a good picture on whether P4P programs work and how much improvement can be expected beyond other efforts to improve quality. There is consensus among those conducting evaluations that P4P experiments should have rigorous evaluations with comparison groups to determine their impact on health care quality and

resource use; however, given that many P4P programs are implemented universally within a given setting (e.g., statewide, within a health plan, or nationally), finding a comparison group is challenging. Additionally, longitudinal data on performance on the measures that are the focus of the P4P program prior to the start of the intervention rarely exists to perform interrupted time series analyses.

The authors of the published studies we reviewed identified a number of topics for future research (Table 3.10). There was strong support for the need to understand incentive structures (e.g., size, type, target of) and how other program design features affect performance, the effectiveness of different measures, and contextual and provider characteristics that are associated with performance results.

Table 3.10. Critical Gap Areas Identified in the Pay-for-Performance Literature

Areas Identified for Future Research and Evaluation	Pay-for-Performance
Incentive structure	• Conduct research on different incentive designs/structures (size, type, level of risk, target of incentive) and how various incentive designs influence performance. • Determine how much of the positive effect on low-performing providers derives from rewarding both absolute performance and improvement.
Measures	• Assess effectiveness of individual performance measures. • Evaluate which measures contributed most to the decrease in the cost trend. • Determine what are the right measures to use to drive the desired behavior changes and achieve goals.
Disparities	• Examine whether reduced variation in quality leads to reduction in inequalities. • Examine the effects of P4P programs on disparities and how to mitigate those effects.
Outcomes	• Track outcomes expected to result from P4P interventions that focus on improved care processes. • Examine the impact of financial incentives on quality when the incentives are implemented for the purpose of controlling resource use or cost of care.
Provider characteristics and contextual factors and their relationship to P4P effects	• Assess how the effects of P4P programs vary with respect to factors such as patient population, health plan-physician contracts, and physician practice characteristics. • Explore whether staff, infrastructure, and IT support lead to more improvement on process measures and outcomes. • Examine results on performance measures by degree of systemic support (e.g., disease management programs, community initiatives).
Unintended/spillover effects	• Monitor P4P programs and the effects of different reward structures on performance and the distribution of incentive payments. • Assess the potential negative effects of certain types of measures on provider behavior (e.g., measures with specific control targets, outcomes). • Evaluate P4P's potential spillover effects on unmeasured areas.

Areas Identified for Future Research and Evaluation	Pay-for-Performance
Other gaps	• Examine relative contribution of public reporting on the quality of care versus P4P on improvements in quality or outcomes. • Assess how the design of the incentive program affects its impact on performance. • Explore the conditions under which implementation of incentives for clinical targets or patient registries yields sufficient improvements in quality to justify the investment. • Identify driving force(s) of the improvement or lack of improvement across incentivized measures. • Assess physician understanding of the P4P program. • Understand what changes providers are making in response to P4P.

14. What are the structural and implementation features of the most successful P4P programs?

The design and implementation of P4P programs (or any VBP program) matters in terms of how successful the intervention will be. Only a handful of studies addressed this question.

In the study of a physician group P4P program by Mullen et al.,[42] the impact on performance increased with the size of the average expected reward.

Werner et al.[56] compared 260 Premier HQID hospitals with a group of comparison hospitals (n=780) and found that HQID hospitals initially improved more than the control group, but by the end of five years, the two groups' scores were virtually identical. The authors noted that larger incentives had a greater effect on changing performance. The response to P4P incentives was larger, and appeared to be more sustained, among hospitals eligible for a large bonus, compared with those eligible for a small bonus.

The study by Pearson et al.[5] of P4P programs introduced into physician group contracts from 2001 to 2003 by five major commercial health plans in Massachusetts found no relationship between the magnitude of quality improvement and specific P4P contracts. Their qualitative analysis did not find any obvious distinctive features of "successful" or "unsuccessful" P4P contracts. The authors flagged that for one of the groups that had high performance, the combined potential incentives were worth approximately $1,900 per PCP for performance on two diabetes measures.

In a small RCT of physician P4P in California, Chung et al.[103] found that varying the frequency of bonus payment (annual versus quarterly) did not affect performance on the process or outcome measures.

Another study of P4P in five Medicaid plans identified two characteristics that were associated with the more successful programs: (1) incentives that were high enough to compensate for the effort required by a provider to obtain them and (2) good communication with providers. The study also reported that plans' efforts to support quality improvement were helpful when provided, such as lists of children about to turn 15 months old and pre-addressed reminder cards that the physician offices could send to patients.[44]

Several studies also comment on the need to involve key stakeholders in the P4P system design and implementation.[4, 105]

Strength of Evidence: Insufficient.

15. Within VBP programs, how can practices from the highest-performing providers/institutions be disseminated?

Strength of Evidence: Insufficient. The literature review did not address this question. Please refer to Chapter Six for the summary of the TEP's discussion.

16. To what extent can VBP programs that have a positive impact in health care be improved and expanded?

Strength of Evidence: Insufficient. The literature review did not address this question. Please refer to Chapter Six for the summary of the TEP's discussion.