Published in final edited form as:

Med Care Res Rev. 2017 February ; 74(1): 3–78. doi:10.1177/1077558715619282.

Pay-for-Performance: Disappointing Results or Masked Heterogeneity?

Adam A. Markovitz and

University of Michigan Medical School and University of Michigan School of Public Health

Andrew M. Ryan

University of Michigan School of Public Health

Abstract

Research on the effects of pay-for-performance (P4P) in health care indicates largely disappointing results. This central finding, however, may mask important heterogeneity in the effects of P4P. We conducted a literature review to assess whether hospital and physician performance in P4P varied by patient and catchment area factors, organizational and structural capabilities, and P4P program characteristics. Several findings emerged: organizational size, practice type, teaching status, and physician age and gender modify performance in P4P. For physician practices and hospitals, a higher proportion of poor and minority patients is consistently associated with worse performance. Other theoretically influential characteristics – including information technology and staffing levels – yield mixed results. Inconsistent and contradictory effects of bonus likelihood, bonus size, and marginal costs on performance in P4P suggest organizations have not responded strategically to financial incentives. We conclude that extant heterogeneity in the effects of P4P does not fundamentally alter current assessments about its effectiveness.

Keywords

pay-for-performance; financial incentives; quality; organizational variation

Introduction

Health care's market failures are myriad. Information asymmetries, agency problems, supplier-induced demand, the non-marketability of health, and the many sources of uncertainty limit the power of supply and demand to set prices, determine quantities, advance quality, and efficiently allocate limited health care resources (Arrow, 1963; Cutler, 2000; Newhouse, 2002). In an effort to promote higher quality health care, private and public payers have experimented with an array of supply- and demand-side programs to improve quality and cost outcomes in a failing marketplace. Although initially undertaken by private payers in piecemeal fashion, payment experiments have grown in number and scope. These efforts include hospital report cards, accountable care organizations (ACOs), tiered insurance networks, bundled payments, and P4P programs. P4P programs, also known

as value-based purchasing (VBP), link provider reimbursement with performance on a set of defined quality measures (Damberg et al., 2014). The logic that underlies P4P – that individuals and institutions respond rationally to incentives and that physicians will take charge of patient outcomes and improve quality if they have “skin in the game” – is attractive.

Given the urgency of improving value in health care, the appeal of financial incentives, and the lack of other politically expedient options, it should perhaps come as no surprise that governmental payers have seized upon P4P as an integral component of health care reform (Tanenbaum, 2009). In 2003, the UK National Health Service introduced the Quality of Outcomes Framework (QOF), a comprehensive P4P program among essentially all of its general practitioners. That same year, the US Centers for Medicare & Medicaid Services (CMS) began the Hospital Quality Incentive Demonstration (HQID), a P4P experiment among several hundred hospitals. Most recently, the 2010 Patient Protection and Affordable Care Act set forth an ambitious agenda for implementing VBP across Medicare settings, including hospitals, physician groups, skilled nursing facilities, home health agencies, hospices, and dialysis facilities. The Hospital Readmissions Reduction Program (HRRP) and the Hospital Value-Based Purchasing Program (HVBP) both began in 2012, while the Physician Value-Based Payment Modifier Program (PVBPM) began in 2015 among a subset of physicians and will be expanded to include all physicians and practices by 2017 (Ryan & Press, 2014).

Political support for P4P is substantial and bipartisan (Tanenbaum, 2009), but empirical support is more limited. A recent review of P4P by the RAND Corporation found mixed and relatively weak evidence that P4P improved clinical outcomes or reduced costs (Damberg et al., 2014). The variation that underlies these inconsistent and generally modest effects, however, is not well understood. Given the widespread adoption of P4P in the US, it is imperative that we gain a more refined sense of the patient, provider, practice, and programmatic characteristics associated with P4P performance.

New Contribution

As P4P programs have proliferated over the past two decades, so too have their attendant analyses, systematic reviews, and, most recently, a systematic review of systematic reviews (Eijkenaar, Emmert, Scheppach, & Schoffski, 2013). These reviews generally conclude that P4P has yielded largely disappointing results. The average effects, however, may mask important sources of heterogeneity. While many empirical studies include data on P4P variation across various factors, investigators often fail to discuss the significance of this heterogeneity. Instead, these sub-analyses are often left uncommented upon in results tables or appendices. Reviews of P4P have followed suit. Although several systematic reviews have included sections on variation, most focus solely on racial or ethnic differences (Alshamsan, Lee, Majeed, Netuveli, & Millett, 2012; Steel & Willems, 2010). Van Herck and colleagues conducted an extensive systematic review that included variation by system, payer, physician, and patient characteristics (Van Herck et al., 2010); however, the review was conducted in 2010 and did not present the details of their findings, making it difficult to assess the credibility of the authors' conclusions. Furthermore, several potentially important

sources of heterogeneity (e.g., information technology, share of low-income patients) were not addressed in the review.

The aim of our review is to systematically assess whether hospital and physician performance in P4P varies by patient and catchment area factors, organizational and structural capabilities, and P4P program characteristics. Understanding the determinants of P4P performance will help payers to more rationally design future iterations of P4P programs and will enable providers to more efficiently allocate resources so as to improve quality and respond to payer demands. This will prove particularly important as P4P is extended across governmental and commercial settings in the coming years.

Conceptual framework

Research to date has identified several factors that may directly affect clinical quality and modify the effectiveness of P4P programs. Drawing upon and modifying a theoretical framework developed by Rittenhouse and colleagues (Rittenhouse et al., 2010), these factors can be broadly categorized as patient and catchment area factors, organizational structure and capabilities, and P4P program characteristics (see Figure 1). Patient factors along with organizational structure and capabilities determine the degree to which practices can effectively respond to incentives such as P4P. Catchment area factors, meanwhile, represent more distal resources that shape providers' environment and affect care delivery. Implicit in this framework is that the effects of P4P programs affect quality and cost outcomes in concert with patient, provider, and organizational factors.

There are a number of mechanisms by which patient and catchment factors and organizational structure and capabilities may both directly influence quality and modify the effects of P4P programs. Organizations with greater shares of poor or minority patients may have fewer financial resources available to invest in quality improvement due to worse payer mix and higher costs of caring for vulnerable patients (Casalino, Elster, et al., 2007). Organizational capabilities such as greater information technology use and higher physician-to-patient ratios may enable organizations to better respond to (and document) the complex initiatives demanded by P4P programs (Bardach et al., 2013). Baseline performance, quality improvement (QI) participation, and organizational culture likely represent past and present organizational learning and motivation and future capacities for adapting to P4P demands. The direct effects of P4P incentives may also vary across organizations, as incentive structure can interact with organizational characteristics to affect bonus likelihood, bonus size, and marginal costs. Economic theory predicts that organizations faced with larger bonuses and greater chances of reaching performance targets should respond more robustly than those with smaller and less realistic incentives.

Conceptually, patient and organizational factors may either directly affect baseline quality levels or indirectly modify quality changes over time in response to P4P program. Well-executed longitudinal studies can disentangle direct effects from effect modification by assessing how levels and changes in quality vary across a given organizational factor. In cross-sectional studies of new P4P programs, however, the degree to which success in P4P reflects historical performance that preceded the program compared to responsiveness to the program is often unclear. This makes it challenging for researchers to identify whether

differences in baseline quality or responsiveness to the program incentives is the primary driver of variation in performance in P4P programs. The relative importance of variation in levels versus changes in P4P is ultimately a matter of perspective. Those who either sponsor (payers) or evaluate (researchers) P4P programs generally care most about predictors of P4P effectiveness and thus prioritize decomposing baseline from longitudinal effects. Yet “effectiveness” from the provider perspective is more likely defined by achieving bonus payments through meeting quality targets, an outcome that is a function of provider performance at baseline and over time. In light of these theoretical, empirical, and policy considerations, we include data from both longitudinal and cross-sectional analyses and treat sources of heterogeneity as “modifying” or “direct” effects, respectively.

Method

We sought to identify all published P4P evaluations that included sub-analyses by patient and catchment area factors, organizational structure and capabilities, and P4P program characteristics. The search utilized a strategy employed by a recent RAND review of P4P publications (Damberg et al., 2014) in order to capture studies published since this comprehensive review. The following combination of MeSH terms was used for MEDLINE: “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])).

This search strategy identified a total of 432 studies published between December 07, 2012 and January 15, 2015. In addition, we screened the 195 studies included in the RAND review. These 627 titles and abstracts underwent preliminary screening for relevant content. We next utilized the PubMed related citations function and conducted several ad hoc Google Scholar searches to identify additional potentially pertinent publications. Last, we supplemented these searches by identifying several organizational reports not available on PubMed, including those released by Medicare Payment Advisory Commission (MedPAC), Kaiser Permanente, and the Commonwealth Fund.

Studies were included if they satisfied all of the following criteria: (1) full text available in English; (2) published between 2000 and 2015; (3) evaluated a program that provided financial bonuses or penalties for performance on pre-specified measures; (4) program was implemented in the US, UK, or Canada; (5) provided qualitative or quantitative empirical analysis; (6) was conducted among physician practices and hospitals; and (7) reported sub-analyses by patient and catchment factors, organizational structure and capabilities, or P4P program characteristics. For the third criterion, our definition excludes ACO programs, which we view as broader programs aimed at systems change. Due to challenges posed by comparing results across diverse care settings, we excluded studies from dialysis facilities, skilled nursing homes, home health agencies, and hospice facilities.

Results

We identified 58 studies that satisfy our inclusion criteria (see Appendix Table A1 for further description of studies analyzed in this review). These studies cover 21 unique P4P programs (see Appendix Table A2 for full descriptions of programs):

- National programs: CMS's HQID, HRRP, and HVBP; NHS's QOF and Advancing Quality (AQ) programme;
- Private health plan sector programs: PacifiCare Quality Incentive Program (QIP), Hawaii Blue Cross Blue Shield P4P, Bronx CHAMPION P4P, Midwestern generic prescription rate (GPR) program, Palo Alto Medical Clinic (PAMC) P4P;
- Regional collaboratives: California Integrated Healthcare Association (IHA) physician organization (PO) program, Veterans Affairs (VA) hypertension P4P, ACCESS Community Health Network of Federally Qualified Health Centers P4P, Minnesota smoking referral P4P, Ontario Primary Care Reform (PCR), Ontario Diabetes Management Incentive (DMI); and
- Medicaid and state programs: Hudson Health Plan Immunization P4P, Hudson Health Plan Diabetes P4P, New York City Department of Health and Mental Hygiene Primary Care Information Project (PCIP), Local Initiative Rewarding Results (LIRR) Collaborative Demonstration, Massachusetts MassHealth.

As “performance” under P4P is a multidimensional concept and varied by program and study definitions, we considered the full spectrum of metrics (e.g., achievement of quality measures, receipt of financial bonuses) and sought to specify performance measures throughout this review. Following our conceptual framework, we present factors in three broad categories: patient and catchment factors, organizational structure and capabilities, and P4P program characteristics.

Patient and catchment area factors

34 papers evaluated the relationship between patient and catchment area factors and performance in P4P programs. The most commonly assessed dimensions were socioeconomic status (SES) (20 studies), followed by race or ethnicity (17 studies) and age (7 studies). 17 of these studies were in the context of hospital programs and 27 were in the context of physician practice programs.

P4P programs targeting hospitals

There is substantial evidence that low patient SES is associated with lower baseline quality performance and attenuated improvement over time. The results are strongest for hospital programs and generally weaker for physician programs. Hospital executives expressed concern about the resources required to respond to P4P programs, while safety net executives were particularly worried that P4P could worsen resource constraints (Weinick, Chien, Rosenthal, Bristol, & Salamon, 2010). Quantitative evidence reinforces these concerns. Hospitals with greater shares of Medicaid patients or dual-eligible patients demonstrated higher and more variable cross-sectional readmission rates and are thus predicted to face higher odds of being penalized under HRRP (Berenson & Shih, 2012;

CMS, 2012; Gu et al., 2014; Joynt & Jha, 2013; Shih, Ryan, Gonzalez, & Dimick, 2014). Two recent studies of early HRRP and HVBP data appear to confirm projections that safety net hospitals, defined as those in the top quartile of Disproportionate Share Hospital (DSH) index, will experience disproportionately worse financial consequences under HRRP and HVBP (Gilman et al., 2014; Ryan, 2013). Recently, MedPAC demonstrated a strong and practically monotonic relationship between hospitals' share of patients on supplemental security income (SSI) and HRRP readmission penalties (MedPAC, 2014). California safety net hospitals achieved lower mortality rates but were nonetheless more likely to be penalized under HBVP due to worse patient experience and process scores (Gilman et al., 2014). Similarly, while safety net hospitals improved more over time under the HQID (Jha, Orav, & Epstein, 2010), they received lower bonuses due to lower baseline performance (Ryan, Blustein, Doran, Michelow, & Casalino, 2012); this gap in bonus achievement decreased when the incentive structure shifted from absolute thresholds to relative improvement but in part reflected a broader diminution in improvement among all participants (Ryan, Blustein, Doran, et al., 2012).

Evidence of the relationship between patients' race and ethnicity and P4P performance is less consistent. Executives of safety net hospitals indicated that P4P might shift resources away from programs important for minority patients (e.g., interpreters) and exacerbate racial disparities by penalizing lower performing safety net hospitals and promoting cherry picking (Weinick et al., 2010). Contrary to these concerns and earlier reports of CABG access disparities widening following public reporting programs (Werner, Asch, & Polsky, 2005), it does not appear that the HQID worsened access among racial or ethnic minorities over time (Ryan, 2010). Early studies of the HRRP, however, indicate that minority-serving hospitals have worse and more variable cross-sectional readmission rates (CMS, 2012) and will thus fare worse financially (Shih, Ryan, et al., 2014). In contrast to hospital findings, there is no consistent evidence that physician practice performance is modified by share of minority patients.

Several studies have assessed the differential effects of the HQID by market competitiveness under the hypothesis that P4P responses should be attenuated in more competitive markets. Finding that HQID hospitals in more competitive markets initially improved less over time on performance-score threshold achievement, Werner, Kolstad, Stuart, and Polsky argued that, while financial incentives may not increase the incremental incentive to improve in competitive markets that already have robust public reporting, P4P may represent a direct mechanism by which stimulate quality improvement in less competitive markets where public reporting may be less effective – implying that competitiveness may negatively modify P4P improvement (Werner, Kolstad, Stuart, & Polsky, 2011). In two other studies of the HQID, however, neither 30-day mortality (Jha, Joynt, Orav, & Epstein, 2012) nor CABG access (Epstein, Joynt, Jha, & Orav, 2014) improvements were modified by market competitiveness.

P4P programs targeting physician practices and organizations

Physician practice studies generally use share of low-income patients, area-level SES, or patient insurance status as proxies for patient-level SES. A mixed-methods analysis of

Medicaid-focused health plans revealed that patients' lack of transportation, housing instability, and focus on economic subsistence posed significant barriers to physician practices improving quality in response to P4P programs (Felt-Lisk, Gimm, & Peterson, 2007). Larger shares of low-income patients (Doran et al., 2006; Doran, Fullwood, Reeves, Gravelle, & Roland, 2008; Gravelle, Sutton, & Ma, 2008) and lower catchment area SES (Chien, Wroblewski, et al., 2012; Doran, Fullwood, Kontopantelis, & Reeves, 2008; Kontopantelis, Buchan, Reeves, Checkland, & Doran, 2013) were generally associated with lower performance at baseline and over time. It may be of little surprise that physicians express concern that quality measures do not adequately adjust for patient SES (Casalino, Alexander, Jin, & Konetzka, 2007).

Medicaid revenue was not associated with levels of early IHA performance on individual measures (Damberg et al., 2010) but was negatively associated with composite IHA performance levels in a follow-up analysis (Chien, Wroblewski, et al., 2012). Performance changes over time were not modified by insurer type in either the ACCESS (Coleman, Reiter, & Fulwiler, 2007) or Hudson Health Plan immunization P4P programs (Chien, Li, & Rosenthal, 2010).

There is some suggestive evidence catchment area factors such as higher population density or rurality may have harmed quality performance at baseline (Gravelle et al., 2008; Kontopantelis et al., 2013) and response to the QOF (Doran, Fullwood, Kontopantelis, et al., 2008; Doran, Fullwood, Reeves, et al., 2008). Evidence of the direct and modifying effects of patient age (Chien, Eastman, Li, & Rosenthal, 2012; Doran et al., 2006; Doran, Fullwood, Kontopantelis, et al., 2008; Doran, Fullwood, Reeves, et al., 2008; Gravelle et al., 2008; Kontopantelis et al., 2013; Ryan & Doran, 2012) and health status (Bardach et al., 2013; Chien, Eastman, et al., 2012; Chien et al., 2010; Hamilton et al., 2014) is mixed and inconclusive.

Organizational structure and capabilities

39 papers evaluated the modifying effect of organizational structure and capabilities on P4P performance. The most frequently included dimensions were organizational size (16 studies) baseline performance (15 studies), and practice type (6 studies). 15 of these studies were in the context of hospital programs and 24 were in the context of physician practice programs.

P4P programs targeting hospitals

Larger hospital size has been consistently related to lower baseline quality in the HRRP (Joynt & Jha, 2013; MedPAC, 2014) and impaired quality improvements over time under the HQID (Lindenauer et al., 2007) and the UK's AQ (Sutton et al., 2012). Although the relationship between the number of beds and readmission rates may indicate supply-sensitive care (Wennberg, 2002), the finding that higher-occupancy hospitals experienced higher readmissions (MedPAC, 2014) suggests that this relationship may instead reflect the greater difficulties faced by larger hospitals in assuring uniform quality.

Hospital teaching status was associated with smaller HQID improvements (Lindenauer et al., 2007) and higher levels of readmissions rates in the HRRP (Joynt & Jha, 2013; MedPAC, 2014). In a qualitative study of high and low HQID performers, cross-sectional performance

was related to physician leadership (e.g., recruitment of QI champions), organizational culture (e.g., support for care coordination), and QI factors (e.g., use of multidisciplinary teams) (Vina, Rhew, Weingarten, Weingarten, & Chang, 2009). There is surprisingly little evidence on the effect of information technology on P4P performance among hospitals. HQID performance was positively related to computerized physician order entry but not compliance with report collection (Vina et al., 2009).

The hypothetical effects of baseline performance are uncertain. While theory predicts that initial low performers should face higher costs (Mullen, Frank, & Rosenthal, 2010), regression to the mean and ceiling effects imply that the marginal costs of quality improvement should be increasing in quality. Empirical investigations into the modifying effects of baseline performance have yielded similarly unpredictable and contradictory results. In the HQID, one study failed to observe variation by initial performance quintile (Nicholas, Dimick, & Iwashyna, 2011) while another found that quality improvement was inversely related to initial performance quintile (Lindenauer et al., 2007). Under the AQ, higher baseline hospital quality (as rated by the national regulator against core standards, existing national targets, and new national targets for quality) was related to greater mortality reductions for pneumonia but not AMI or HF (Sutton et al., 2012). In the HVBP, patient experience or clinical process changes did not vary by initial performance (Ryan, Burgess, Pesko, Borden, & Dimick, 2014).

P4P programs targeting physician practices and organizations

The effect of practice size varied across and within P4P programs, measures, and patient groups. Larger IHA POs consistently outperformed smaller POs in cross-sectional IHA studies (Chien, Wroblewski, et al., 2012; Damberg et al., 2010), with larger POs reporting greater capacity for quality monitoring (Damberg, Raube, Teleki, & Dela Cruz, 2009). In the Ontario PCR, larger practices increased provision of incentivized measures modestly more than their small counterparts (Li, Hurley, DeCicca, & Buckley, 2014). Under the QOF, larger practice size generally attenuated practice performance at baseline (Doran et al., 2006; Doran, Fullwood, Kontopantelis, et al., 2008; Doran, Fullwood, Reeves, et al., 2008; Kontopantelis et al., 2013) and over time (Doran, Fullwood, Kontopantelis, et al., 2008). There is some indication that larger practices disproportionately excluded patients from quality measurements, potentially to “game” the QOF incentive design (Doran, Fullwood, Reeves, et al., 2008).

Neither clinic nor physician specialty modified responses to the smoking cessation (An et al., 2008) or ACCESS (Coleman, Reiter, & Fulwiler, 2007) programs. For Hudson Health Plan, private practices improved relatively more than Federally Qualified Health Centers in the immunization program (Chien et al., 2010) but not diabetes program (Chien et al., 2012). Medical groups outperformed independent practice associations in cross-sectional IHA analyses (Chien, Wroblewski, et al., 2012; Damberg et al., 2010).

Perhaps surprisingly, there is no consistent evidence that information technology affects quality or modifies P4P responses. An early study of Massachusetts Health Quality Partners found that EMR usage was not associated with QI initiatives targeting a given HEDIS measure (Mehrotra et al., 2007). High performing IHA physician organizations (POs) placed

relatively greater importance on IT capability (Damberg et al., 2009). However, both high and low performers reported difficulties monitoring quality at PO and physician levels. Under the QOF, meanwhile, the type of computing system was a stronger predictor of performance changes over time than any other practice or population characteristic (Kontopantelis et al., 2013). Intriguingly, differences across systems were stronger for intermediate health outcomes than for measurement or treatment indicators.

In general, lower baseline performance positively modified P4P responses at both the physician (Chen et al., 2010; Coleman et al., 2007) and practice (Doran, Fullwood, Kontopantelis, et al., 2008) level. However, the effects of the IHA program did not systematically vary by baseline performance (Mullen et al., 2010). History of QI involvement was the only characteristic that predicted a differential response to a smoking cessation P4P program: clinics with less QI involvement responded more robustly to the P4P than did peer controls (An et al., 2008). Participation in external QI initiatives was also related to better cross-sectional IHA process scores (Damberg et al., 2010).

Despite the challenges of drawing inferences about physician demographics from practice-level data, there is reason to believe that physician age and gender modify the effect of P4P. Practices with greater shares of younger physicians and female physicians demonstrated better baseline quality and responded more robustly to the QOF (Doran, Fullwood, Kontopantelis, et al., 2008; Doran, Fullwood, Reeves, et al., 2008; Gravelle et al., 2008; Kontopantelis et al., 2013). The negative modifying effects of physician age were also observed in differential responses to Ontario's PCR (Li et al., 2014). Greater response to P4P may not map onto greater support for P4P, however: a survey of general internists found that male, rather than female, physicians were more likely to support financial incentives and physician-specific public reporting (Casalino, Alexander, et al., 2007).

P4P Program Characteristics

12 articles evaluated the relationship between P4P program characteristics and performance. Potential bonus size (9 studies) and bonus likelihood (3 studies) were the most commonly included dimensions. Six of these articles described hospital P4P programs and six described physician practice programs.

Payers generally implement P4P programs in a uniform fashion across a defined population of hospitals or physician organizations. This lack of programmatic variation defies empirical attempts to capture heterogeneity across potentially important but theoretically ambiguous P4P program characteristics (e.g., selection of process vs. outcome measures). Nonetheless, a few studies utilized mid-program changes to estimate the causal effect of P4P program characteristics within relatively stable populations of providers (Mullen et al., 2010; Ryan, Blustein, Doran, et al., 2012). Several more studies exploited the fact that organizational characteristics (e.g., baseline quality performance) and incentive structures (e.g., absolute vs. relative achievement) interact to effectively create variation in the P4P incentives faced by organizations (Bhattacharyya, Freiberg, Mehta, Katz, & Ferris, 2009; Dowd, Feldman, & Nersesian, 2013; Epstein et al., 2014; Jha et al., 2012; Kantarevic & Kralj, 2013; Nicholas, Dimick, & Iwashyna, 2011; Rosenthal, Frank, Li, & Epstein, 2005; Werner et al., 2011).

P4P programs targeting hospitals

One randomized trial examined whether the effects of P4P was modified by whether financial incentives were targeted at the physician or the clinic level in an outpatient hospital setting (Petersen et al., 2013). Individual-level but not clinic-level or combined incentives led to greater blood pressure control or appropriate response to uncontrolled blood pressure. Meanwhile, none of the incentive schemes was related to greater use of guideline-recommended medications or the development of hypotension, a potential unintended consequence.

Most studies of hospital P4P characteristics drew on HQID data and yielded results that were largely contrary to economic theory. The lowest performing hospitals at baseline improved the most under HQID Phase 1 (Glickman et al., 2007; Lindenauer et al., 2007), despite an incentive structure that precluded all but the highest performing hospitals from achieving bonuses. Yet, when the HQID changed incentive structures so as to allow lower baseline performing hospitals to share in financial bonuses, improvements among the lower performers fell off, while those that stood to gain the most (above the median cutoff but below the achievement thresholds) did not respond more robustly than those below the median cutoff (Ryan, Blustein, & Casalino, 2012).

In a study that exploited differential exposure to HQID incentives, Nicholas, Dimick, and Iwashyna found no consistent evidence that hospitals preferentially improved on “easy” (low-cost, easy-to-improve) versus “difficult” (high-cost, difficult-to-improve) tasks for incentivized measures (Nicholas et al., 2011). Although P4P was associated with a one percentage point gain in compliance for easy versus hard tasks for heart attack admissions, this pattern did not hold for heart failure admissions and may have reversed for pneumonia admissions. Nicholas et al. further exploited the fact that not all hospitals faced the same costs to substitute on such tasks; namely, those hospitals with relatively more patients eligible for an easy task (i.e., lower marginal cost ratios for substitution) should be more likely to substitute toward easier tasks than hospitals with higher marginal cost ratios. Yet this did not occur.

Potential HQID bonus size, meanwhile, was conceptualized in different ways by different studies, potentially leading to inconsistent estimates of variation across that dimension. HQID responses worth modified by the share of Medicare payments linked to incentivized conditions (Werner et al., 2011) but not share of Medicare beneficiaries (Jha et al., 2012); the former measured achievement of performance thresholds 2004–2008, however, while the latter measured 30-day mortality 2003–2009. Hospital volume, meanwhile, was positively related to cross-sectional HQID surgical quality (Bhattacharyya et al., 2009) but did not modify P4P responses for AML, HF, or pneumonia quality (Nicholas, Dimick, & Iwashyna, 2011). Interestingly, surgical quality was not significantly associated with cross-sectional mortality, UTI rates, or iatrogenic complications. Despite perverse incentives to the contrary, these analyses collectively suggest that hospitals did not strategically alter resource allocations so as to maximize HQID performance scores while minimizing monetary costs and effort.

P4P programs targeting physician practices and organizations

The effect of payment frequency was assessed in a physician-level trial of financial incentives (Chung, Palaniappan, Wong, Rubin, & Luft, 2010). The P4P program did not improve quality in either treatment arm nor did the null effect vary by payment frequency (either 4 quarterly bonuses or a single year-end bonus). However, the study design could not separate the effect of quarterly payment from the effect of quarterly reporting. A more recent cluster-randomized trial of small primary care clinics assessed the effect of disbursing greater payments for meeting quality targets in patient populations considered more difficult to treat (i.e., patients with comorbidities, Medicaid, or uninsurance). There was a non-significantly greater improvement among difficult-to-treat patients as compared to other patients, suggesting larger payments may motivate greater quality improvement; nonetheless, the lack of a non-tiered incentive group precludes analytic decomposition of patient and payment factors.

A study of medical groups in California contracting with a large network HMO, PacifiCare, exploited the successive implementation of two programs, PacifiCare's QIP and IHA's P4P, to examine the differential effects of bonus size and absolute vs. relative thresholds. Under PacifiCare's QIP threshold incentive structure, the lowest performers improved most but captured only 5% of the bonuses, while the highest performers improved least but captured 75% of the bonuses (Rosenthal et al., 2005). Meanwhile, those practices with the most to gain by improvement (below but within 10% of the target threshold) improved significantly less than lower performing groups. Moreover, when PacifiCare switched from the threshold incentives of the QIP to the IHA (in which larger payments were awarded for improvements in progressively higher strata), groups in higher baseline strata did not consistently improve more than those in lower strata despite greater financial incentives to do so (Mullen et al., 2010). And, although potential bonus size increased (rising from roughly 10% to 60% of capitated revenues) and became more achievable (as bonuses were awarded for relatively lower performance), groups did not appear to respond more robustly to the IHA than the QIP.

Dowd et al. recently asked whether low performing practices "give up" when faced with an absolute threshold, and, if so, at what "distance" from the targeted rate of prescribing generic medications (Dowd et al., 2013). Practices below the target were equally likely to improve as those above the target, and practices up to 22% below the target continued to increase their rate of prescribing generic medications. This could reflect the fact that providers react to perceived longer-term incentives, rather than short-term incentives for a given performance period. Lastly, the effect of the Ontario PCR was paradoxically negatively modified by expected income gain and share of enrolled patients at baseline (Kantarevic & Kralj, 2013).

Discussion

There is a pressing need to move from volume-based to value-based payments to health care providers. Theoretically, P4P presents a clear mechanism by which payers can reward clinical outcomes rather than clinical activity. Empirically, however, P4P presents a murkier picture.

Among patient factors, one of the clearest findings was that practices and hospitals that serve poor patients and patients of color fared worse under P4P (Jha et al., 2010; Ryan, Blustein, Doran, et al., 2012). Performance did not systematically vary by any other patient factors, including insurance type, gender, age, or patient health (Coleman et al., 2007; Hamilton et al., 2014). A few factors related to organizational structure and capabilities did appear to modify performance, with large teaching hospitals generally demonstrating worse performance at baseline and over time under P4P (Joynt & Jha, 2013; Lindenauer et al., 2007). Under the QOF, smaller practices and those with higher proportions of female and younger physicians responded more robustly to P4P than other practices (Doran, Fullwood, Reeves, et al., 2008; Gravelle et al., 2008; Kontopantelis et al., 2013). In cross-sectional IHA analyses, medical groups and larger POs outperformed independent practice associations and smaller POs, respectively (Chien, Wroblewski, et al., 2012; Damberg et al., 2010).

Several organizational factors yielded conflicting results despite strong *a priori* reasons to expect enhanced response to P4P: prior QI engagement (An et al., 2008; Damberg et al., 2010), higher staffing levels (An et al., 2008; McHugh, Berez, & Small, 2013; Ryan & Doran, 2012), and greater financial strength (Epstein et al., 2014; Sutton et al., 2012). The evidence for information technology was particularly inconsistent. While high performing IHA POs placed greater importance on IT capability than did low performing POs (Damberg et al., 2009), both high and low performing POs reported difficulty in monitoring quality performance, and EMR capabilities were not related to performance levels (Damberg et al., 2010). Cross-sectional performance under the HQID was related to computerized physician order entry but not compliance report collection and reporting (Vina et al., 2009). On the other hand, one study of the QOF found that the type of computing system was a stronger modifier of P4P responses than any other practice and population characteristic (Kontopantelis et al., 2013).

Analyses of the modifying effects of P4P program characteristics suggest that P4P is not motivating quality improvement in a manner anticipated by payers or predicted by economic theory. Organizations with the most to gain financially did not necessarily improve; those that improved the most did not necessarily stand to receive large financial awards; the tasks on which institutions improved were not necessarily the easiest; nor were the institutions that did improve those with the lowest marginal costs of substitution (Kantarevic & Kralj, 2013; Mullen et al., 2010; Nicholas et al., 2011; Rosenthal et al., 2005; Ryan, Blustein, & Casalino, 2012). For better or worse, low performing practices do not appear to give up, even when they are unrealistically far from reaching high performance targets (Dowd et al., 2013).

There is a paucity of research that isolates and quantifies heterogeneity across P4P program characteristics. This gap proceeds from two causes. First, P4P program sponsors do not generally vary program characteristics while implementing incentive programs. When variation is introduced during program implementation, it is generally unintentional and non-random (exceptions include (Bardach et al., 2013; Petersen et al., 2013). Second, and relatedly, articles on P4P have almost exclusively focused on the effects of particular P4P programs in defined populations of providers and patients; we found only two articles that compared head-to-head the effects of different P4P programs (Mullen et al., 2010) or mid-

implementation changes (Ryan, Blustein, & Casalino, 2012) within the same P4P program among a single group of providers. Thus, published research has primarily captured variation along provider and patient dimensions, not P4P program characteristics.

A number of commentaries have attempted to fill this empirical vacuum by making normative claims about how P4P programs ought be designed (Jha, 2013; Mehrotra, Sorbero, & Damberg, 2010). Several literature reviews, meanwhile, have sought to indirectly measure the effect of various P4P design choices by comparing the effectiveness of different P4P programs and then inferring the relative influence of various program characteristics (e.g., whom, what, how, and when to target financial incentives) (Eijkenaar et al., 2013; Van Herck et al., 2010). Yet, this approach does not allow investigators to measure the effect of one programmatic characteristic (e.g., frequency of incentives) while holding constant other factors.

It is unlikely that payers will begin randomizing providers and patients to different combinations of P4P design schemes. The few that have used such an approach have, at times, lacked treatment arms critical for distinguishing between fundamental programmatic features (e.g., payment levels vs. patient factors, payment frequency vs. reporting frequency). Instead, capturing heterogeneity across program characteristics will require that investigators creatively combine multiple quality databases so as to compare the effects of multiple P4P programs in the same provider populations. Although this would not control for the effect of time-varying factors – the sequential nature of P4P programs would make it particularly difficult to separate programmatic effects from temporal effects – this would allow investigators to analyze heterogeneity across program characteristics while holding time-invariant characteristics relatively constant.

Our review has several limitations. The external validity of our findings is limited by restriction of our search to P4P programs conducted among physician practices and hospitals in the US, UK, and Canada. While this reflected our concerns for comparability, our findings should not be generalized to other countries or practice settings. Although our search was comprehensive and all results systematically collected, paper inclusion was conducted by one author and thus subject to greater bias than a more systematic search approach. Our results are limited to primarily qualitative conclusions insofar as we did not seek to quantitatively combine and weight effect modification across studies. Although we note concordance or disagreement between studies, we felt that it would be misleading to provide weighted averages of interaction terms given the heterogeneity in programmatic measures and incentive design and analytical estimation strategies and variable definition.

The threat of scientific error further limits our conclusions. While studies of P4P programs should ideally be designed with sub-analyses in mind, the presence of null results may simply reflect the possibility that analyses were not powered to detect effect modification. On the other hand, the literature may reflect a publication bias resulting from researchers failing to publish the results of null sub-analyses. Given these countervailing forces, it is unclear whether we should place greater confidence in our positive findings (e.g., share of low-income patients) or in our negative or inconclusive findings (e.g., hospital financial strength). It is important to note that several of the effect modifiers considered in our

analysis likely interact with one another, complicating attempts to cleanly delineate variation under P4P. For instance, we found that safety-net hospitals exhibit worse baseline performance but may improve more under P4P. Yet the likelihood of floor and ceiling effects and regression to the mean suggest that, in theory and in practice, it is quite difficult to disentangle the effects of baseline quality and true P4P responsiveness. Although some articles endeavored to assess the variation in responsiveness to P4P incentives according to baseline performance using quantile regression (Ryan, Burgess, Pesko, Borden, & Dimick, 2014), stratified analysis (Jha et al., 2012), and other methods (Chien, Eastman, et al., 2012), the high correlation between factors such as baseline quality and DSH index remain a fundamental empirical challenge.

Although many studies provided tangential evidence for heterogeneity of outcomes under P4P, relatively few studies sought to systematically examine whether P4P varied by patient or catchment factors, organizational structure and capabilities, or P4P program characteristics. Among those studies that explored the determinants of P4P performance, most researchers opportunistically selected the subset of characteristics available in the P4P dataset (e.g., Hospital Compare), rather than broader or more theoretically consequential characteristics that might require external linking (e.g., area-level income or literacy). There is further lack of measure and variable standardization between studies, as well as a dearth of patient- and provider-level data. These limitations reflect, in part, a departure in priorities and standards for program evaluation between academic researchers and P4P program administrators (Damberg et al., 2014). But in the absence of standardized, patient-level data, it is unclear, for instance, whether the inverse relation between DSH and P4P performance is due to within-unit variation (i.e., low-income patients receive worse care than high-income patients within the same hospitals), between-unit variation (i.e., low-income patients receive care at lower quality hospitals), or both. Estimation of heterogeneity in P4P performance and effectiveness more generally will require that payers and providers define and gather a standardized set of performance measures and organizational characteristics. Yet there remain tremendous opportunities to examine potentially important sources of heterogeneity given existing data sources. Herrin et al. recently found that county characteristics, including area SES, physician mix, and nursing home quality, explained nearly half of the total variation in hospital readmission nationwide (Herrin et al., 2015). Such studies highlight how broader conceptualization of the socioeconomic influences on P4P performance could substantially enrich the meager and mostly insignificant stock of variables that currently constitute catchment area factors, e.g., market competitiveness (Epstein et al., 2014), rurality (Gravelle et al., 2008), and population density (Doran, Fullwood, Kontopantelis, et al., 2008).

Marsh and McConnell put forward a framework for establishing policy success that focuses on three dimensions: process success, political success, and programmatic success (Marsh, 2010). Programmatic success, in turn, hinges on operational success (implementation as per objectives), outcome success (achievement of intended goals), economic success (efficient use of resources) and equity (not disproportionately benefiting or harming a particular class, community, or institution) (Marsh, 2010; Nuckols, 2015). P4P has not substantially improved quality (Damberg et al., 2014), and, when it has, the quality gains and concomitant financial bonuses appear largely concentrated among wealthier patients and the

organizations that serve them. These programmatic failures jeopardize P4P's political success and the other healthcare reforms that draw upon political capital for program implementation.

Our review gives suggestive evidence that much of the variation in performance is due to factors over which organizations have limited control. This finding threatens both programmatic and political success of P4P initiatives and suggests that we are treating these sources of performance heterogeneity as modifiers of quality performance when they may in fact represent confounders of quality measurement. Recently, an expert panel convened by the National Quality Forum recommended that the NQF reverse its policy to allow for SES risk adjustment of certain measures under certain conditions (NQF, 2014). This reflected concerns that measures that fail to account for patient complexity – whether clinical or sociodemographic – encourage provider avoidance of disadvantaged patients, promote payer and consumer avoidance of seemingly “poor” quality providers, and ultimately threaten the validity and fairness of comparative performance assessment. Of greatest concern is the real possibility that P4P will exacerbate health disparities by rewarding the highest achievers with scarce resources taken from organizations serving disadvantaged patients. How to best address this concern, however, depends in part on the nature of the apparent quality gap. Within-organization variation could be addressed by case-mix adjustment for sociodemographic measures or paying larger amounts for reaching quality goals for disadvantaged patients (Bardach et al., 2013). To the extent that between-hospital variation exists and that disadvantaged patients disproportionately receive care from lower-quality providers (Dimick, Ruhter, Sarrazin, & Birkmeyer, 2013), paying for quality improvement in addition to achievement might help but is likely insufficient to fully prevent payment redistributions across organizational characteristics (Damberg, Elliott, & Ewing, 2015). However, reinforcement theory suggests that behavioral change is most robust when the link between behavior and incentives is clear (Luthans, 1985); thus, the complexity of incentive structures that reward both achievement and improvement may confuse providers and ultimately fail to promote quality improvement (Ryan, Blustein, Doran, et al., 2012).

Our two primary findings – that institutions are not responding strategically to P4P incentives and that safety net providers are performing worse under P4P programs – pose a classic policy dilemma. P4P programs need to clarify incentive structures and increase financial incentives to motivate large and sustained quality improvements, but simplifying and raising the financial stakes will, in turn, escalate the risk of unintended consequences. Yet the tension between effectiveness and equity may prove a false dilemma. Rather than address the inequalities among P4P participants through complicated and ineffective tiered incentive structures, P4P programs could create homogenous competition pools for institutions that do not differ meaningfully with respect to the confounders of performance measurement (Ryan, 2013). This would minimize the undesired redistributive effects of P4P and avoid “masking” potential socioeconomic or racial disparities (MedPAC, 2014). While this presents a significant obstacle for P4P programs with smaller physician bases, national programs could easily offer peer institution groups of sufficient size.

Ultimately, it does not appear that P4P has succeeded in satisfying its goals either efficiently or equitably. While it seems that the main effects of P4P may be masking some

heterogeneity across organization types, this variation does not appear sufficient to alter the conclusion that P4P has largely failed to realize substantial quality improvements. There is something more fundamentally amiss with P4P. Understanding the origins of these failures will prove crucial as P4P is extended across Medicare and commercial insurers in the coming years.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Bibliography

- Alshamsan R, Lee JT, Majeed A, Netuveli G, Millett C. Effect of a UK pay-for-performance program on ethnic disparities in diabetes outcomes: interrupted time series analysis. *Ann Fam Med*. 2012; 10(3):228–234. DOI: 10.1370/afm.1335 [PubMed: 22585887]
- An LC, Bluhm JH, Foldes SS, Alesci NL, Klatt CM, Center BA, Manley MW. A randomized trial of a pay-for-performance program targeting clinician referral to a state tobacco quitline. *Arch Intern Med*. 2008; 168(18):1993–1999. DOI: 10.1001/archinte.168.18.1993 [PubMed: 18852400]
- Arrow KJ. Uncertainty and the Welfare Economics of Medical Care. *American Economic Review*. 1963; 53(5):941–973.
- Bardach NS, Wang JJ, De Leon SF, Shih SC, Boscardin WJ, Goldman LE, Dudley RA. Effect of pay-for-performance incentives on quality of care in small practices with electronic health records: a randomized trial. *JAMA*. 2013; 310(10):1051–1059. DOI: 10.1001/jama.2013.277353 [PubMed: 24026600]
- Berenson J, Shih A. Higher readmissions at safety-net hospitals and potential policy solutions. *Issue Brief (Commonwealth Fund)*. 2012; 34:1–16. [PubMed: 23289161]
- Bhalla R, Schechter CB, Strelnick AH, Deb N, Meissner P, Currie BP. Pay for performance improves quality across demographic groups. *Qual Manag Health Care*. 2013; 22(3):199–209. DOI: 10.1097/QMH.0b013e31829a6b4f [PubMed: 23807132]
- Bhattacharyya T, Freiberg AA, Mehta P, Katz JN, Ferris T. Measuring the report card: the validity of pay-for-performance metrics in orthopedic surgery. *Health Aff (Millwood)*. 2009; 28(2):526–532. DOI: 10.1377/hlthaff.28.2.526 [PubMed: 19276012]
- Casalino LP, Alexander GC, Jin L, Konetzka RT. General internists' views on pay-for-performance and public reporting of quality scores: a national survey. *Health Aff (Millwood)*. 2007; 26(2):492–499. DOI: 10.1377/hlthaff.26.2.492 [PubMed: 17339678]
- Casalino LP, Elster A, Eisenberg A, Lewis E, Montgomery J, Ramos D. Will pay-for-performance and quality reporting affect health care disparities? *Health Aff (Millwood)*. 2007; 26(3):w405–414. DOI: 10.1377/hlthaff.26.3.w405 [PubMed: 17426053]
- Chen JY, Kang N, Juarez DT, Hodges KA, Chung RS, Legorreta AP. Impact of a pay-for-performance program on low performing physicians. *J Healthc Qual*. 2010; 32(1):13–21. quiz 21–12. [PubMed: 20151587]
- Chien AT, Eastman D, Li Z, Rosenthal MB. Impact of a pay for performance program to improve diabetes care in the safety net. *Prev Med*. 2012; 55(Suppl):S80–85. DOI: 10.1016/j.ypmed.2012.05.004 [PubMed: 23046985]
- Chien AT, Li Z, Rosenthal MB. Improving timely childhood immunizations through pay for performance in Medicaid-managed care. *Health Serv Res*. 2010; 45(6 Pt 2):1934–1947. DOI: 10.1111/j.1475-6773.2010.01168.x [PubMed: 20849554]
- Chien AT, Wroblewski K, Damberg C, Williams TR, Yanagihara D, Yakunina Y, Casalino LP. Do physician organizations located in lower socioeconomic status areas score lower on pay-for-performance measures? *J Gen Intern Med*. 2012; 27(5):548–554. DOI: 10.1007/s11606-011-1946-8 [PubMed: 22160817]

- Chung S, Palaniappan L, Wong E, Rubin H, Luft H. Does the frequency of pay-for-performance payment matter?--Experience from a randomized trial. *Health Serv Res.* 2010; 45(2):553–564. DOI: 10.1111/j.1475-6773.2009.01072.x [PubMed: 20059568]
- CMS. Medicare Hospital Quality Chartbook 2012: Performance Report on Outcome Measures. Prepared by Yale New Haven Health Services Corporation Center for Outcomes Research and Evaluation for the Centers for Medicare and Medicaid Services. 2012. Retrieved November 14, 2014, from <http://cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/MedicareHospitalQualityChartbook2012.pdf>
- Coleman K, Reiter KL, Fulwiler D. The impact of pay-for-performance on diabetes care in a large network of community health centers. *J Health Care Poor Underserved.* 2007; 18(4):966–983. DOI: 10.1353/hpu.2007.0090 [PubMed: 17982218]
- Cutler, DM., Zeckhauser, RJ. The anatomy of health insurance. In: Culyer, AJN., P, J., editors. *Handbook of Health Economics.* 1. Vol. 1. Buckingham, U.K.: Elsevier; 2000. p. 563-643.
- Damberg CL, Elliott MN, Ewing BA. Pay-for-performance schemes that use patient and provider categories would reduce payment disparities. *Health Aff (Millwood).* 2015; 34(1):134–142. DOI: 10.1377/hlthaff.2014.0386 [PubMed: 25561654]
- Damberg CL, Raube K, Teleki SS, Dela Cruz E. Taking stock of pay-for-performance: a candid assessment from the front lines. *Health Aff (Millwood).* 2009; 28(2):517–525. DOI: 10.1377/hlthaff.28.2.517 [PubMed: 19276011]
- Damberg CL, Shortell SM, Raube K, Gillies RR, Rittenhouse D, McCurdy RK, Adams J. Relationship between quality improvement processes and clinical performance. *Am J Manag Care.* 2010; 16(8): 601–606. [PubMed: 20712393]
- Damberg, CL., Sorbero, ME., Lovejoy, SL., Martsolf, G., Raaen, L., Mandel, D. *Measuring Success in Health Care Value-Based Purchasing Programs: Summary and Recommendations.* Santa Monica, CA: RAND Corporation; 2014.
- Dimick J, Ruhter J, Sarrazin MV, Birkmeyer JD. Black patients more likely than whites to undergo surgery at low-quality hospitals in segregated regions. *Health Aff (Millwood).* 2013; 32(6):1046–1053. DOI: 10.1377/hlthaff.2011.1365 [PubMed: 23733978]
- Doran T, Fullwood C, Gravelle H, Reeves D, Kontopantelis E, Hiroeh U, Roland M. Pay-for-performance programs in family practices in the United Kingdom. *N Engl J Med.* 2006; 355(4): 375–384. DOI: 10.1056/NEJMsa055505 [PubMed: 16870916]
- Doran T, Fullwood C, Kontopantelis E, Reeves D. Effect of financial incentives on inequalities in the delivery of primary clinical care in England: analysis of clinical activity indicators for the quality and outcomes framework. *Lancet.* 2008; 372(9640):728–736. DOI: 10.1016/S0140-6736(08)61123-X [PubMed: 18701159]
- Doran T, Fullwood C, Reeves D, Gravelle H, Roland M. Exclusion of patients from pay-for-performance targets by English physicians. *N Engl J Med.* 2008; 359(3):274–284. DOI: 10.1056/NEJMsa0800310 [PubMed: 18635432]
- Dowd B, Feldman R, Nersesian W. Setting pay for performance targets: do poor performers give up? *Health Econ.* 2013; 22(2):168–179. DOI: 10.1002/hec.2773 [PubMed: 22302418]
- Eijkenaar F, Emmert M, Scheppach M, Schoffski O. Effects of pay for performance in health care: a systematic review of systematic reviews. *Health Policy.* 2013; 110(2–3):115–130. DOI: 10.1016/j.healthpol.2013.01.008 [PubMed: 23380190]
- Epstein AM, Joynt KE, Jha AK, Orav EJ. Access to coronary artery bypass graft surgery under pay for performance: evidence from the premier hospital quality incentive demonstration. *Circ Cardiovasc Qual Outcomes.* 2014; 7(5):727–734. DOI: 10.1161/CIRCOUTCOMES.114.001024 [PubMed: 25160840]
- Felt-Lisk S, Gimm G, Peterson S. Making pay-for-performance work in Medicaid. *Health Aff (Millwood).* 2007; 26(4):w516–527. DOI: 10.1377/hlthaff.26.4.w516 [PubMed: 17595198]
- Gilman M, Adams EK, Hockenberry JM, Wilson IB, Milstein AS, Becker ER. California safety-net hospitals likely to be penalized by ACA value, readmission, and meaningful-use programs. *Health Aff (Millwood).* 2014; 33(8):1314–1322. DOI: 10.1377/hlthaff.2014.0138 [PubMed: 25092831]

- Glickman SW, Ou FS, DeLong ER, Roe MT, Lytle BL, Mulgund J, Peterson ED. Pay for performance, quality of care, and outcomes in acute myocardial infarction. *JAMA*. 2007; 297(21):2373–2380. DOI: 10.1001/jama.297.21.2373 [PubMed: 17551130]
- Gravelle, H., Sutton, M., Ma, A. CHE Research Paper. York: Centre for Health Economics, University of York; 2008. Doctor Behaviour Under a Pay for Performance Contract: Further Evidence from the Quality and Outcomes Framework.
- Gray J, Millett C, Saxena S, Netuveli G, Khunti K, Majeed A. Ethnicity and quality of diabetes care in a health system with universal coverage: population-based cross-sectional survey in primary care. *J Gen Intern Med*. 2007; 22(9):1317–1320. DOI: 10.1007/s11606-007-0267-4 [PubMed: 17594128]
- Gu Q, Koenig L, Faerberg J, Steinberg CR, Vaz C, Wheatley MP. The Medicare Hospital Readmissions Reduction Program: potential unintended consequences for hospitals serving vulnerable populations. *Health Serv Res*. 2014; 49(3):818–837. DOI: 10.1111/1475-6773.12150 [PubMed: 24417309]
- Hamilton FL, Lavery AA, Gluvajic D, Huckvale K, Car J, Majeed A, Millett C. Effect of financial incentives on delivery of alcohol screening and brief intervention (ASBI) in primary care: longitudinal study. *J Public Health (Oxf)*. 2014; 36(3):450–459. DOI: 10.1093/pubmed/fdt121 [PubMed: 24375203]
- Herrin J, St Andre J, Kenward K, Joshi MS, Audet AM, Hines SC. Community factors and hospital readmission rates. *Health Serv Res*. 2015; 50(1):20–39. DOI: 10.1111/1475-6773.12177 [PubMed: 24712374]
- Jha AK. Time to get serious about pay for performance. *JAMA*. 2013; 309(4):347–348. DOI: 10.1001/jama.2012.196646 [PubMed: 23340633]
- Jha AK, Joynt KE, Orav EJ, Epstein AM. The long-term effect of premier pay for performance on patient outcomes. *N Engl J Med*. 2012; 366(17):1606–1615. DOI: 10.1056/NEJMsa1112351 [PubMed: 22455751]
- Jha AK, Orav EJ, Epstein AM. The effect of financial incentives on hospitals that serve poor patients. *Ann Intern Med*. 2010; 153(5):299–306. DOI: 10.7326/0003-4819-153-5-201009070-00004 [PubMed: 20820039]
- Joynt KE, Jha AK. Characteristics of hospitals receiving penalties under the Hospital Readmissions Reduction Program. *JAMA*. 2013; 309(4):342–343. DOI: 10.1001/jama.2012.94856 [PubMed: 23340629]
- Kantarevic J, Kralj B. Link between pay for performance incentives and physician payment mechanisms: evidence from the diabetes management incentive in Ontario. *Health Econ*. 2013; 22(12):1417–1439. DOI: 10.1002/hec.2890 [PubMed: 23203722]
- Kontopantelis E, Buchan I, Reeves D, Checkland K, Doran T. Relationship between quality of care and choice of clinical computing system: retrospective analysis of family practice performance under the UK's quality and outcomes framework. *BMJ Open*. 2013; 3(8):doi: 10.1136/bmjopen-2013-003190
- Li J, Hurley J, DeCicca P, Buckley G. Physician response to pay-for-performance: evidence from a natural experiment. *Health Econ*. 2014; 23(8):962–978. DOI: 10.1002/hec.2971 [PubMed: 23861240]
- Lindenauer PK, Remus D, Roman S, Rothberg MB, Benjamin EM, Ma A, Bratzler DW. Public reporting and pay for performance in hospital quality improvement. *N Engl J Med*. 2007; 356(5):486–496. DOI: 10.1056/NEJMsa064964 [PubMed: 17259444]
- Luthans, F., Kreitner, R. *Organizational Behavior Modification and Beyond: An Operant and Social Learning Approach*. Glenview, IL: Scott Foresman; 1985.
- Marsh D, McConnell A. Towards a framework for establishing policy success. *Public Administration*. 2010; 88(2):564–583.
- McHugh MD, Berez J, Small DS. Hospitals with higher nurse staffing had lower odds of readmissions penalties than hospitals with lower staffing. *Health Aff (Millwood)*. 2013; 32(10):1740–1747. DOI: 10.1377/hlthaff.2013.0613 [PubMed: 24101063]
- MedPAC. Report to the Congress: Medicare and the Health Care Delivery System. 2014. Retrieved December 17, 2014, from http://www.medpac.gov/documents/reports/jun14_entirereport.pdf?sfvrsn=0

- Mehrotra A, Pearson SD, Coltin KL, Kleinman KP, Singer JA, Rabson B, Schneider EC. The response of physician groups to P4P incentives. *Am J Manag Care*. 2007; 13(5):249–255. [PubMed: 17488190]
- Mehrotra A, Sorbero ME, Damberg CL. Using the lessons of behavioral economics to design more effective pay-for-performance programs. *Am J Manag Care*. 2010; 16(7):497–503. [PubMed: 20645665]
- Millett C, Gray J, Saxena S, Netuveli G, Khunti K, Majeed A. Ethnic disparities in diabetes management and pay-for-performance in the UK: the Wandsworth Prospective Diabetes Study. *PLoS Med*. 2007; 4(6):e191.doi: 10.1371/journal.pmed.0040191 [PubMed: 17564486]
- Mullen KJ, Frank RG, Rosenthal MB. Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. *Rand J Econ*. 2010; 41(1):64–91. [PubMed: 21667575]
- Newhouse JP. Why is there a quality chasm? *Health Aff (Millwood)*. 2002; 21(4):13–25.
- Nicholas LH, Dimick JB, Iwashyna TJ. Do hospitals alter patient care effort allocations under pay-for-performance? *Health Serv Res*. 2011; 46(1 Pt 1):61–81. DOI: 10.1111/j.1475-6773.2010.01192.x [PubMed: 21029088]
- NQF. Risk Adjustment for Socioeconomic Status or Other Sociodemographic Factors. 2014. Retrieved February 23, 2015, from http://www.qualityforum.org/Publications/2014/08/Risk_Adjustment_for_Socioeconomic_Status_or_Other_Sociodemographic_Factors.aspx
- Nuckols TK. County-level variation in readmission rates: implications for the Hospital Readmission Reduction Program's potential to succeed. *Health Serv Res*. 2015; 50(1):12–19. DOI: 10.1111/1475-6773.12268 [PubMed: 25630850]
- Petersen LA, Simpson K, Pietz K, Urech TH, Hysong SJ, Profit J, Woodard LD. Effects of individual physician-level and practice-level financial incentives on hypertension care: a randomized trial. *JAMA*. 2013; 310(10):1042–1050. DOI: 10.1001/jama.2013.276303 [PubMed: 24026599]
- Rittenhouse DR, Shortell SM, Gillies RR, Casalino LP, Robinson JC, McCurdy RK, Siddique J. Improving chronic illness care: findings from a national study of care management processes in large physician practices. *Med Care Res Rev*. 2010; 67(3):301–320. DOI: 10.1177/1077558709353324 [PubMed: 20054057]
- Rosenthal MB, Frank RG, Li Z, Epstein AM. Early experience with pay-for-performance: from concept to practice. *JAMA*. 2005; 294(14):1788–1793. DOI: 10.1001/jama.294.14.1788 [PubMed: 16219882]
- Ryan AM. Has pay-for-performance decreased access for minority patients? *Health Serv Res*. 2010; 45(1):6–23. DOI: 10.1111/j.1475-6773.2009.01050.x [PubMed: 19840137]
- Ryan AM. Will value-based purchasing increase disparities in care? *N Engl J Med*. 2013; 369(26):2472–2474. DOI: 10.1056/NEJMp1312654 [PubMed: 24369072]
- Ryan AM, Blustein J. The effect of the MassHealth hospital pay-for-performance program on quality. *Health Serv Res*. 2011; 46(3):712–728. DOI: 10.1111/j.1475-6773.2010.01224.x [PubMed: 21210796]
- Ryan AM, Blustein J, Casalino LP. Medicare's flagship test of pay-for-performance did not spur more rapid quality improvement among low-performing hospitals. *Health Aff (Millwood)*. 2012; 31(4):797–805. DOI: 10.1377/hlthaff.2011.0626 [PubMed: 22492897]
- Ryan AM, Blustein J, Doran T, Michelow MD, Casalino LP. The effect of Phase 2 of the Premier Hospital Quality Incentive Demonstration on incentive payments to hospitals caring for disadvantaged patients. *Health Serv Res*. 2012; 47(4):1418–1436. DOI: 10.1111/j.1475-6773.2012.01393.x [PubMed: 22417137]
- Ryan AM, Burgess JF Jr, Pesko MF, Borden WB, Dimick JB. The Early Effects of Medicare's Mandatory Hospital Pay-for-Performance Program. *Health Serv Res*. 2014; doi: 10.1111/1475-6773.12206
- Ryan AM, Doran T. The effect of improving processes of care on patient outcomes: evidence from the United Kingdom's quality and outcomes framework. *Med Care*. 2012; 50(3):191–199. DOI: 10.1097/MLR.0b013e318244e6b5 [PubMed: 22329994]
- Ryan AM, Press MJ. Value-based payment for physicians in Medicare: small step or giant leap? *Ann Intern Med*. 2014; 160(8):565–566. DOI: 10.7326/M13-1715 [PubMed: 24733199]

- Shih T, Nicholas LH, Thumma JR, Birkmeyer JD, Dimick JB. Does pay-for-performance improve surgical outcomes? An evaluation of phase 2 of the Premier Hospital Quality Incentive Demonstration. *Ann Surg*. 2014; 259(4):677–681. DOI: 10.1097/SLA.0000000000000425 [PubMed: 24368657]
- Shih T, Ryan AM, Gonzalez AA, Dimick JB. Medicare's Hospital Readmission Reduction Program in Surgery May Disproportionately Affect Minority-Serving Hospitals. *Ann Surg*. 2014; doi: 10.1097/SLA.0000000000000778
- Steel N, Willems S. Research learning from the UK Quality and Outcomes Framework: a review of existing research. *Qual Prim Care*. 2010; 18(2):117–125. [PubMed: 20529473]
- Sutton M, Nikolova S, Boaden R, Lester H, McDonald R, Roland M. Reduced mortality with hospital pay for performance in England. *N Engl J Med*. 2012; 367(19):1821–1828. DOI: 10.1056/NEJMsa1114951 [PubMed: 23134382]
- Tanenbaum SJ. Pay for performance in Medicare: evidentiary irony and the politics of value. *J Health Polit Policy Law*. 2009; 34(5):717–746. DOI: 10.1215/03616878-2009-023 [PubMed: 19778930]
- Vamos EP, Pape UJ, Bottle A, Hamilton FL, Curcin V, Ng A, Millett C. Association of practice size and pay-for-performance incentives with the quality of diabetes management in primary care. *CMAJ*. 2011; 183(12):E809–816. DOI: 10.1503/cmaj.101187 [PubMed: 21810950]
- Van Herck P, De Smedt D, Annemans L, Remmen R, Rosenthal MB, Sermeus W. Systematic review: Effects, design choices, and context of pay-for-performance in health care. *BMC Health Serv Res*. 2010; 10:247.doi: 10.1186/1472-6963-10-247 [PubMed: 20731816]
- Vina ER, Rhew DC, Weingarten SR, Weingarten JB, Chang JT. Relationship between organizational factors and performance among pay-for-performance hospitals. *J Gen Intern Med*. 2009; 24(7): 833–840. DOI: 10.1007/s11606-009-0997-6 [PubMed: 19415390]
- Weinick RM, Chien AT, Rosenthal MB, Bristol SJ, Salamon J. Hospital executives' perspectives on pay-for-performance and racial/ethnic disparities in care. *Med Care Res Rev*. 2010; 67(5):574–589. DOI: 10.1177/1077558709354522 [PubMed: 20054056]
- Wennberg JE. Unwarranted variations in healthcare delivery: implications for academic medical centres. *BMJ*. 2002; 325(7370):961–964. [PubMed: 12399352]
- Werner RM, Asch DA, Polsky D. Racial profiling: the unintended consequences of coronary artery bypass graft report cards. *Circulation*. 2005; 111(10):1257–1263. DOI: 10.1161/01.CIR.0000157729.59754.09 [PubMed: 15769766]
- Werner RM, Kolstad JT, Stuart EA, Polsky D. The effect of pay-for-performance in hospitals: lessons for quality improvement. *Health Aff (Millwood)*. 2011; 30(4):690–698. DOI: 10.1377/hlthaff.2010.1277 [PubMed: 21471490]

Figure 1.
Conceptual Framework for Determinants of Performance in Pay-for-Performance Programs

Table 1

Articles Evaluating Whether Patient or Catchment Area Factors Modify P4P Performance

P4P Programs Targeting Hospitals					
Effect	Study	Program	Design	Variable	Results
Modifying effects	(Weinick et al., 2010)	Multiple programs	Qualitative	Safety net provider	Safety net hospitals were particularly concerned that P4P could worsen resource constraints in safety net hospitals. Executives were broadly concerned about the resources required to respond to P4P programs and questioned whether P4P was intended to improve quality or contain costs.
	(Jha et al., 2010)	HQID	DID	DSH	The association between DSH and improvement was significantly higher among participants than non-participants for both AMI and pneumonia. By the end of Phase 1, higher DSH hospitals had caught up to their HQID counterparts on AMI and pneumonia.
	(Ryan, Blustein, Doran, et al., 2012)	HQID	DID	DSH	Large gaps existed across DSH quartiles for both receipt of any payment and payment per discharge in Phase 1. These gaps decreased significantly between Phases 1 and 2.
	(Ryan, 2010)	HQID	DDD	Patient race/ethnicity	Changes in rates of CABG or AMI, HF, or pneumonia admissions did not vary among any racial or ethnic group (i.e., black, Hispanic, non-white) relative to the white group.
	(Werner et al., 2011)	HQID	LA	Market competitiveness	HQID participants in less competitive markets (HHI) responded more vigorously to incentives than participants in more competitive markets between 2003 and 2006. These differences eroded as non-participants caught up to participants.
	(Jha et al., 2012)	HQID	DID	Market competitiveness	30-day, RA mortality did not vary across market competitiveness (HHI) across all incentivized conditions (AMI, CABG, HF, pneumonia) and for each condition individually.
	(Epstein et al., 2014)	HQID	DID	Market competitiveness	Change in CABG rates among AMI patients did not vary by market competitiveness (HHI) between participant and control hospitals.
	(Ryan & Blustein, 2011)	MassHealth	DID	Urbanicity	The effect of the P4P did not vary by urbanicity.
	(Joynt & Jha, 2013)	HRRP	CSA	DSH	Safety net hospitals were projected as twice as likely to be penalized as non-safety net hospitals.
	(Berenson & Shih, 2012).	HRRP	CSA	DSH	Safety net hospitals were 1.34, 1.29, and 1.26 times more likely than lowest DSH quartile hospitals to have above-average readmission rates for AMI, HF, and pneumonia.
Direct effects	(Gu et al., 2014)	HRRP	CSA	Share of dual-eligible patients	Both patient dual-eligibility status and higher hospital shares of dual patients were related to higher readmission rates for all HRRP conditions (AMI, HF, and pneumonia).
	(CMS, 2012)	HRRP	CSA	Share of Medicaid patients	Hospitals with the highest shares of Medicaid patients demonstrated greater variation and slightly higher RSRRs than lowest share hospitals. Hospitals with the most Medicaid patients demonstrated higher overall RSRR (17.1% vs. 16.5%) and higher median RSRRs

						for AMI (20.2% vs. 19.5%), HF (25.2% vs. 24.5%), pneumonia (18.7% vs. 17.9%), but not for THR/TKR.
	(MedPAC, 2014)	HRRP	CSA	SSI decile		There was a strong and almost monotonic relationship between share of Medicare patients qualifying for SSI and readmission penalties. SSI was a stronger predictor of readmissions than race or DSH, with average readmission penalties more than double in the highest SSI decile than in the lowest SSI decile (0.45% vs. 0.21%) in 2013.
	(Gilman et al., 2014)	HRRP & HVB	CSA	DSH		Safety net hospitals in California were projected to be more likely to be penalized under HRRP than non-safety net hospitals (88.3% vs. 68.1%). A marginally greater proportion of safety net hospitals may face penalties under HVB (70.0% vs. 58.2%) due to significantly worse patient experience scores (24.4% vs. 34.2%) and marginally worse process scores (53.8% vs. 60.0%). Safety net hospitals in California had significantly lower mortality rate indices across the three incentivized conditions (0.91 vs. 1.02) and marginally higher efficiency (\$7,688 vs. \$7,973).
	(Ryan, 2013)	HVB	CSA	DSH		DSH was inversely related to achievement, improvement, and payment adjustments.
	(CMS, 2012)	HRRP	CSA	Share of African-American patients		Hospitals with the highest shares of African-American patients demonstrated higher overall RSRR (17.5% vs. 16.6%) and higher median RSRRs for AMI (median RSRR of 20.4% vs. 19.2%), HF (25.8% vs. 24.4%), pneumonia (19.1% vs. 17.9%), but not THR/TKR as compared to lowest share hospitals.
	(Shih, Ryan, et al., 2014)	HRRP	CSA	Share of African-American patients		Minority-serving hospitals demonstrated higher CABG readmission rates than non-minority serving hospitals (19.2% vs. 17.2%). Minority-serving hospitals were likely to be penalized more frequently (60.8% vs. 32.2%) and to pay higher penalties (\$112 million vs. \$41 million).
	(MedPAC, 2014)	HRRP	CSA	Rurality		Rural hospitals fared worse than urban hospitals for average penalties as either a share of operating payments (0.30% vs. 0.34%) or a share of all inpatient payments (0.28% vs. 0.23%).
P4P Programs Targeting Physician Practices and Organizations						
Modifying effects	(Casalino, Alexander, et al., 2007)	Multiple programs	QA	Patient SES		Most physicians believed that quality measures did not adequately adjust for SES. Physicians' share of Medicaid patients was not related to support for financial incentives or public reporting.
	(Felt-Lisk et al., 2007)	LIRR	QA	Medicaid-focused health plans		Lack of excess capacity and funds to expand outreach, failure of walk-in clinics to schedule routine appointments, and lack or underuse of technology were cited as barriers to practice improvement.
	(**Harrison et al., 2014)	QOF	LA	Share of income-deprived patients		Practices with above median levels of income-deprived patients achieved slightly greater reductions in admissions for incentivized ambulatory care sensitive conditions.
	(Doran, Fullwood, Kontopantelis, et al., 2008)	QOF	LA	Area-level deprivation		Area-level deprivation was associated with reported achievement but not improved achievement over time after adjusting for initial performance.

(Ryan & Doran, 2012)	QOF	LA	Area-level deprivation	Area deprivation did not modify the relationship between composite process changes and composite outcome changes.
(Hamilton et al., 2014)	QOF+	LA	Area-level deprivation	Patients from the most deprived areas were more likely to receive alcohol screening compared to those from the least deprived areas.
(Doran, Fullwood, Kontopantelis, et al., 2008).	QOF	LA	Population density	Population density negatively modified reported achievement in years 1–2 but not years 2–3 of the QOF.
(Chien et al., 2010)	Hudson Health	DID and ITS	Patient insurance status	Immunization rates did not vary by insurance type (Medicaid vs. the State Children's Health Insurance Program).
(Coleman et al., 2007)	ACCESS	Pre-post	Patient insurance status	Private insurance was positively related to diabetes care and control at baseline but not improvements over time.
(Bardach et al., 2013)	PCIP	LA	Patient insurance status	Among the subgroup of uninsured or Medicaid (non-HMO) patients, intervention clinics demonstrated non-significantly greater improvements for all performance measures (range: +7.9%–12.9%) except cholesterol control (–0.33%).
(Doran, Fullwood, Kontopantelis, et al., 2008).	QOF	LA	Share of ethnic minorities	Race/ethnicity was associated with greater increases in reported exclusion rates and smaller improvements in reported achievement between years 2 and 3 (but not 1 and 2).
(Weinick et al., 2010)	No specific program	QA	Share of ethnic minorities	Hospital executives worried that P4P might exacerbate racial disparities by penalizing lower performing safety net hospitals and by promoting cherry picking. Safety net hospital executives expressed concern that P4P programs may shift resources away from other programs important for minority patients (e.g., patient education, interpreters).
(Coleman et al., 2007)	ACCESS	Pre-post	Patient race/ethnicity	Race/ethnicity was related to baseline diabetes care/control but not longitudinal changes.
(Chien et al., 2010)	Hudson Health	DID and ITS	Patient race/ethnicity	Full and timely immunization was higher among patients of Hispanic descent but did not vary between Black vs. White patients. Preexisting racial/ethnic disparities were not significantly altered by the intervention.
(Bhalla et al., 2013)	CHAMPION	LA	Patient race/ethnicity	Quality improvement did not vary by demographic subgroups (including African American/black race, Hispanic/Latino ethnicity, and Spanish preferred language).
(Millett et al., 2007)	QOF	LA	Patient race/ethnicity	Black Caribbean patients demonstrated significantly less improvement in HbA1c (AOR: 0.75; 95% CI: 0.57, 0.97) and BP control (AOR: 0.65; 95% CI 0.53, 0.81) as compared to white British patients.
(Alshamsan et al., 2012).	QOF	LA	Patient race/ethnicity	Changes in intermediate outcome measures did not vary consistently by race/ethnicity.
(Hamilton et al., 2014)	UK QOF+	LA	Patient race/ethnicity	Black patients were more likely receive alcohol screening than white patients (73.5% vs. 67.9%) but white patients were more likely screen positive than other ethnicities (14.6% vs. rates ranging from 3.4%–8.6%). Among those screening positive, black patients had lower odds of receiving behavioral interventions (AOR 0.37; 95% CI: 0.27, 0.52).

Direct effects	(Chien et al., 2012)	Hudson Health	DID and ITS	Patient age	Changes in quality of care or care utilization did not vary by patient age.
	(Doran, Fullwood, Kontopantelis, et al., 2008).	QOF	LA	Share of older patients	Share of patients over 65 years of age was inversely related to reported exception rates.
	(Ryan & Doran, 2012)	QOF	LA	Share of older patients	Share of patients over 65 years of age modified the process-outcome relationship for diabetes and epilepsy but not stroke, hypertension, or coronary heart disease.
	(Chien et al., 2010)	Hudson Health	DID and ITS	Patient health status	Children with chronic conditions were more likely to receive full immunization than children without chronic conditions.
	(Chien et al., 2012)	Hudson Health	DID and ITS	Patient health status	Changes in quality of care or utilization did not vary by number of comorbidities (data not shown).
	(Bardach et al., 2013)	PCIP	LA	Patient health status	The intervention group improved more than controls across most comorbidity categories.
	(Epstein et al., 2014)	HQID	DID	Patient health status	CABG rates fell similarly among high and low risk AMI patients in HQID and control hospitals.
	(Hamilton et al., 2014)	UK QOF+	LA	Patient health status	Patients with mental health conditions were much less likely to be offered alcohol screening as compared to patients with cardiovascular disease.
	(Casalino, Alexander, et al., 2007)	Multiple programs	QA	Patient health status	Most physicians believed that quality measures did not adequately adjust for patients' medical conditions and that quality measurement programs could cause physicians to avoid high-risk patients.
	(Damberg et al., 2010)	IHA	CSA	Share of Medicaid patients	Medicaid revenue was not related to composite measures of diabetes management, care processes, or intermediate diabetes and cardiovascular outcomes.
	(Chien, Wroblewski, et al., 2012)	IHA	CSA	Share of Medicaid patients	Increased Medicaid revenue was associated with poorer performance. For every 10% in Medicaid revenue, POs were 32% less likely to rank in the top 40% of performance.
	(Doran et al., 2006)	QOF	CSA	Share of income-deprived patients	Reported achievement was lower among practices with higher shares of patients living in income-deprived households, with a 0.05 standard deviation decrease in reported achievement for each standard deviation increase in household income deprivation.
	(Doran, Fullwood, Reeves, et al., 2008)	QOF	CSA	Share of income-deprived patients	Exception reporting rates were not related to practice share of income-deprived patients.
	(Gravelle et al., 2008)	QOF	CSA	Share of income-deprived patients	Population achievement was inversely related to share of income-deprived patients.
	(Chien, Wroblewski, et al., 2012)	IHA	CSA	Area-level SES	A PO's area-based SES was positively related to IHA performance.
	(Kontopantelis et al., 2013)	QOF	CSA	Area-level deprivation	Greater area-level deprivation was associated with somewhat worse performance for reported achievement, population achievement, and achievement of quality points.

(Gravelle et al., 2008)	QOF	CSA	Share of ethnic minorities	Reported achievement was lower in Scottish practices with larger shares of minorities.
(Doran et al., 2006)	QOF	CSA	Share of ethnic minorities	Reported achievement was not related to race/ethnicity in the first year of the QOF.
(Doran, Fullwood, Reeves, et al., 2008)	QOF	CSA	Share of ethnic minorities	Race/ethnicity was strongly associated with exception reporting for asthma and treatment provision but not related to overall exception reporting rates in Year 1.
(Gray et al., 2007).	QOF	CSA	Patient race/ethnicity	Black patients and South Asian patients had lower odds of achieving all three intermediate clinical outcomes (i.e., cholesterol, BP, diabetes). There was no consistent relationship between ethnicity and process measures of care.
(Gravelle et al., 2008)	QOF	CSA	Area-level diversity	Rates of exception reporting were higher among practices that drew from areas with higher proportions of ethnic minorities.
(Kontopantelis et al., 2013)	QOF	CSA	Rurality	Rural practices achieved lower percentages of quality points and demonstrated worse reported achievement but not population achievement.
(Gravelle et al., 2008)	QOF	CSA	Rurality	Rates of exception reporting in Scotland were lower among practices in rural areas.
(Doran, Fullwood, Reeves, et al., 2008).	QOF	CSA	Population density	Population density was related to modestly higher rates of exception reporting.
(Gravelle et al., 2008)	QOF	CSA	Share of older patients	Higher proportions of patients over 65 years of age were positively correlated to exception rates among Scottish practices.
(Doran, Fullwood, Reeves, et al., 2008)	QOF	CSA	Share of older and younger patients	The proportion of patients over 65 years of age and under 16 years of age was inversely related to reported exception rates.
(Doran et al., 2006).	QOF	CSA	Share of older patients	Overall reported achievement was lower among practices with higher proportions of patients over 65 years of age.
(Kontopantelis et al., 2013)	QOF	CSA	Share of older patients	The proportion of patients over 65 years of age was negatively associated with reported achievement but positively associated with population achievement and the percentage of available quality points achieved.
(Doran et al., 2006)	QOF	CSA	Share of female patients	Reported achievement was not significantly associated with patient gender.
(Kontopantelis et al., 2013)	QOF	CSA	Share of female patients	Practices with larger shares of female patients demonstrated better performance across all three outcomes (reported achievement, population achievement, and percentage of quality points attained).

Note. QA = qualitative analysis; DID = difference-in-differences analysis; LA = longitudinal analysis; ITS = interrupted time series analysis; DDD = difference-in-difference-in-differences; CSA = cross-sectional analysis; PO = physician organization; DSH = disproportionate share hospital; AMI = acute myocardial infarction; HF = heart failure; RA=risk-adjusted; BSRR = risk-standardized readmission ratios; HHI = Herfindahl-Hirschman Index; SSI = supplemental security income; THR/TKR = total hip replacement/total knee replacement; GABG = coronary artery bypass graft; HbA1c = hemoglobin A1c; BP = blood pressure.

Table 2
Articles Evaluating Whether Organizational Structure and Capabilities Modify P4P Performance

Effect	Study	Program	Design	Variable	Results
P4P Programs Targeting Hospitals					
Modifying effects	(Lindenauer et al., 2007)	HQID	DID	Hospital size	Hospital size appeared inversely related to composite scores in stratified analyses. Formal interaction tests were not presented.
	(Sutton et al., 2012)	AQ	DID, DDD	Hospital size	Mortality reductions were greatest for small hospitals.
	(Ryan & Blustein, 2011)	MassHealth	DID	Hospital size	The effect of the P4P did not vary by number of beds.
	(Lindenauer et al., 2007)	HQID	DID	Teaching hospital	Non-teaching hospitals improved more than teaching hospitals for overall performance (4.9% vs. 3.5%) and composite process scores for AMI (5.9% vs. 2.0%) and HF (6.6% vs. 2.6%) but not pneumonia (3.6% vs. 4.4%) compared to HQA peers.
	(Ryan & Blustein, 2011)	MassHealth	DID	Teaching hospital	The effect of the P4P did not vary by teaching status.
	Ryan, Burgess, Pesko, Borden, & Dimick, 2014)	HVBP	DID	Baseline performance	Changes in patient experience or clinical process did not vary by initial performance.
	(Sutton et al., 2012)	AQ	DID, DDD	Baseline performance	High but not low quality AQ hospitals (as rated by national regulator) reduced pneumonia mortality more than similarly rated, non-AQ hospitals.
	(Glickman et al., 2007)	HQID	DID	Baseline performance	Baseline score was inversely related to changes on HQID and non-HQID composites. As this trend occurred for all hospitals, P4P effects did not vary by baseline.
	(Nicholas, Dimick, & Iwashyna, 2011)	HQID	LA	Baseline performance	The effect of the HQID did not vary significantly by initial quality.
	(Lindenauer et al., 2007)	HQID	DID	Baseline performance	Low performers improved more than high performers in participants vs. controls for HF processes but not AMI, pneumonia, or overall process measures (data not shown).
	(Jha et al., 2012)	HQID	DID	Baseline performance	There were no differences in overall mortality at baseline or change in mortality over time among the overall sample or low performers at baseline (i.e., highest mortality quartile). Low HQID performers improved more than peers for pneumonia mortality.
	(Shih, Nicholas, Thumma, Birkmeyer, & Dimick, 2014)	HQID	DID	Baseline performance	Change in incentive structure did not decrease adverse events among either the overall sample or hospitals in the lowest quintile for mortality or serious complications.
	(Ryan, Blustein, & Casalino, 2012)	HQID	DID	Baseline performance	The effects of the change in incentive structure did not vary by initial performance quartiles for the three incentivized measures.
	(Werner et al., 2011)	HQID	LA	Financial strength	HQID hospitals in better financial shape (i.e., above-average margins) improved more than those in worse shape. This gap grew over time and reached 22% by 2008.
	(Jha et al., 2012)	HQID	DID	Financial strength	Mortality did not vary by hospital financial health (i.e., total margins per 1% change).

Effect	Study	Program	Design	Variable	Results
Direct effects	(Epstein et al., 2014)	HQID	DID	Financial strength	Medicare margins did not modify CABG access changes between participants and controls.
	(Sutton et al., 2012)	AQ	DID, DDD	Financial strength	Neither financial management rating nor designation as a "foundation trust" (i.e., NHS hospitals with greater financial and managerial freedom) modified mortality changes.
	(Vina et al., 2009)	HQID	QA	Organizational support	Organizational Support factors were generally similar between high and low performing hospitals. However, higher performer status was associated with greater nursing staff support for QI adherence and adequate human resources for QI.
	(Vina et al., 2009)	HQID	QA	Physician Leadership	Recruitment of physician QI champions was significantly higher in top vs. bottom performance groups. High performers were not more likely to have a chief medical officer whose responsibilities included QI.
	(Vina et al., 2009)	HQID	QA	Organizational culture	Top performers reported having an organization culture that supported care coordination across departments, an emphasis on identifying system errors rather than blaming individuals, and a willingness to try new QI projects.
	(Vina et al., 2009)	HQID Phase	QA	QI participation	Top performers were significantly more likely than bottom performers to use: clinical pathways for AMI, HF, pneumonia, and THR/TKR; multidisciplinary teams for AMI and HF; and order sets for THR/TKR.
	(Joynt & Jha, 2013)	HRRP	CSA	Teaching hospital	Teaching hospitals were predicted to be somewhat more likely to receive penalties.
	(MedPAC, 2014)	HRRP	CSA	Teaching hospital	Major teaching hospitals fared worse than non-teaching hospitals for average penalties as a share of operating payments (0.45% vs. 0.31%) or total inpatient payments (0.28% vs. 0.25%).
	(Vina et al., 2009)	HQID	QA	Information technology	Top performers were more likely to use computerized physician order entry systems. Collection and reporting of hospital- and physician-specific reports was not related to performance.
	(McHugh et al., 2013)	HRRP	CSA	Nurse-to-patient ratios	Hospitals with higher nurse staffing had lower odds of penalties compared to similar hospitals with lower nurse staffing levels. Each additional nurse hour per adjusted patient was associated with 10% lower odds of being penalized.
	(Joynt & Jha, 2013)	HRRP	CSA	Hospital size	Large (>400 beds) and medium (200–399 beds) size hospitals were predicted to be about two times more likely to be penalized than small hospitals (<200 beds).
	(MedPAC, 2014)	HRRP	CSA	Hospital size	Large hospitals (>300 beds) had higher penalties as a share of operating payments (0.33% vs. 0.28%) or total inpatient payments (0.28% vs. 0.24%) compared to small hospitals (<100 beds).
	(MedPAC, 2014)	HRRP	CSA	Occupancy rate	High occupancy hospitals (>75%) had higher penalties than low occupancy hospitals (<50%) as a share of operating payments (0.43% vs. 0.28%) or total inpatient payments (0.31% vs. 0.22%).
P4P Programs Targeting Physician Practices and Organizations					
Modifying effects	(Doran, Fullwood, Kontopantelis, et al., 2008)	QOF	LA	Practice size	Larger practices improved more slowly over time as compared to smaller practices.

Effect	Study	Program	Design	Variable	Results
	(Vamos et al., 2011)	QOF	LA	Practice size	The effect of the QOF varied by practice size for some intermediate outcome measures in some years and did not vary for process measures.
	(Li et al., 2014)	PCR	DID	Practice size	Larger practice size was weakly related to improved provision of flu shots, toddler immunization, mammogram, colorectal cancer screening, but not Pap smears.
	(An et al., 2008)	Smoking cessation	LA	QI participation	Only history of QI involvement predicted a differential response to the P4P: Clinics with a history of being engaged and clinics with a history of being less engaged improved more under P4P than did peer clinics.
	(Chien et al., 2010)	Hudson Health	DID, ITS	Practice type	Children were more likely to be immunized if seen by a private practice than a FQHC.
	(Chien et al., 2012)	Hudson Health	DID, ITS	Practice type	Changes in quality or utilization did not vary by practice type (data not shown).
	(An et al., 2008)	Smoking cessation	LA	Practice type	Clinic specialty type (i.e., family practice, internal medicine, obstetrics-gynecology, multispecialty) was not significantly related to referral rates.
	(Coleman et al., 2007)	ACCESS	Pre-post	Physician specialty	The effect of the intervention did not vary systematically by physician specialty.
	(Kontopantelis et al., 2013)	QOF	LA	Information technology	Type of computing system used was the strongest predictor of reported achievement, population achievement, and point achievement among all practice and population characteristics. Differences across systems were stronger for intermediate outcomes than for measurement or treatment indicators. Payment differences by computing system were small.
	(Damberg et al., 2009)	IHA	QA	Baseline performance	Health plans felt the lowest performing POs improved most but that a gap persisted. Health plans suggested targeting low performers or using improvement-based designs.
	(Doran, Fullwood, Kontopantelis, et al., 2008)	QOF	LA	Baseline performance	Lower previous practice performance predicted greater subsequent quality but also reported exclusion rates, which, in turn, were associated with better performance.
	(Chen et al., 2010)	Blue Cross Hawaii	DID	Baseline performance	Physicians in the lowest quartile improved most between years 1 and 2 among P4P and comparison groups. Only those in P4P continued to improve in subsequent years.
	(Mullen et al., 2010)	QIP & IHA	DID, LTT	Baseline performance	The QIP and IHA did not systemically vary by baseline performance. Lowest quintile improved more on breast cancer screening but less on cervical cancer screening.
	(Coleman et al., 2007).	ACCESS	Pre-post	Baseline performance	Lower process but not outcome baseline performance predicted greater improvements.
	(Li et al., 2014)	PCR	DID	Baseline performance	Lower physician baseline performance predicted greater provision of flu shots, mammograms, and colorectal cancer screenings but not immunization or Pap smears.
	(Li et al., 2014)	PCR	DID	Physician age	Younger physicians responded more vigorously to P4P for pap smears, mammograms, and colorectal cancer screening but not for provision of senior flu shots or toddler immunizations.

Effect	Study	Program	Design	Variable	Results
Direct effects	(Doran, Fullwood, Kontopantelis, et al., 2008)	QOF	LA	Share of female physicians	Higher proportions of female practitioners were associated with greater improvements in population achievement between years 1 and 2 (but not 2 and 3).
	(Doran, Fullwood, Reeves, et al., 2008)	QOF	LA	Share of female physicians	Gender was not associated with overall exception rate reporting.
	(Doran, Fullwood, Kontopantelis, et al., 2008)	QOF	LA	Physician-to-patient ratios	Physician-to-patient ratios were negatively correlated with reported achievement improvements between Years 1 and 2 but not between Years 2 and 3.
	(Ryan & Doran, 2012)	QOF	LA	Physician-to-patient ratios	Higher physician-to-patient ratios strengthened the relationship between process and outcomes for diabetes and CAD but weakened the relationship for hypertension.
	(An et al., 2008)	Smoking cessation	LA	Physician number	Absolute number of physicians was not related to referral rate improvements.
	(Coleman et al., 2007)	ACCESS	Pre-post	Training location	The effect of the intervention did not vary systematically by country of physician training.
	(Doran, Fullwood, Reeves, et al., 2008)	QOF	LA	Training location	Share of UK-trained practitioners was not associated with overall rates of exception reporting.
	(Mehrotra et al., 2007)	Multiple programs	QA	PO size	PO size was positively related to using QI initiatives that targeted a HEDIS measure.
	(Chien, Wroblewski, et al., 2012)	IHA	CSA	PO size	Larger POs were more likely to rank in the top 40% of IHA composite score.
	(Damberg et al., 2010)	IHA	CSA	PO size	PO size was positively associated with processes-of-care composite performance.
	(Damberg et al., 2009)	IHA	QA	PO size	Larger POs reported greater ability to monitor quality as compared to smaller POs.
	(Doran et al., 2006)	QOF	CSA	Practice size	Larger practices demonstrated lower reported achievement in Year 1 of the QOF.
	(Doran, Fullwood, Reeves, et al., 2008)	QOF	CSA	Practice size	Larger practice size was related to modestly higher rates of exception reporting in Scotland, with larger practices more likely to be in the top 1% of exception reporting.
	(Kontopantelis et al., 2013)	QOF	CSA	Practice size	Larger practices demonstrated worse reported achievement and population achievement but achieved a higher percentage of available quality points.
	(Casalino, Alexander, et al., 2007)	Multiple programs	QA	Practice sizes	Very small practices (1–2 physicians) were less likely to support financial incentives and were more likely to support public reporting on groups but not physicians.
	(Damberg et al., 2009)	IHA	QA	Organizational culture	Low performing POs reported less organizational support for addressing quality issues. High and low performing POs reported similarly strong views on organizational culture of quality.
	(Damberg et al., 2010)	IHA	CSA	QI participation	Participation in external QI initiatives was related to higher composite process scores.

Effect	Study	Program	Design	Variable	Results
	(Mehrotra et al., 2007)	Multiple programs	QA	Practice type	Affiliation with a physician network and employment of a majority of its physicians were all strongly related to physician group use of QI initiatives.
	(Chien, Wroblewski, et al., 2012)	IHA	CSA	Practice type	Medical groups were more likely than IPAs to rank in the top 40% of performance.
	(Damberg et al., 2010)	IHA	CSA	Practice type	Medical group organizations outperformed IPAs on two of three composite measures.
	(Mehrotra et al., 2007)	Multiple programs	QA	Information technology	Use of EMR by majority of physicians was not significantly related to QI initiatives.
	(Damberg et al., 2009)	IHA	QA	Information technology	High performing POs placed greater importance on IT capability than did low performing POs. Both high and low performing POs reported difficulty monitoring quality performance at the PO and physician levels. Most POs reported aggregated utilization and HEDIS or IHA measures to physicians. Few POs were capable of reporting physician-specific performance results.
	(Damberg et al., 2010)	IHA	CSA	Information technology	EMR capability was not significantly related to clinical performance.
	(Kontopantelis, Buchan, Reeves, Checkland, & Doran, 2013)	QOF	CSA	Practices' mean age of physician	Mean physician age was related to modestly lower reported achievement, population achievement, and percentage of available quality points obtained.
	(Gravelle et al., 2008)	QOF	CSA	Share of younger physicians	Practices with higher proportions of younger physicians demonstrated significantly greater population achievement in Scotland.
	(Doran, Fullwood, Reeves, et al., 2008)	QOF	CSA	Share of older physicians	Overall rate of exception reporting was not associated with share of physicians over 55.
	(Casalino, Alexander, et al., 2007)	Multiple programs	QA	Physician age	Physician age was not related to physician support for financial incentives or public reporting.
	(Gravelle et al., 2008)	QOF	CSA	Share of female physicians	Practices with higher proportions of female physicians demonstrated significantly better population achievement in Scotland.
	(Kontopantelis et al., 2013)	QOF	CSA	Share of female physicians	Higher proportions of female physicians were related to slightly better reported achievement, population achievement, and percentage of available quality points attained.
	(Casalino, Alexander, et al., 2007)	Multiple programs	QA	Female physicians	Female physicians were less likely to support financial incentives and public reporting for individual physicians but not medical groups.
	(Doran et al., 2006)	QOF	CSA	Financial arrangements	Contracting with Personal Medical Services was related to lower overall achievement.
	(Kontopantelis et al., 2013)	QOF	CSA	Financial arrangements	Performance was positively related to Personal Medical Services contracting and negatively related to contracting with Alternative Physician or Primary Care Trust Medical Services.
	(Doran, Fullwood, Reeves, et al., 2008)	QOF	CSA	Financial arrangements	Contracting with Personal Medical Services was not associated with exception rates.
	(Doran, Fullwood, Reeves, et al., 2008)	QOF	CSA	Financial arrangements	Holding a budget predicted higher rates of exception reporting in Scotland.

Effect	Study	Program	Design	Variable	Results
	(Doran et al., 2006)	QOF	CSA	Physician-to-patient ratios	Physician-to-patient ratios were positively correlated with reported achievement.
	(Doran, Fullwood, Reeves, et al., 2008)	QOF	CSA	Physician-to-patient ratios	Physician-to-patient ratios were not significantly related to exception rate reporting.
	(Doran et al., 2006)	QOF	CSA	Training location	Share of UK-trained practitioners was positively related to reported achievement in Year 1.
	(Ryan & Doran, 2012)	QOF	CSA	Training location	Share of UK-trained practitioners was not associated with improvements over time.
	(Kontopantelis et al., 2013)	QOF	CSA	Training location	Share of UK-trained practitioners was marginally and inconsistently related to performance.

Note. DID = difference-in-differences analysis; DDD = difference-in-difference-in-differences; CSA = cross-sectional analysis; QA = qualitative analysis; LA = longitudinal analysis; ITS = interrupted time series analysis; LTT = linear time trends; QI = quality improvement; PO = physician organization; HEDIS = Healthcare Effectiveness Data and Information Set; AMI = acute myocardial infarction; HF = heart failure; THR/TKR = total hip replacement/total knee replacement; IPA = independent practice associations; FQHC = federally qualified health center; EMR = electronic medical records; GABG = coronary artery bypass graft; CAD = coronary artery disease.

Table 3
Articles Evaluating Whether P4P Program Characteristics Affect P4P Performance

Effect	Study	Program	Design	Variable	Results
P4P Programs Targeting Hospitals					
Modifying effects	(Jha et al., 2012)	HQID	DID	Bonus size	Mortality improvements did not vary by the hospital's proportion of Medicare patients.
	(Epstein et al., 2014)	HQID	DID	Bonus size	Change in CABG rates among AMI patients did not vary by hospital share of Medicare patients.
	(Nicholas, Dimick, & Iwashyna, 2011)	HQID	LA	Bonus size	The effect of the HQID did not vary by hospital volume.
	(Werner et al., 2011)	HQID	LA	Bonus size	Improvements were positively related to share of Medicare payments linked to HQID conditions.
	(Ryan & Blustein, 2011)	MassHealth	DID	Bonus size	The effect of the P4P did not vary by Medicaid share or Medicare share for pneumonia or SIP. (Payouts were proportional to number of eligible Medicaid beneficiaries.)
Direct effects	(Nicholas, Dimick, & Iwashyna, 2011)	HQID	LA	Marginal costs	Participants with lower marginal cost ratios for substitution (calculated from relative numbers of eligible patients) were not more likely to substitute toward easier tasks across most measures.
	(Nicholas, Dimick, & Iwashyna, 2011)	HQID	LA	Task difficulty	There was no consistent evidence that hospitals preferentially improved on "easy" vs. "difficult" tasks. Hospitals strategically shifted resources to improve AMI scores; the HQID was associated with a 1% gain in compliance for easy vs. hard tasks. Improvement did not vary by difficulty for HF or pneumonia.
	(Bhattacharyya et al., 2009)	HQID	CSA	Bonus size	Surgical volume was positively related to composite THR/TKR quality and hospital performance tier. HQID quality was not related to mortality, iatrogenic complications, or UTI rates.
P4P Programs Targeting Physician Practices and Organizations					
Modifying effects	(Chung et al., 2010)	PAMC	LA	Bonus frequency	The effect of the randomized P4P did not vary by payment frequency. Physicians were randomized to either 4 quarterly bonuses or a single year-end bonus.
	(Petersen et al., 2013)	VA	LA	Bonus target	Individual-level but not clinic-level or combined incentives led to greater blood pressure control or appropriate response to uncontrolled blood pressure. None of the incentive schemes was related to greater use of guideline-recommended medications or the development of hypotension. Performance gains were not sustained after the washout.
	Mullen, Frank, & Rosenthal, 2010)	QIP & IHA	DID, LTT	Bonus size	The QIP did not vary by potential QIP bonus size for either cervical cancer screening or breast cancer screening. Total enrollment marginally increased QIP responses for cervical screening.
	(Mullen et al., 2010).	QIP & IHA	DID, LTT	Bonus size	Groups in higher baseline strata did not consistently improve more than those in lower strata despite incentives that awarded larger payments for improvements in progressively higher strata.
	(Kantarevic & Kralj, 2013)	PCR	DID	Bonus size	The effect of joining FHO (vs. FHG) practices was positively modified by expected income gain and share of diabetes patients on the intensive margin (i.e., physician's share of DMI patients) and negatively modified on the extensive margin (i.e., physician participation in the DMI).
	(Mullen et al., 2010)	QIP & IHA	DID, LTT	Bonus likelihood	P4P led to cross-substitution between cervical cancer screening and breast cancer screening for some baseline quantiles; groups with low cervical cancer screening and breast cancer screening just below targets increased breast cancer screening and decreased cervical cancer screening.
	(Rosenthal et al., 2005)	QIP & IHA	DID	Bonus likelihood	Practices with the most to gain (below but within 10% of the target) improved less than lower performers for cervical cancer screening, mammography, and HbA1c testing. Lowest baseline

Effect	Study	Program	Design	Variable	Results
					performers improved most but captured 5% of bonuses, while highest performers improved least but captured 75% of bonuses.
	(Dowd et al., 2013)	GPR	LA	Bonus likelihood	Practices below the absolute target threshold were equally likely to improve as those above the target for rates of prescribing generation medication. Practices with calculated equilibrium values up to 22% below the target continued to increase their rates of prescribing generation medication.

Note. DID = difference-in-differences analysis; LA = linear time trends; CSA = cross-sectional analysis; FHO = Family Health Organizations (blended capitation model); FHG = Family Health Groups (enhanced FFS model); CABG = coronary artery bypass graft; AMI = acute myocardial infarction; THR/TKR = total hip replacement/total knee replacement; SIP = surgical infection prevention; HbA1c = hemoglobin A1c.