This report is part of the RAND Corporation research report series. RAND reports present research findings and objective analysis that address the challenges facing the public and private sectors. All RAND reports undergo rigorous peer review to ensure high standards for research quality and objectivity.

RESEARCH REPORT

# Measuring Success in Health Care Value-Based Purchasing Programs

## Findings from an Environmental Scan, Literature Review, and Expert Panel Discussions

*Cheryl L. Damberg • Melony E. Sorbero • Susan L. Lovejoy*

*Grant Martsolf • Laura Raaen • Daniel Mandel*

Sponsored by the Office of the Assistant Secretary for Planning and Evaluation

The research described in this report was sponsored by the Office of the Assistant Secretary for Planning and Evaluation in the U.S. Department of Health and Human Services, and was conducted in RAND Health, a division of the RAND Corporation.



RAND OFFICES

SANTA MONICA, CA • WASHINGTON, DC

PITTSBURGH, PA • NEW ORLEANS, LA • JACKSON, MS • BOSTON, MA

CAMBRIDGE, UK • BRUSSELS, BE

# Preface

The U.S. Department of Health and Human Services (HHS) is advancing the implementation of value-based purchasing (VBP) across an array of health care settings in the Medicare program in response to requirements in the 2010 Patient Protection and Affordable Care Act. VBP refers to a broad set of performance-based payment strategies that link financial incentives to providers' performance on a set of defined measures in an effort to achieve better value by driving improvements in quality and slowing the growth in health care spending. Policymakers are grappling with many policy decisions about how best to design and implement VBP programs so that they are successful in achieving stated goals.

To inform future policymaking by HHS regarding the implementation of VBP in the Medicare program, the Office of the Assistant Secretary for Planning and Evaluation in HHS asked RAND to review what has been learned over the past decade with performance-based payment models. Three types of VBP models were the focus of the review: (1) pay-for-performance programs, (2) accountable care organizations (ACOs), and (3) bundled payment programs.

This report summarizes the current state of knowledge based on a review of the published literature, a review of publicly available documentation from VBP programs, and discussions with an expert panel composed of VBP program sponsors, health care providers and health systems, and academic researchers with VBP evaluation expertise.

The contents of this report will be of interest to public and private payers of health care who sponsor VBP programs, health care providers, policymakers, and health researchers who work to build the evidence base.

ASPE sponsored this work under contract No. 12-233-SOL-00418. The work was conducted in RAND Health, a division of the RAND Corporation. A profile of RAND Health, abstracts of its publications, and ordering information can be found at www.rand.org/health.

# Contents

# Figures

# Tables

# Executive Summary

Value-based purchasing (VBP) refers to a broad set of performance-based payment strategies that link financial incentives to providers' performance on a set of defined measures. Both public and private payers are using VBP strategies in an effort to drive improvements in quality and to slow the growth in health care spending. Nearly ten years ago, the Department of Health and Human Services (HHS) and the Centers for Medicare and Medicaid Services (CMS) began testing VBP models with their hospital pay-for-performance (P4P) demonstrations, known as the Premier Hospital Quality Incentive Demonstration (HQID) and the Physician Group Practice (PGP) Demonstration, which provided financial incentives to physician groups that performed well on quality and cost metrics. The use of financial incentives as a strategy to drive improvements in care dates back even further among private payers[2] and Medicaid programs, with limited experimentation occurring in the early 1990s; more widespread use of P4P began to pick up steam in the late 1990s and early 2000s.

Although the published evidence from P4P programs implemented by private-sector payers between 2000 and 2010 showed mostly modest results in improving performance,[3–10] public and private payers have continued to experiment with the use of financial incentives as a policy lever to drive improvements in care. Many of the early P4P program designs have evolved over time to include a larger and broader set of measures, including resource use and cost metrics, in an effort to reward providers for delivering value,[*] and many programs are deploying a wider range of incentives. Additionally, other VBP models have since emerged and are currently being tested, including accountable care organizations (ACOs) and bundled payment programs that include both quality and cost design features. VBP models are relatively new to the health system, and they represent a work in progress in terms of understanding how best to design these programs to achieve desired goals, the optimal conditions that support successful implementation, and provider response to the incentives.

## Policy Context and Study Purpose

The Medicare program has gradually been moving toward implementing VBP across various care settings, starting with pay-for-reporting programs (e.g., the Hospital Inpatient Quality Reporting program and the Physician Quality Reporting Initiative) and P4P demonstrations to

[*] Value is defined as the outcomes (outputs) achieved divided by the cost or resources used (inputs) to generate those outcomes.

gain experience. The 2010 Patient Protection and Affordable Care Act[11] significantly expands VBP by requiring the Medicare program to implement, develop plans for, and test in the context of demonstrations the use of VBP across a broad set of providers and settings of care.

As HHS actively considers the federal government's near- and long-term strategy for how to design and implement VBP programs within the Medicare program, the department is seeking to apply the best available evidence to guide policymaking. Because of the substantial investments that HHS is making regarding VBP, it is an opportune moment to reflect on what has been learned from the past decade of experimentation that could guide current and future federal efforts. It is also a good time to consider the type of monitoring and systematic evaluation work that is needed to generate the information that policymakers require to fine-tune VBP program designs and to understand the impact these programs are having related to stated goals.

In 2012, the Office of the Assistant Secretary for Planning and Evaluation (ASPE) in HHS asked RAND to review what has been learned about VBP over the past decade that might help inform policymaking. The goal of the review was to understand whether VBP programs have been successful, what the elements of successful programs are, and the gaps in the knowledge base that need to be addressed to improve the design and functioning of VBP programs moving forward. This report summarizes the findings from RAND's review. We direct readers to the companion document to this summary report, *Measuring Success in Health Care Value-Based Purchasing Programs: Summary and Recommendations.*

## Conceptual Framework for Assessing the Effects of Value-Based Purchasing Programs

To help us consider the research questions that ASPE asked RAND to address, we developed a conceptual framework for VBP. The model is adapted from a conceptual model by Dudley et al.[12] and includes three core elements that interplay and affect the response to VBP:

- **Program design features** (i.e., measures, incentive structure, target of incentive, and quality improvement support/resources)
- **Characteristics of the providers and the settings in which they practice** that may predispose them to a response
- **External factors** (e.g., other payment policies, other quality initiatives, regulatory changes) that can enable or hinder provider response to the incentive.

The conceptual framework offers a foundation for considering the design features of the incentive program, as well as other mediating factors that influence whether and how providers may respond to the incentives and whether programs are successful in reaching stated goals. Largely, VBP programs are natural experiments, and the associated research is observational in nature. Dudley (2005) underscores that, as a result, it is critical that evaluators select theory-driven hypotheses about how incentives affect behavior to identify potential confounding factors

that could explain observed effects.[13] Policymakers and researchers could use this framework to develop theory-driven hypotheses.

## Study Approach

We defined VBP programs as private or public programs that link financial reimbursement to performance on measures of quality (i.e., structure, process, outcomes, access, and patient experience) and cost or resource use. We focused our review on three types of VBP models: (1) P4P, which includes both "pay for quality" and "pay for quality and resource use, efficiency, or costs"; (2) shared savings models that typically, but not exclusively, are being deployed in the context of ACOs; and (3) bundled payments for episodes of care (only when paired with holding providers accountable for performance on quality measures). We excluded from review pay-for-reporting and demand-side programs (e.g., tiered networks and consumer incentives).

We define each of the three broad types of VBP models as follows:

- **Pay-for-performance** refers to a payment arrangement in which providers are rewarded (bonuses) or penalized (reductions in payments) based on meeting pre-established targets or benchmarks for measures of quality and/or efficiency.
- **Accountable care organization** refers to a health care organization composed of doctors, hospitals, and other health care providers who voluntarily come together to provide coordinated care and agree to be held accountable for the overall costs and quality of care for an assigned population of patients. The payment model ties provider reimbursements to performance on quality measures and reductions in the total cost of care. Under an ACO arrangement, providers in the ACO agree to take financial risk and are eligible for a share of the savings achieved through improved care delivery provided they achieve quality and spending targets negotiated between the ACO and the payer.
- **Bundled payments**[*] are a method in which payments to health care providers are based on the expected costs for a clinically defined episode or bundle of related health care services. The payment arrangement includes financial and quality performance accountability for the episode of care.

ASPE identified 16 research questions that were the focus of this review, organized by three broad areas of inquiry: (1) measuring the performance of VBP programs; (2) the results of performance in VBP programs; and (3) improving the performance of VBP programs. We used three approaches to gather information to address the questions:

- **Environmental scan of existing value-based purchasing programs:** We reviewed information that was publicly available for 129 VBP programs (91 P4P programs, 27

[*] Other common terms used for bundled payment arrangements are *episode-based payment*, *episode payment*, *episode-of-care payment*, *case rate*, *evidence-based case rate*, *global bundled payment*, and *global payment*.

ACOs, and 11 bundled payment programs) sponsored by private health plans, regional collaboratives, Medicaid agencies or states, and the federal government. The VBP programs we reviewed do not represent the universe of all VBP programs in current operation in the United States, and the documentation for some programs we reviewed was not complete given the propriety nature of the information.

- **Review of the published evaluation literature on value-based purchasing:** We examined the peer-reviewed published literature for studies that evaluated the impact of P4P, ACO, or VBP-type bundled payment programs.
- **Input from a technical expert panel:** We convened a technical expert panel (TEP), composed of VBP program sponsors, providers from health systems who have been the target of VBP programs, and health services researchers with expertise in examining the effects of VBP programs, to help address many of the study questions where the literature was void of information. We provided the TEP with the findings from the environmental scan and the literature review as background information for the panel's discussions.[14]

## Summary of Findings

We summarize the findings from the environmental scan of existing programs, the literature review, and our discussions with the TEP in an integrated manner. The findings are organized by the topic areas we were asked to address in the scope of work for this project. We direct readers of this report to its companion report, *Measuring Success in Health Care Value-Based Purchasing Programs: Summary and Recommendations*, which provides a set of recommendations that emerged from our review and TEP discussions.

### *Goals of Value-Based Purchasing Programs*

Based on our review of VBP programs in operation, VBP program sponsors tend to identify multiple high-level goals that focus on improving clinical quality (75 percent of the programs we reviewed) and cost/affordability (53 percent of the programs we reviewed). Less commonly reported were goals related to improving patient outcomes (34 percent) and patient experience (17 percent). There was some variation in goals among VBP program type, with goals focused on coordination of care and patient experience more prevalent in ACO and bundled payment programs as compared with P4P programs.

In most cases, the goals specified by VBP program sponsors were not quantified or measurable (e.g., "breakthrough improvement in quality" or "bend the cost curve"). In a handful of cases (five of the 129 programs we reviewed), we found quantified goals related to desired cost savings (e.g., "keep 2010 health care premium costs flat" and "reduce the annual increase in cost of care by two percentage points"). Our inability to find the specific performance goals for many of the VBP programs, particularly programs sponsored by private-sector payers, is likely a function of the proprietary nature of this information. Performance measures and thresholds are embedded within the contracts negotiated between providers (i.e., physicians, physician organizations, hospitals) and payers.

The absence of quantifiable goals for many programs makes it difficult to determine whether programs have been successful in meeting their goals; instead, evaluators and program sponsors typically examine whether performance on the incentivized measures improved over time. Given this difficulty, the TEP recommended that individual VBP program sponsors establish well-defined, measurable intermediate goals (i.e., program performance targets) derived from external benchmarks and use these to assess success.

Our discussions with the TEP also revealed support for VBP programs having broad goals, and panelists commented that beyond driving improvements in quality and costs, the larger goal of VBP is to transform the way care is delivered to enhance performance. TEP members outlined the following additional goals that they believed would be important to establish and potentially measure to assess VBP program success:

- **Stimulate organizational nimbleness to rapidly learn and improve in order to achieve a new performance target.** TEP members indicated that a key goal of VBP is improving the functional capacity of providers to learn and improve. Therefore, it is important to understand whether there is capacity in health systems and provider organizations to improve quality against a moving target, and whether performance levels can be maintained once targets are achieved. TEP members commented that VBP programs should affect providers' willingness to change, their measurement capacity to identify problems, and their ability to respond to correct quality defects.
- **Promote innovation.** The panelists commented that part of the value of VBP is the innovation that occurs to fix the fundamental problems leading to poor quality and outcomes within provider organizations and, ideally, across providers in response to the incentive scheme. Examples they cited were the creation of more integrated data systems to improve communication between providers, the development of care management protocols that span care settings to improve transitions in care between the hospitals and ambulatory settings, investments in registries that allow physicians to track and better manage high risk populations, the development and use of risk assessment tools, and provision of clinical decision support. There was interest among the TEP panelists in capturing whether and how VBP initiatives are stimulating innovation.

Although the TEP identified a desire to understand whether VBP is successful in helping to make providers "more nimble" and to "improve their functional capacity for learning and improvement," it remains unclear at this stage what providers would need to demonstrate to prove that these aspirational goals had been met. To the extent that these are desired characteristics that VBP program sponsors want to encourage, work is required to define what is meant by these concepts so that VBP sponsors could determine whether this evolution has occurred.

The TEP also discussed whether success should be defined by levels (i.e., absolute performance achieved) or by the counterfactual (i.e., the extent of improvement in performance compared with what it would have been absent the VBP program). A VBP program sponsor may consider a program successful if a certain level of performance is met, whereas researchers would consider a program successful if greater improvements in performance occurred for those

providers exposed to VBP as compared with those who were not (i.e., the comparison group). The latter perspective is important because quality may be improving broadly over time as a function of a variety of factors, such as quality improvement interventions and infrastructure improvements distinct from actions undertaken in response to the VBP program, so providers may reach the stated goals in the absence of a VBP program. This discussion highlighted important differences in what program sponsors, policymakers, and researchers are interested in evaluating and what defines success.

The VBP program sponsors on the TEP felt that study designs need to be adapted to fit with the needs for making policy change, such as more rapid but less rigorous initial evaluation cycles to guide decisions about fine-tuning program design. They cited the initial Premier HQID design, which was changed based on less rigorous evidence; the changes were needed to restructure the incentives to achieve more engagement from poorly performing hospitals.

### *Measures Included in Value-Based Purchasing Programs*

Our review of public documents from VBP programs revealed there is a relatively narrow set of measures included in VBP programs that are used as the basis for differential payments. The measures vary somewhat by the health care settings in which they are being deployed as well as by the type of VBP model.* Historically, P4P programs have focused on quality performance, while the newer VBP models (ACOs and bundled payments) incentivize providers for both cost and quality; however, P4P programs have been evolving over time to include more cost and use measures. P4P programs typically include measures of clinical process and intermediate outcomes (e.g., Healthcare Effectiveness Data and Information Set [HEDIS] or Joint Commission measures), patient safety measures (e.g., surgical infection prevention), utilization (generic prescribing, emergency department use, length of stay, ambulatory care sensitive hospital admissions), patient experience (i.e., Consumer Assessment of Healthcare Providers and Systems survey, Hospital Consumer Assessment of Healthcare Providers and Systems survey), and, to a more limited degree, outcomes (e.g., readmissions, mortality, complications, total cost of care or cost per episode) and structural elements (e.g., HIT adoption or meaningful use of HIT requirements for CMS incentive payments, National Committee for Quality Assurance certification or patient-centered medical home certification, staffing, inspections). Clinical measures in the ambulatory setting focus heavily on preventive care and management of heart disease and diabetes, while in the hospital setting, the focus has been on heart attack, congestive heart failure (CHF), pneumonia, and surgical infection prevention.

* For example, for fiscal year 2014, CMS has 59 clinical and patient experience measures in its Hospital Inpatient Quality Reporting program and 18 clinical measures for nursing homes under its Nursing Home Quality Initiative.

The three ACO program models being tested by CMS use 33 measures, which include HEDIS clinical processes and intermediate outcomes; Consumer Assessment of Healthcare Providers and Systems survey questions on patient experience; all-cause hospital readmission; ambulatory sensitive care hospital admissions; patient safety; and electronic health record (EHR) functionality. Private-sector ACOs are using a similar set of measures, and again the clinical focus has been on three highly prevalent chronic conditions (i.e., heart disease, diabetes, and hypertension), cancer screening, and immunizations. The measures included in bundled payment programs tend to vary by the condition or procedure included in the episode as well as the setting(s) in which care is delivered. Cost measures are most commonly used. In the hospital setting, where most bundled payment programs occur, measures include clinical process, patient safety, readmissions, mortality, length of stay, and total cost of care. Some programs avoid tying physician compensation to outcome measures, so that physicians will not hesitate to treat patients who are more complicated. Little public information is available regarding the measures that are being used in ambulatory care bundled payment programs. Some of the VBP programs we reviewed are signaling that they intend to move to patient-reported outcomes in the next few years, but they are struggling to find market-ready measures that can be readily applied.

The discussions with the TEP highlighted problems with the narrow set of measures typically being used in VBP programs. The TEP estimated that only a small fraction (less than 20 percent) of all care that is delivered by providers is addressed by performance measures in VBP programs. An exception is "total cost of care" contracts (which as of late 2013 apply to only a small number of organizations) that hold providers accountable for the cost of all or most care delivered but which only measure quality performance for a fraction of all care delivered by providers. It was the panelists' opinion that the current, narrow set of measures tends to encourage providers to narrowly focus improvement efforts on the things that are measured (teaching to test) rather than wholesale improvement. The TEP also expressed concern that it is hard to demonstrate that VBP programs lead to performance improvements when the incentivized measures are the same set of measures that have been used for nearly a decade (i.e., Joint Commission measures, HEDIS); many of these measures have less room for improvement and, in some cases, have topped out. Panelists commented that shifting measurement focus to areas where performance is lagging[15] would better address the question of whether VBP can improve the delivery of care in areas not previously the focus of reporting and incentives. With respect to what is measured, the TEP questioned whether VBP programs are addressing areas with the greatest impact on health. While medical care can influence health outcomes, the TEP observed that lifestyle behaviors (diet, exercise, smoking, etc.) contribute roughly 50 percent to determining health outcomes.

Another measurement challenge the TEP flagged was the inability to assess value because of the lack of an agreed-upon definition of value and that providers' lack of cost accounting systems that enable them to know the true cost of delivering care. Many organizations have struggled with how best to measure and convey value to providers and consumers, highlighting

the need for measure development in this area. Although they did not offer a definition of value, the TEP members thought that a first step would be to achieve consensus on an overarching view of what value means; then VBP sponsors could develop value measures in the context of their own programs.

Many members of the TEP thought that a broad and more comprehensive set of measures in VBP programs would create incentives for providers to perform well across the board, rather than focus narrowly on a small number of areas, which promotes "teaching to the test"—that is, focusing only on improving areas that are measured and incentivized by the VBP program and ignoring clinically important areas that are not. However, neither the literature nor the TEP addressed how many measures are reasonable or practical to implement or when the data collection burden on providers becomes excessive. Expanding the set of measures included in VBP programs to more comprehensively assess care delivered and to include infrequently captured measure domains will require the development of new measures and new types of measures. Developing new measures is a time- and resource-intensive activity. Measurement concepts must be defined, specifications developed, data collection processes piloted, and data validated, among other steps. Recognizing this, the TEP recommended that it would be important to develop a framework to guide future directions about what to measure and, in turn, what measures need to be developed. They stated that the framework should address the multiple levels at which behavioral change needs to occur and where interventions should be directed (i.e., health system, institution, and individual provider).

The TEP identified several areas, discussed below, that should be the focus of future measure expansion work in the context of VBP.

### Measuring Patient Outcomes and Functional Status

The TEP members agreed that the ultimate objective of VBP is to hold providers accountable for and financially incentivize provider performance primarily based on measures of health outcomes. CMS expressed that is moving toward increased accountability for outcomes in its hospital and physician VBP programs, and is seeking to find a balance of structure, process, and outcome measures in its programs. An example of this transition to outcomes is illustrated in the hospital VBP program. In the first year of hospital VBP, 70 percent of the measures were process measures, whereas in the second year the percentage drops to 30 percent, as currently outlined in CMS's proposed Notice of Rule Making.[16, 17] Questions remain about the pace at which CMS should push toward outcomes measurement, the types of outcomes to use, and the consequences of those actions.

There was sentiment among the TEP members that functional status/health status is an important, feasible measure and that inclusion of these types of measures would shift VBP programs in the direction of incentivizing performance on outcomes. TEP members pointed to several health care settings and providers that are already measuring functional status on a regular basis: Medicare ACO programs are paid for reporting patient-reported functional

limitations, and CMS collects health status information in nursing homes and home health agencies. The Dartmouth Institute is measuring quality-adjusted life years and has built functional status, which is considered a vital sign, into a provider order for life-sustaining care for patients who are at or near the end of life. Other provider representatives stated they are also measuring health status for some conditions. The TEP suggested that CMS could implement the Patient Reported Outcome Measures (PROMs), as the National Health Service in the United Kingdom has done, to measure the performance of hospitals regarding the functioning of patients undergoing selected procedures.

### Measuring Appropriateness of Care

TEP members were supportive of including measures of appropriateness (i.e., overuse) in VBP programs, but panelists recognized that additional work is required to develop the definitions and engage providers in using these measures. They cautioned that without an external impetus, providers have little incentive to use practice guidelines or protocols that might withhold care due to the current fee-for-service and malpractice systems, which instead provide an incentive to increase the use of diagnostics and procedures. The TEP commented that providers under risk-sharing arrangements (e.g., ACO and total cost of care contracts) will be more likely to implement appropriateness guidelines, because the financial incentives they face are aligned with focusing on reducing the overuse of services that are not deemed appropriate. Based on direct experience, members of the TEP observed that when implementing appropriateness criteria measures in a health system, it can take years to get providers to buy-in related to establishing the criteria and being held accountable for performance against the criteria. TEP members suggested that measurement of shared decisionmaking is one of the keys to implementing appropriateness of care. A TEP representative of one health system noted the provider is piloting a process of "patient appropriate order entry" where the specialist has to attest that he or she held a discussion with the patient about the appropriateness of the care being recommended. Another TEP member recognized the challenge that physicians could face if appropriateness of care metrics are in conflict with patient preferences.[18]

### Enhancing the Ability of Electronic Health Records to Support Performance Measurement and Improvement

There was widespread agreement among the TEP members that it is important to incentivize and help providers build the infrastructure for quality improvement. EHRs may facilitate measurement and improvement, but the TEP did not see this happening in the near term. Based on their experiences to date, the panelists expressed concern that most EHRs are far from

including a comprehensive set of standardized data in data fields that can readily produce data needed to support the construction of performance measures, in part because providers who are the customers for EHRs are not demanding that EHRs be able to generate this type of information. Meaningful use requirements* currently require that EHR vendors build functionalities in EHRs to support reporting from a select list of quality measures. This is very different than freeing up the EHR data for use by providers for their own performance monitoring, improvement, and broader performance measurement. For example, some delivery systems have EHRs and registries that give providers alerts at the point of care on the patients' status with respect to a given measure and/or that allow providers to benchmark their performance on measures against their peers. ASPE staff commented that ASPE is working with the Office of the National Coordination for Health Information Technology, which is the lead federal agency responsible for meaningful use requirements, to make EHRs function more effectively to facilitate automated capture and reporting of quality measures, but this will be a long process.

### *Types of Incentives*

The review of public documents from program sponsors found that the types of financial incentives offered to providers have expanded beyond bonuses that have been commonly used in P4P programs, and which work at the margin, to a stronger set of incentives that more fundamentally alter payment arrangements. Examples include changes to fee schedules, shared savings arrangements (either alone or combined with bonuses or shared risk, in which the ACO loses money if targets for reducing patient costs are not met), and global budgets (i.e., overarching payment for all care delivered to a patient, similar to capitation). Most of the ACOs reviewed in our environmental scan have shared savings arrangements, and a few have shared risk. VBP programs often use combinations of financial incentives to drive change. The Blue Cross Blue Shield of Massachusetts Alternative Quality Contract (AQC)—an ACO-type arrangement—allows for shared savings and shared risk and offers a bonus payment up to 10 percent above the global budget based on performance on quality measures. The majority of the bundled payment programs for which we were able to identify information are offering shared savings to providers, while others adjust the episode fee based on quality performance.

---

* The Medicare and Medicaid EHR Incentive Programs provide incentive payments to eligible professionals, eligible hospitals, and critical access hospitals as they adopt, implement, upgrade, or demonstrate meaningful use of certified EHR technology. Eligible professionals can receive up to $44,000 through the Medicare EHR Incentive Program and up to $63,750 through the Medicaid EHR Incentive Program. (CMS, "Medicare and Medicaid EHR Incentive Program Basics," web page, no date. As of November 15, 2013: http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Basics.html.)

Although our review of the literature on VBP did not include a review of the use of consumer incentives, the TEP highlighted the importance of working to align incentives for consumers. Panelists commented that creating incentives to drive patients toward higher-performing providers could strengthen the impetus for providers to improve and might be more effective in shifting performance up than current P4P incentives that attempt to influence provider performance at the margin. CMS commented that it is already taking a number of actions in its VBP programs to affect consumer market behavior. For example, if a Medicare Advantage plan is consistently low-performing for three years, beneficiaries are not allowed to enroll online in that plan. Additionally, CMS sends letters to beneficiaries who are enrolled in low-performing Medicare Advantage plans and encourages them to shift to high-performing "five Star" plans; to facilitate plan switching, beneficiaries in low-performing contracts have the option of changing plans any time during the year. Panelists recommended that CMS continue to explore using tools like these to push quality improvement in a strategic way.

### *Type of Benchmarks/Thresholds*

An important design element of any VBP program is the performance benchmarks or thresholds that are used to determine who will receive an incentive payment. In some cases, these are absolute, fixed benchmarks (e.g., provider must have at least 90 percent performance on mammography screening), while in other cases benchmarks are relative (e.g., the provider's performance must in in the top 20th percentile of performance), and as a result the absolute score required to reach the percentile cut-point changes year to year. Some VBP programs reward providers for attaining specific benchmarks, improving over time, or a combination of attainment and improvement.

We were only able to find information about the types of benchmarks used for a third of the VBP programs in our environmental scan. There was no publicly available information about the benchmarks being used by bundled payment programs. Among P4P programs, the most common benchmark used was an absolute threshold only, followed by relative thresholds only, which may be based on the performance of peers in the market, the state, or nationally. Other programs, such as the CMS Hospital VBP program, have two paths to earning incentives: attainment against an absolute threshold or showing improvement over time.

Very little information was publicly available about the types of benchmarks being used for ACO models, as these are developed in the context of private negotiations between payers and providers. The exception was the three CMS ACO demonstration models. In its shared savings programs, CMS is establishing the cost benchmark for each agreement period for each ACO using three-years-prior expenditure data. Quality benchmarks are based on national percentile rankings from the year prior, and points are assigned on a sliding scale based on the ACO's performance. For 2013, the Pioneer ACO program measures and rewards improvement on the quality measures. The Physician Group Practice demonstration, the precursor ACO demonstration that CMS ran, utilized absolute thresholds for quality measures.

The literature highlights some of the issues associated with use of different types of benchmarks. Providers report disliking relative thresholds,[19, 20] for several reasons. First, providers do not know ahead of time what actual level of performance is required to obtain the incentive payment, creating much uncertainty about whether their performance is "good enough." Second, when topped-out measures are included in the VBP program, providers may have very high performance that does not meet the necessary threshold to receive the incentive, but yet is not meaningfully different from the performance of providers that do receive the incentive payment. For example, the initial design of the Premier HQID in Phase 1 of the program's implementation only paid hospitals that were in the top 20th percentile of performance. Performance rates for a large proportion of the hospitals hovered around 99 percent on a number of the measures, and which hospitals received the incentive payment was based on differences in performance at the second decimal point. In response to this problem, CMS changed the incentive structure in Phase 2 of the Premier HQID to reward above-average achievement and improvement.

A relative incentive structure can promote a "race to the top," creating perverse incentives for providers to allocate resources to improvement on a measure that may not yield the greatest clinical benefit and which may lead to overtreatment of patients. Achieving 100 percent performance on a measure also may not be appropriate and may lead to overtreatment. No matter how well the performance measure is constructed, and despite attempts to exclude from the denominator patients who should be excluded, it is unlikely that any process measure will be applicable to 100 percent of the population. In practice, there are often sound reasons why some small percentage of patients does not receive recommended processes of care. These reasons include patient preferences regarding treatment, contraindications to recommended therapy (e.g., allergies or intolerance of medications), prior rare side effects, and the clinical challenges of balancing treatment of multiple clinical conditions and interactions between medications. Typically, the patients in the upper tail of the distribution differ from patients in the other 95 percent of the distribution in ways that performance measurement typically is not very good at systematically capturing through exclusion criteria. In these cases, not providing the recommended care is not an error in care. In the UK Quality Outcomes Framework P4P program, where providers are allowed to exclude patients from the measure calculation (i.e., exception reporting), a median of 5.3 percent of patients were excluded from performance measure calculations. Exception reporting occurred most often for performance measures related to providing treatments and achieving target levels of intermediate outcomes.[21] U.S.-based VBP programs do not typically allow providers to exclude patients from reporting.

TEP members noted that while establishing absolute attainment thresholds is preferred by providers, some payers express concern that this approach removes the motivation for providers to continue to improve once the threshold has been attained. Paying all who achieve an absolute attainment target also creates budgeting challenges for payers, who will not be able to estimate how many providers they will need to pay; if the payer sets a fixed incentive pool, the more

providers who succeed results in a smaller incentive payment per provider. Some VBP sponsors have set multiple absolute targets along a continuum to motivate improvement at all levels of performance and to continue to motivate improvement at the top end of the performance distribution.

### *Performance of Value-Based Purchasing Programs*

VBP program sponsors and evaluators have primarily assessed whether improvements have occurred in the measures that were incentivized through VBP. Efforts to disentangle the VBP effect from other interventions designed to improve the delivery of health care locally and nationally (e.g., investments in HIT, enhanced quality improvement, and public reporting) have proven more challenging to study, because the natural experiments typically lack robust comparison groups. Furthermore, contextual factors and how they may contribute to any observed impacts are rarely considered.

The TEP highlighted some of the challenges with evaluations conducted over the past decade: (1) the measures included in a VBP program are often also included in national performance measurement and public reporting programs (e.g., CMS) and the VBP programs by other private sponsors, making it difficult to tease out the effect of any individual VBP program; (2) the presence of other incentives (e.g., public reporting/transparency of performance results) make it difficult to isolate the effects on incentivized measures of the financial incentives; (3) there is usually no comparison population when a VBP program is implemented statewide or nationally; (4) the size of payment incentives is often small; (5) VBP programs typically have used the same core measures (i.e., HEDIS, Joint Commission measures) that have been used for more than a decade and are largely "topped out"; and (6) there is a substantial lag for the data required to assess impact, such as data on avoiding admissions and readmissions.

#### Clinical Quality

##### *Pay-for-Performance*

We identified 49 studies that examined the effect of P4P on process and intermediate outcome measures: 37 studies examined the effect of P4P on process measures for physicians or physician groups;[5, 8, 10, 22–52] 11 studies examined the effect of P4P on process measures in the hospital setting;[53–60] and a single study examined the effect of P4P on process measures in other care settings.[61] The published studies have focused on assessing a few large P4P interventions (e.g., the Premier demonstration, the Physician Group Practice demonstration, the Integrated Healthcare Association P4P program, the Blue Cross Hawaii P4P program, the Massachusetts multi-plan P4P program, the UK Quality Outcomes Framework P4P program, and more recently the Blue Cross Blue Shield of Massachusetts AQC) and a number of very small-scale incentive experiments that were of short duration.

Overall, the results of the studies were mixed, and studies with stronger methodological designs were less likely to identify significant improvements associated with the P4P programs. Any identified effects were relatively small. Studies with weaker study designs mostly found that P4P was significantly associated with higher levels of quality, and many reported substantial effect sizes.

#### *Accountable Care Organizations*

We identified six evaluations (of five distinct ACO programs) examining the effect on quality of care associated with implementing an ACO or ACO-like model (e.g., the Blue Cross Blue Shield of Massachusetts AQC, which is a global budget total cost of care contract, and the CMS Physician Group Practice demonstration, which was a precursor to the CMS ACO demonstrations). Five of the studies investigated the effect of the ACO on a small number of process-of-care measures[62–66] and showed greater improvements than controls on some but not all of the measures. In addition to these evaluations, CMS issued a press release on the early experiences of the Medicare Pioneer ACO on July 16, 2013.[67] In the first performance year, the Pioneer ACOs had higher performance overall than the Medicare fee-for-service beneficiary comparison population on the 15 quality of care measures reported, but it was not reported whether the Pioneer ACOs had greater improvements or just higher baseline performance. At this stage, it is difficult to discern the effects of ACOs on quality, given the newness of the ACO model and the short period of implementation.

#### *Bundled Payments*

Of the three studies of bundled payments that include value-based payment design elements (cost and quality components), only one study examined the effect of bundled payments on process measures. The study found that adherence on 40 clinical process measures increased from 59 percent to 100 percent.[68] However, this study was conducted in a single integrated health system with unique characteristics that make generalizing the findings to other providers difficult. A recent systematic review of the bundled payment literature showed inconsistent effects on quality measures associated with implementing bundled payment arrangements. Most of the bundled payment programs reviewed in this study did not include quality elements as part of the incentive formula; in these instances, the evaluators sought to determine whether the application of bundled payments resulted in undesired effects on quality.[1]

### Outcomes

We reviewed 21 studies that evaluated the effect of P4P on outcomes in physician groups (12), hospitals (6), and other settings (3). In the physician practice setting, the studies generally focused on a small number of intermediate diabetes outcomes and found mixed results. Of the studies we rated as fair- and poor-quality in terms of their design, three[29, 33, 46] found between 2 and 22 percent improvement in the percentage of patients with HbA1c control, while another

studies found no effect.[27] There was only a single study rated as good-quality,[69] and it found that changes in diabetes intermediate outcome measures (e.g., percent of patients with HbA1c and lipid control) were not statistically significant from the comparison group. Four studies focused on other types of health outcome measures. One good-quality study[70] found that a P4P program focused on prenatal care for pregnant members of a union health plan led to a reduction in admissions to the neonatal intensive care unit (NICU), but no reduction in low birth weight. Three fair- and poor-quality studies[24, 39, 50] found no effect on mortality, readmission, or incident of major health events (e.g., stroke or heart attack), but did find a slight reduction in initial hospitalizations.

The studies in the hospital setting focused primarily on measuring the effects on mortality. Three of the studies that focused on outcomes were deemed to be of good methodological quality and found mixed results. Glickman[53] found no evidence that in-hospital mortality improvements were incrementally greater at P4P hospitals in the CMS Premier HQID program, while Ryan[71] found no evidence that the HQID had a significant effect on risk adjusted 30-day mortality acute myocardial infarction, CHF, pneumonia, or coronary artery bypass graft (CABG). Sutton et al.[72] found that risk-adjusted mortality for the conditions included in the P4P program decreased by 1.3 percent compared with controls in a study evaluating a program in the UK modeled after CMS HQID. Another study by Jha et al.,[73] which we deemed to be of fair quality, found no differences in a composite measure of 30-day mortality between hospitals in the HQID demonstration and hospitals exposed to pay-for-reporting. Mortality declined similarly across the two groups of hospitals (0.04 percent per quarter), and mortality rates were similar after six years of the pay-for-reporting demonstration. When considering the results from this study, it is important to note that hospitals exposed to the pay-for-reporting incentive increased their performance on the process measures similarly to pay-for-reporting hospitals, and both sets of hospitals topped out performance on these measures, so that there was no variation in performance to detect a differential effect.

One study,[74] which we rated as good, evaluated five states' Medicaid P4P programs in nursing homes and found that three of six outcome measures (the percentage of residents being physically restrained, in moderate to severe pain, and having developed pressure sores) improved a negligible amount, between 0.3 and 0.5 percent one year after P4P implementation. Performance on other targeted quality measures either did not change or worsened. Based on this study, it is unclear what the effects of P4P in the nursing home setting are. We also reviewed two studies that we deemed to be of fair quality. Hittle et al.[75] found that only two measures (improvement in pain interfering with activity and improvement in urinary incontinence), which were non-incentivized, showed significant differences between treatment and control home health agencies across one intervention year; otherwise, no differences were found in the incentivized measures. Shen[76] found that P4P was associated with a reduction in the proportion of clients in substance abuse clinics classified as most severely ill for three years post-intervention.

Among the studies evaluating ACOs, there is limited evidence that ACOs may reduce hospital readmission rates.[62, 63] Only one bundled payment study investigated the effect on health outcomes, and it found no effect.[68]

### Costs

#### *Pay-for-Performance*

Few studies have investigated the impact of P4P on costs. The studies with the strongest study designs report mixed effects on costs in the physician or physician group setting.[40, 70] Two studies with weak designs[3, 39] found evidence of significant cost savings and a positive return on investment. We found only two studies that specifically investigated changes in costs in the hospital setting. Both of these studies were based on the HQID, and neither found any significant effects on hospital costs, revenues, margins or Medicare payments.[77, 78]

#### *Accountable Care Organizations*

All of the studies we reviewed attribute various degrees of cost savings for the shared savings payment model, but not all of the individual ACOs were able to generate statistically significant savings relative to controls.[65, 66, 62–64] CMS also reported that the costs for the Pioneer ACO beneficiaries increased 0.3 percent in 2012 compared with 0.8 percent growth for similar Medicare fee-for-service beneficiaries. While 13 of the 32 ACOs shared savings with CMS, two Pioneer ACOs had shared losses. Two Pioneer ACOs were leaving the ACO program, and an additional seven were switching to the Medicare Shared Savings Program, which involved less risk to providers. Because there were only six studies of four programs, the studies were of short duration, and several had poor or no comparison group, the evidence is insufficient to make conclusions about the impact of ACO payment structures on costs.

#### *Bundled Payments*

Of the two studies investigating the impact of bundled payments, both identified reductions in costs. One found a reduction in hospital charges of around five percent,[68] while another found a reduction in costs per case of roughly $2,000 over a two-year period.[79] The systematic review that documented the impact of implementation of 19 bundled payment programs[1] found that all programs showed declines of 10 percent or less in spending and utilization.

## *Unintended Effects*

We examined undesired behaviors (often referred to as unintended consequences) and spillover effects to assess any unintended effects from these programs. Undesired effects include provider gaming of the data used to generate scores, ignoring other clinically important areas that are not measured and incentivized by the P4P program, avoiding sicker or more challenging patients when providing care, providing care that is not clinically recommended, and overtreating patients. Other undesired effects are an increase in disparities in treatment or outcomes among

patients and the VBP program having harmful effects on providers who serve more challenging patient populations. Spillover effects occur when changes made to improve areas measured by VBP programs extend to other areas not included in the VBP program. The literature was sparse related to undesired and spillover effects; few studies have looked at the main effects of VBP interventions, let alone their side effects.

### Pay-for-Performance

We identified 21 articles that examined undesired behaviors and spillover effects in P4P programs. Most of the published evidence regarding undesired effects related to application of P4P shows either small or no effects. However, recent studies in the Veteran's Administration found evidence of overtreatment of patients with hypertension and diabetes associated with use of intermediate outcome measures that use thresholds.[80–82] These authors have called for moving from the current class of dichotomous target measures (i.e., met or didn't meet a threshold such as HbA1c <7), where there is a push to get all patients to the threshold, to a set of improved performance measures that focus on giving providers credit for appropriate clinical actions taken (intensification of medications, being on maximal medications, contraindications to further treatment, etc.) and which account for individual risks and preferences. An improved set of performance measures could help reduce incentives to overtreat patients. In addition to the selection of appropriate performance measures, VBP program sponsors should conduct monitoring studies[83] to assess whether and how often patients may be receiving inappropriate treatment so that they can adjust the measures included in VBP programs to mitigate these effects. The lack of evidence on observed negative effects in other P4P studies may be due to the fact that many of the P4P interventions studied were small in scale, of short duration, and did not have substantial amounts of revenue at risk that might encourage providers to engage in undesired behaviors.

Our review of the literature found a small number of studies (n=5) that examine whether P4P programs have spillover effects. The P4P studies have found mixed effects, with some finding no effects (either positive or negative) on measures that were non-incentivized,[53, 84] one finding negative effects,[85] and, in a few cases, evidence of improvement on non-incentivized measures within the same conditions that were the target of the incentives.[42, 86] The evaluation of the UK Quality Outcomes Framework P4P program found that that both incentivized and non-incentivized measures improved between 2004 and 2005 for asthma, diabetes, and heart disease, but that the mean quality scores for aspects of care that were not linked to incentives (only for asthma and heart disease) declined between 2005 and 2007 while the mean scores for the incentivized measures continued to increase. Group practices participating in the CMS Physician Group Practice demonstration reported implementing a variety of quality improvement and care management programs, information technology, and patient registries, all of which have the potential to improve quality of care beyond the measures included in the demonstration; however, no spillover effects were measured.

### Accountable Care Organizations

Because these models are newly being implemented and have yet to gain experience, there are no studies that have examined unintended consequences in ACO models, and only one study that assessed spillover effects. A recent study by McWilliams et al.[87] found spillover effects to the Medicare population from implementation of the Blue Cross Blue Shield of Massachusetts's AQC, which targeted commercial HMO enrollees. This study examined changes associated with the AQC in spending and quality of care for traditional fee-for-service Medicare beneficiaries and found that the AQC was associated with lower spending for Medicare beneficiaries but not with consistently improved quality. The AQC evaluation research team also has examined the effect on quality measures not included in AQC, particularly for children with special needs; in this case, they observed more improvement for generic prescribing measures, but no effect on other measures that were not incentivized. Within the AQC practices, improvements were larger for ACQ members (HMO members), and there did not seem to be spillover effects to the Blue Cross Blue Shield of Massachusetts PPO members; by extension, the study team doubted there would be spillover improvements for PPO patients for other health plans. A TEP member who represented the AQC cited two possible reasons for the absence of spillover effects: (1) Blue Cross Blue Shield of Massachusetts has provided physician practices with better data on ACQ members than other plans' members, so a provider's behavior changes only for the AQC patients, since they have better data to manage those patients; and (2) the practices have used case managers and other resources for high-risk subgroups covered by the AQC, and these resources are not available for other high-risk patient populations they serve. Other TEP members agreed that this is a common occurrence, as health plans focus on providing resources for their members who are the focus of the VBP programs.

ACOs are expected to implement a variety of quality improvement and care management programs, information technology, and patient registries, which have the potential to improve quality of care more broadly and which could generate positive spillover effects. Some researchers and policymakers have expressed concerns that the formation of ACOs may lead to greater market concentration and have the adverse effect of raising prices; the TEP expressed similar concerns. One TEP member commented that in Massachusetts, a law was passed in 2012 that sets a maximum rate of growth in health care spending by providers and hospitals, which holds providers accountable. This law established guardrails and protects against the effects of excessive consolidation. The TEP suggested that a similar law in other states or nationally could be a strong policy lever to guard against this type of behavior.

### Bundled Payments

We found no evidence of unintended effects or spillover effects from the three studies of bundled payments that included quality measures. The Hussey et al.[1] review of the broader bundled payment literature highlighted the types of undesired effects that it has been hypothesized might occur in the context of bundled payment arrangements: increasing the number of bundles

(volume), underuse of appropriate care services that may lead to poorer outcomes for patients, selection of low-risk patients into the bundles and avoidance of high-risk (potentially more expensive) patients, upcoding to maximize payment for the bundle, and moving services in time or location to qualify for separate reimbursement. However, Hussey et al. found limited evidence on unbundling services and upcoding, but consistent evidence regarding shifting services to other settings of care (e.g., from inpatient to outpatient). There was little evidence that there were major effects on quality; rather, the findings were mixed, with some measures having improved while other worsened.

The TEP supported the need to monitor spillover effects in VBP programs. To assess spillover effects on quality requires access to data for other measures (within the same clinical condition or addressing other clinical conditions) that were not incentivized by the program, something that most programs do not routinely collect. The TEP also identified multiple possible unintended consequences, the occurrence of which should be monitored, including the loss of revenue for providers caring for disadvantaged populations, the excessive exclusion of patients when that is an option in the program, access barriers and patient turnover from practices related to providers avoiding more difficult patients, and market concentration and price effects in the context of ACOs.

### *Effect on Disparities*

Many P4P studies have commented about possible unintended effects for patients of low socioeconomic status (SES) and the providers that serve these populations (e.g., safety net clinics and hospitals). Examinations of whether VBP programs work to reduce or increase disparities are challenged by the lack of information at the patient level on race, ethnicity, education, SES, and other markers of vulnerable populations prone to disparities.

We found only five empirical studies that assessed the effects of P4P on disparities. Among the four studies that evaluated U.S. P4P programs, three found no effects related to increasing or decreasing racial/ethnic or SES disparities while one[88] poor-quality study found very small significant differences in baseline performance for hospitals with a high disproportionate share hospital (DSH) index comparing HQID P4P and pay-for-reporting hospitals (between –0.5 percent and –1.1 percent lower performance for high DSH-index hospitals versus non-high-DSH-index hospitals).* Three years post-HQID-intervention based solely on attaining performance in the top 20th percentile of performance distribution, there were modestly greater gains (only a few significant) for the high-DSH-index hospitals compared with the non-high-

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

DSH-index hospitals exposed to P4P (e.g., 0.6 percent to 1.2 percent higher), and no differences in performance were observed between high-DSH-index and non-high-DSH-index hospitals exposed to P4P. This study should be interpreted in light of the fact that differences at baseline were negligible, and nearly all hospitals in both the P4P and pay-for-reporting groups topped out their performance on the clinical process measures that were the focus of this study.

The 2010 Ryan study,[89] which had a strong design, found no negative access effects related to avoiding treating minority patients after introduction of the Premier HQID. A more recent (2012) study by Ryan et al.[58] found that changes to the HQID incentive structure between Phase I and II of the program resulted in a redistribution of available incentive payments, with a greater proportion going to hospitals with greater socioeconomic disadvantage (as measured by the DSH index). This effect was a function of changes in the structure of the incentive and not due to lower-performing hospitals actually improving more.[90] This study found that disparities neither had worsened nor reduced. A study from the United Kingdom[91] showed a lessening of the disparities gap in performance among primary care practices, with measures largely topping out on performance; however, the results of this study are not generalizable to the United States due to substantial differences in the delivery system (national health system, national HIT platform in primary care practices) and design of the P4P program. There are currently no empirical studies on disparities for either ACO or bundled payment VBP models.

A TEP member from one large commercial health plan noted that a global-budget contract model with strong quality incentives had driven important gains in closing racial and ethnic disparities. This is because a few medical groups with a low-SES patient mix worked to innovate with their population and to get their doctors to improve quality. These provider groups with low-SES patient populations actually achieved some of the highest gains and absolute quality scores in the state. However, this was not a universal finding among all groups with low-SES patients.

While the TEP recognized the importance of monitoring the effects of VBP programs on disparities in care, panelists also noted that assessing the effect of VBP on disparities is difficult to monitor due to the lack of routinely collected data on the demographic and socioeconomic characteristics of patients. TEP members indicated that they had faced challenges in capturing this information, despite their interest in capturing self-reported language, health literacy, and indicators of patient vulnerability to help improve their ability to work with patients. However, several providers on the TEP stated they were making inroads in the data they capture to be able to examine disparities. For example, one delivery system has a mandatory data gathering protocol for zip code, race, and ethnicity.

### *Characteristics of High- and Low-Performing Providers*

There is limited evidence characterizing high- and low-performing providers under VBP. The few studies that do describe characteristics of high- and low-performing providers have been opportunistic in defining the characteristics based on the variables that were available to them

(e.g., provider size and type), rather than considering a broad set of factors that might differentiate high and low performers. The TEP noted that the American Medical Group Association has developed a set of elements for what defines the characteristics of a high-performing health system;[92] however, it remains untested whether these elements differentiate high and low performers under VBP.

Most of the studies that looked at provider characteristics focused on physician or physician group P4P programs. The limited literature shows that higher-performing providers tend to be large provider organizations,[7, 43, 69] have a medical group rather than an independent practice association organizational structure, have more HIT infrastructure,[93–96] and have been historically high performers. Other studies find that high performers engage in more care management processes,[7] use order sets and clinical pathways for measured areas,[97] have nursing staff's support for quality indicators, have adequate human resources for initiatives to improve performance,[97] and engage in more external quality improvement initiatives.[7] High performers also served a smaller fraction of low-SES or Medicaid patients.[43, 88] Lower-performing providers under P4P programs tended to serve a lower-SES population (i.e., physician organizations with more Medicaid patients[43, 69, 98] or hospitals with a high DSH index[88]). Hospitals that achieved the largest improvements under P4P are characterized as being well financed, operating in less competitive markets,[56] having lower performance at baseline,[58, 59] and having a higher DSH index.[88]

Although associations have been found between patient population SES and provider performance, it is important to note that some providers that serve low-SES populations are able to perform well. For example, Medicare has found that most hospitals with high proportions of Medicaid patients achieve readmission rates comparable to those with fewer Medicaid patients.[98]

The CMS Physician Group Practice demonstration evaluation highlighted organizational characteristics associated with performance. Physician groups characterized as being either affiliated with an academic medical center or a freestanding physician group practice were more able to achieve both quality and cost targets than groups with only non-academic hospital affiliations. It is unclear whether the results based on the 10 physician groups that self-selected into the Physician Group Practice demonstration would generalize more broadly. Case studies and commentaries suggest that strong physician leadership with a clear strategy and vision is necessary to change practice culture to one that is comfortable with sharing the risk of a predetermined patient population.[99–102] There have been no studies of VBP-type bundled payment models conducted that compare the features of high and low performers under these programs; implementation of these models has proven challenging, and there are few models that have been evaluated.

### *Features of Successful Value-Based Purchasing Programs*

There is very limited published literature to inform what structural and implementation features are associated with successful P4P programs. It is rare to find studies that examine the effects of

alternative design features (e.g., the size or frequency of the incentive payment) to assess their impact on provider behavior; the studies that exist are typically small-scale, of short duration,[103] and in many cases the intervention being tested was not expected to be permanent, so providers would not have been expected to invest in practice redesign to improve outcomes and obtain rewards. Consequently, it is difficult to assess from these studies whether the programs have been successful and would be if scaled up to a larger number of providers (i.e., statewide or nationally), what would have happened if the intervention was sustained, and what can be generalized to implementing P4P in the same setting or other settings.

Based on the review of the published literature, there have been mixed findings on the effectiveness of VBP programs to meet its intended goals to improve quality and control costs. This may be because VBP programs are still a work in progress and sponsors are continuing to evolve these programs in response to what does and does not work when implemented. Despite the fact that many programs have been in operation for the past five to ten years, there is a substantial gap in the knowledge base about what has been learned regarding design and implementation in large P4P programs to inform what features promote success in VBP programs.

ACOs are new, and there has not been sufficient time to test ACOs to know whether they can succeed and what factors must be present to allow them to form and achieve desired goals. There is, as yet, little accumulated knowledge about their formation and, once formed, what types of performance results are accrued and what factors are associated with observed performance results. Evaluations of the private- and public-sector ACO experiments will hopefully generate knowledge to inform what factors need to be present for an ACO to succeed in meeting performance goals. Various challenges associated with implementing bundled payments have been identified,[104] and, similar to ACOs, these models are not well tested or in routine operation.

When we queried the TEP about the features of successful VBP programs based on their knowledge from having designed and operated these programs, most panelists agreed that the evidence is thin regarding successful programs and what features characterize these programs. Based on the panelists' anecdotal evidence and the limited literature, we identified six features that appear to influence the success of VBP programs:

- **Sizable incentives:** A limited number of studies have shown that larger incentives were associated with a larger impact on performance.[42, 56] Incentives that were large enough to compensate providers for the effort required to obtain them was identified as one characteristic associated with more successful programs in a study of P4P in five Medicaid plans.[44] Researchers who have found limited effects associated with P4P programs have hypothesized that incentives were too small to garner the attention of providers, but there is uncertainty about how big incentives need to be to garner the desired response and investment for improvement by providers while also minimizing the likelihood of unintended consequences. Absolute incentive size is influenced by the size of the program's incentives (e.g., 1 or 2 percent of base payment), the size of the base payment (e.g., diagnostic-related group [DRG] payment amount) and the number of a

provider's patients who are covered by the program, as incentives are often computed on a per capita basis. An important policy consideration regarding the size of the incentive relates to the fact that in U.S. VBP programs, payers fund the incentive payment in a budget-neutral fashion, meaning that the winnings of high-quality providers are financed by the loss of revenue from poor-quality providers. In this situation, increasing the size of the incentives could potentially lead to large redistributions of resources between providers and have the undesired effect of de-resourcing low-quality providers who may be most in need of resources to be able to improve quality.

- **Measure alignment:** A number of TEP members discussed the importance of measure alignment across VBP programs to give providers a clear signal of what is important. However, if different VBP programs cover different patient populations, then it is more important for measures to align with the population's conditions than with other VBP programs. If programs are measuring an area where established measures exist, they should use the measures as defined and not tweak the measures to promote alignment.
- **Provider engagement:** A few studies have identified the involvement of key stakeholders in the P4P system design and implementation as important.[4, 105] Similarly, a number of TEP members discussed the importance of provider engagement in design and implementation of VBP (e.g., providing input on the design of the program, participating in choosing performance measures and targets).
- **Performance targets:** TEP members discussed the importance of the methodology used to measure and reward performance. Members stressed the importance of rewarding both achievement and improvement (such as was used in the second phase of the Premier HQID) and that VBP programs should not be designed as a "tournament" wherein relative thresholds are used and providers are pitted against each other (which was how the incentive was structured in Phase 1 of the HQID and in many other P4P programs). Some TEP members recommended that the reward should be based on objective targets that are defined prior to the start of the measurement year in absolute terms; if a provider hits those targets, it should receive an incentive payment. Providers can then strive to achieve a number of targets along a continuum and compete against themselves rather than competing with other providers for a limited number of "winning positions" (e.g., top 20th percentile of performance). This approach provides motivation for all providers to move up the scale.
- **Data and other quality improvement support:** There was an extensive discussion among the TEP of the importance of support to help providers improve, particularly through the use of HIT and data registries. It was also noted that best practices for sharing, consultative support, health coaching, and other infrastructure building are important types of support to make available to providers participating in VBP.

### *Dissemination of Best Practices from Highest-Performing Providers*

TEP members stated that the dissemination of best practices currently occurs through trade conferences and regional quality improvement activities. Although the information from these conferences is not published, several provider organization TEP members observed that they do provide vital information for organizational learning of best practices and improvement strategies. Panelists said that it would be useful to extract and compile lessons learned from providers about best practices they have implemented and to widely disseminate this

information. Some panelists recommended that HHS should conduct case studies of high-performing providers to see what factors they identify as contributing to producing positive results; however, because high performers may be doing many of the same things as low performers, it is necessary to look at both high and low performers to see what differentiates them.

Alternative approaches to disseminating best practices were discussed by the TEP. Some TEP members felt that for dissemination to be effective, awareness is necessary of how low-performing organizations/providers with different resources and capabilities than the high performers will interpret and use the information that is being disseminated. Some providers may be more receptive to the information if the provider is "like them," and benefit from peer-to-peer coaching by providers located in their own community who have similar characteristics to overcome resistance to adoption of certain practices. Other providers who are willing to innovate may look to other organizations for their "good ideas" as a way to continue to improve, regardless of where they are located or their characteristics, and will embrace best practices from dissimilar organizations or practices.

### *Monitoring and Evaluation of Value-Based Purchasing Programs*

#### Qualitative Evaluation

The TEP broadly agreed that there is a need for qualitative research to understand what has been learned by those who design and sponsor VBP programs and by the providers who are targets of the VBP programs. There has been a lot of iterative work by VBP program sponsors, and case studies could shed light on lessons learned that are not making their way into the published literature. Qualitative research focused on understanding what does and does not work regarding design and implementation would be useful to those designing VBP programs. For example, it would be useful to learn how providers have used performance benchmarking data provided by both public and private VBP programs to inform their quality improvement efforts and engage leadership in organizational infrastructure investments to support high-value care. One TEP member suggested Qualitative Comparative Analysis[106, 107] as one qualitative analytic methodology that might be a good fit for VBP evaluations, as it attempts to isolate key factors that are necessary conditions, versus those that are sufficient conditions, to achieve the outcome. This approach acknowledges that there are a number of possible paths or combinations of elements (e.g., alternative designs) that may lead to the desired outcome. The other area flagged by the TEP where qualitative work would be beneficial is understanding what changes providers are making in response to VBP programs. Although the TEP emphasized the need for qualitative evaluation work, there may be challenges in getting private VBP sponsors to share proprietary information, particularly in a competitive marketplace.

### Quantitative Assessment of Impacts

The TEP supported the need to evaluate the impact of VBP programs, and panelists felt that having a common set of variables that potentially influence outcomes, such as program characteristics (e.g., size and type of incentives), market characteristics (e.g., extent of monopoly power among providers in the market), provider characteristics, and other facilitators/enablers, would facilitate this work. They also noted the importance of having a comparison group, as reflected by one TEP member's comment: "We need to avoid marketing techniques that claim to achieve reduction in trends when the trends were happening anyway." A comparison group guards against this possibility.

## Conclusions

Although the past decade has witnessed a fair amount of experimentation with performance-based payment models, primarily P4P programs, we still know very little about how best to design and implement VBP programs to achieve stated goals and what constitutes a successful program. The published evidence regarding improvements in performance from the P4P experiments of the past decade is mixed (i.e., positive and null effects); where observed, improvements were typically modest. Many of the published studies evaluating the impact of P4P programs suffer from methodological weaknesses that make it hard to determine whether the VBP intervention had an effect above and beyond other changes (e.g., investment in quality improvement support, public reporting, health information technology [HIT] investments and support) that were simultaneously occurring to improve quality and restrain spending.

VBP programs are natural experiments and inherently difficult to evaluate because program sponsors rarely withhold the VBP intervention from a matched group of providers to see what would have occurred absent the intervention. There are many weaknesses in the methods often used to evaluate P4P (and now the broader class of VBP programs), including reliance on pre-post comparisons without a comparison group that was not exposed to the intervention, comparisons with populations of providers that are substantially different from the treatment group, and failure to account for other factors that may be contributing to the observed results.

ACOs and bundled payment programs that embed clinical quality measures have only recently emerged and are just now being tested and evaluated. There is currently very limited evidence regarding the impact of these programs and whether they can be successfully implemented. Only a handful of ACO evaluation studies have been published, and these evaluations have been of relatively short duration (i.e., 1–2 years), making it difficult to know whether the results are real and can be sustained. These studies also suffer from similar methodological weaknesses as seen in the P4P literature. The published studies show some improvements in cost and quality; however, several of the ACO studies reported cost savings compared with expected year-over-year trend in spending as opposed to comparing the intervention providers' experience against a matched comparison group of providers. Bundled

payment programs that incorporate a quality component are equally new, and there is virtually no evidence on whether they can be successfully implemented and what their effects are.

The paucity of publicly available information regarding what constitutes a successful VBP program—that is, what VBP design features and other factors (i.e., characteristics of the providers, the health care market where the VBP program is implemented, and policy/regulatory environment) facilitate success in VBP—presents challenges for policymakers who seek to design VBP programs. In practice, more is likely known about what does and does not work in terms of VBP design and implementation than what the published literature suggests. VBP program sponsors (particularly private program sponsors) have gained a great deal of experience through trial and error as they work to operationalize the VBP concept in real-world settings; however, these experiences are not being documented through traditional means. Because VBP programs are relatively new and experimentation is likely beneficial at this stage of VBP development, the question is how to generate information from all the experimentation. Efforts to extract these lessons from VBP sponsors are critically needed to strengthen the knowledge base.

# Acknowledgments

The authors would like to thank Stephanie Glier (project officer), Dr. William Borden, and Dr. Lok Wong Samson from the Office of Health Policy within the Office of the Assistant Secretary for Planning and Evaluation (ASPE) for their valuable guidance and feedback throughout the project. We also thank Nancy De Lew and Dr. Pierre Yong of ASPE, Drs. Timothy Cuerdon and Jordan VanLare of the Centers for Medicare and Medicaid Services (CMS), and Drs. Richard Kronick and Irene Fraser of the Agency for Healthcare Research and Quality (AHRQ) for their insightful comments during the meetings of the technical expert panel and their thoughtful reviews of this report.

We are especially indebted to our expert panelists, who gave generously of their time to participate in panel discussions. These individuals graciously shared the lessons they have learned on the "front line" as VBP program designers and implementers, as providers who have had to respond to VBP programs, and as VBP program evaluators.

**ASPE Measuring Success in Value-Based Purchasing Technical Expert Panel**

| | |
|---|---|
| **Adams Dudley, MD, PhD**<br>University of California at San Francisco | **Andrew Ryan, PhD**<br>Weill Cornell Medical College |
| **Patrick Falvey, PhD**<br>Aurora Health Care | **Dana Safran, PhD**<br>Blue Cross Blue Shield of Massachusetts |
| **Tammy Fisher**<br>Partnership HealthPlan of California | **Barbara Walters, MD**<br>Dartmouth-Hitchcock |
| **John Hirshleifer, MD**<br>Blue Shield of California | **Rachel Werner, MD, PhD**<br>ASPE and University of Pennsylvania |
| **Elizabeth Mort, MD**<br>Partners HealthCare, Inc | **Tom Williams, DrPH, and Dolores Yanagihara**<br>Integrated Healthcare Association |

We also acknowledge the contributions of the following federal participants who provided the public payer perspective during the discussions of the Technical Expert Panel: Elizabeth Goldstein (CMS, Consumer Assessment and Plan Performance), John Pilotte (CMS, Performance-based Payment Policy Group), James Poyer (CMS, Division of Value, Incentives, and Quality Reporting), and Mark Wynn (CMS).

We thank Drs. Jon Christianson (University of Minnesota) and Richard C. Neu (RAND) for their careful review and comments on the draft of this report, contributions that strengthen the final product. We also acknowledge the important role played by RAND team members Roberta Shanman, Margaret Maglione, and Lynn Polite, who provided research assistance, helped with review of the literature, and provided project support.

# Abbreviations

| | |
|---|---|
| ACE | acute care episode |
| ACO | accountable care organization |
| AHRQ | Agency for Healthcare Research and Quality |
| AMI | acute myocardial infarction |
| AQC | Alternative Quality Contract |
| ASPE | Assistant Secretary for Planning and Evaluation |
| CABG | coronary artery bypass graft |
| CalPERS | California Public Employees' Retirement System |
| CHF | congestive heart failure |
| CI | confidence interval |
| CMP | care management process |
| CMS | Centers for Medicare and Medicaid Services |
| DRG | diagnosis related group |
| DSH | disproportionate share hospital |
| EHR | electronic health record |
| FFS | fee-for-service |
| GWTG-CAD | Get With The Guidelines–Coronary Artery Disease |
| HAC | hospital acquired condition |
| HHS | U.S. Department of Health and Human Services |
| HbA1c | glycolated hemoglobin test (i.e., blood glucose level) |
| HEDIS | Healthcare Effectiveness Data and Information Set |
| HMO | health maintenance organization |
| HQA | Hospital Quality Alliance |
| HQID | Hospital Quality Incentive Demonstration (Premier) |
| IHA | Integrated Healthcare Association |
| IPPS | Inpatient Prospective Payment System |
| LDL | low-density lipoprotein |
| LOS | length of stay |
| MEDPAR | Medicare Provider and Analysis Review |
| MMR | measles, mumps, rubella |
| MSSP | Medicare Shared Savings Program |
| NCQA | National Committee for Quality Assurance |
| NICU | neonatal intensive care unit |
| OB/GYN | obstetrics and gynecology |
| OR | odds ratio |
| P4P | pay-for-performance |
| PCMH | patient-centered medical home |
| PCP | primary care practitioner |
| PGP | Physician Group Practice |
| PMPM | per member per month |
| PPO | preferred provider organization |
| QALY | quality-adjusted life year |

| | |
|---|---|
| QI | quality improvement |
| RCT | randomized controlled trial |
| SCIP | Surgical Care Improvement Project |
| SES | socioeconomic status |
| TEP | technical expert panel |
| VBP | value-based purchasing |

# 1. Introduction

Value-based purchasing (VBP) refers to a broad set of performance-based payment strategies that link financial incentives to providers' performance on a set of defined measures. Both public and private payers are using VBP strategies in an effort to drive improvements in quality and to slow the growth in health care spending. Nearly 10 years ago, the Department of Health and Human Services (HHS) and the Centers for Medicare and Medicaid Services (CMS) began testing VBP models with their hospital pay-for-performance (P4P) demonstrations, known as the Premier Hospital Quality Incentive Demonstration (HQID) and the Physician Group Practice (PGP) Demonstration, which provided financial incentives to physician groups that performed well on quality and cost metrics. The use of financial incentives as a strategy to drive improvements in care dates back even further among private payers[2,9] and Medicaid programs, which began to experiment with P4P in the mid-1990s and early 2000s. These early private payer P4P programs generally focused on holding providers accountable for their quality performance and targeted physician groups, individual physicians, and hospitals.[19, 20, 52]

Although the published evidence from P4P programs implemented by private-sector payers between 2000 and 2010 showed mostly modest results in improving performance,[3–10] public and private payers have continued to experiment with the use of financial incentives as a policy lever to drive improvements in care. Many of the early P4P program designs have evolved over time to include a larger and broader set of measures, including resource use and cost metrics, in an effort to reward providers for delivering value,* and many are deploying a wider range of incentives. Additionally, other VBP models have since emerged and are currently being tested, including accountable care organizations (ACOs) and bundled payment programs that include both quality and cost design features.

## Policy Context and Study Purpose

The Medicare program has gradually been moving toward implementing VBP across various care settings starting with pay-for-reporting programs (e.g., the Hospital Inpatient Quality Reporting program and the Physician Quality Reporting Initiative) and P4P demonstrations to gain experience. The 2010 Patient Protection and Affordable Care Act[11] significantly expands VBP by requiring the Medicare program to implement, develop plans for, and test in the context

*Value is defined as the outcomes (outputs) achieved divided by the cost or resources used (inputs) to generate those outcomes.

of demonstrations the use of VBP across a broad set of providers and settings of care (i.e., physicians, skilled nursing homes, home health agencies, ambulatory surgery centers, long-term hospitals, rehabilitation hospitals, cancer hospitals, psychiatric hospitals, and hospice facilities), as shown in Table 1.1. For example, the Patient Protection and Affordable Care Act required HHS to submit plans for implementing VBP in ambulatory surgery centers, home health, and skilled nursing homes to Congress in 2011. The Hospital Value-Based Purchasing Program, which makes payment adjustments (both bonuses and penalties) to hospitals based on performance, began implementation in October 2012, and the Physician Value-Based Payment Modifier will start in January 2015. The Patient Protection and Affordable Care Act further links provider payments to cost reductions and quality improvements through the implementation and testing of VBP models such as ACOs and bundled payments. To that end, Medicare has begun the Bundled Payments for Care Improvement demonstrations and the ACO shared savings programs and demonstrations. Moreover, Congress is actively considering ways to revise the physician fee schedule (i.e., the sustainable growth rate or SGR) to incorporate VBP incentives so that payment policy for physicians paid under fee-for-service (FFS) supports the delivery of high quality care and efficient use of resources.

**Table 1.1. 2010 Patient Protection and Affordable Care Act Value-Based Purchasing Provisions**

| Type of Value-Based Purchasing Program and Setting | Timeline |
|---|---|
| **Pay-for–Performance** | |
| Hospital Value-Based Purchasing Program | October 1, 2012 (current program) |
| Physicians (or groups of physicians) under Physician Value-Based Payment Modifier | January 1, 2015, for a subset of physicians<br>January 1, 2018, for all physicians<br>(program to be implemented) |
| Inpatient critical access hospitals | No later than 2 years after date of act (May 1, 2010) (demonstration program) |
| Hospitals excluded from the Hospital Value-Based Purchasing Program due to insufficient numbers of measures and cases | No later than 2 years after date of act (May 1, 2010) (demonstration program) |
| Long-term care hospitals | No later than January 1, 2016 (pilot program) |
| Hospice programs | No later than January 1, 2016 (pilot program) |
| Psychiatric hospitals | No later than January 1, 2016 (pilot program) |
| Rehabilitation hospitals | No later than January 1, 2016 (pilot program) |
| PPS-exempt cancer hospitals | No later than January 1, 2016 (pilot program) |
| Ambulatory surgical centers | Submit plan to Congress no later than January 1, 2011 (plan for program) |
| Home health agencies | Submit plan to Congress no later than October 1, 2011 (plan for program) |
| Skilled nursing facilities | Submit plan to Congress no later than October 1, 2011 (plan for program) |
| **Shared Savings** | |
| ACOs | no later than January 1, 2012 (current program) |
| **Bundled Payment** | |
| Hospital/physicians/post-acute care | no later than January 1, 2012(demonstration program) |

Despite the widespread enthusiasm for and adoption of P4P programs over the past decade, uncertainty remains about the extent to which these programs have been successful in accomplishing their goals and what design elements work best and under which conditions.[6, 20, 108] Various studies have identified a number of issues, such as uncertainty about the size of incentives required to attain desired changes,[108] whether financial incentives may lead to unintended consequences,[43, 109–111] and measurement issues, including measure reliability and misclassification risk.[112–114]

VBP models are quite new to the health system and represent a work in progress in terms of our understanding of how best to design these programs to achieve desired goals, the optimal conditions that support successful implementation, and provider response to the incentives. Because of the substantial investments that HHS is making to implement and test a variety of VBP models, this is an opportune moment to reflect on what has been learned from the past decade of experimentation that could guide current and future federal policymaking related to VBP program design and implementation. It is also an important moment to consider the type of monitoring and systematic evaluation work that is needed to generate the information that policymakers require to fine-tune VBP program designs and to understand the impact these programs are having related to stated goals.

To that end, the Office of the Assistant Secretary for Planning and Evaluation (ASPE) in HHS asked RAND to review what has been learned about VBP over the past decade that might help inform policymaking. In particular, ASPE asked RAND to address three overarching questions: (1) What is the evidence regarding whether VBP programs have been successful? (2) What are the elements of successful VBP programs? (3) What questions remain unanswered that, if answered, could improve the design and functioning of VBP programs moving forward? This report presents the findings from RAND's review. For recommendations based on the findings, we direct readers to the companion document to this summary report, *Measuring Success in Health Care Value-Based Purchasing Programs: Summary and Recommendations*.

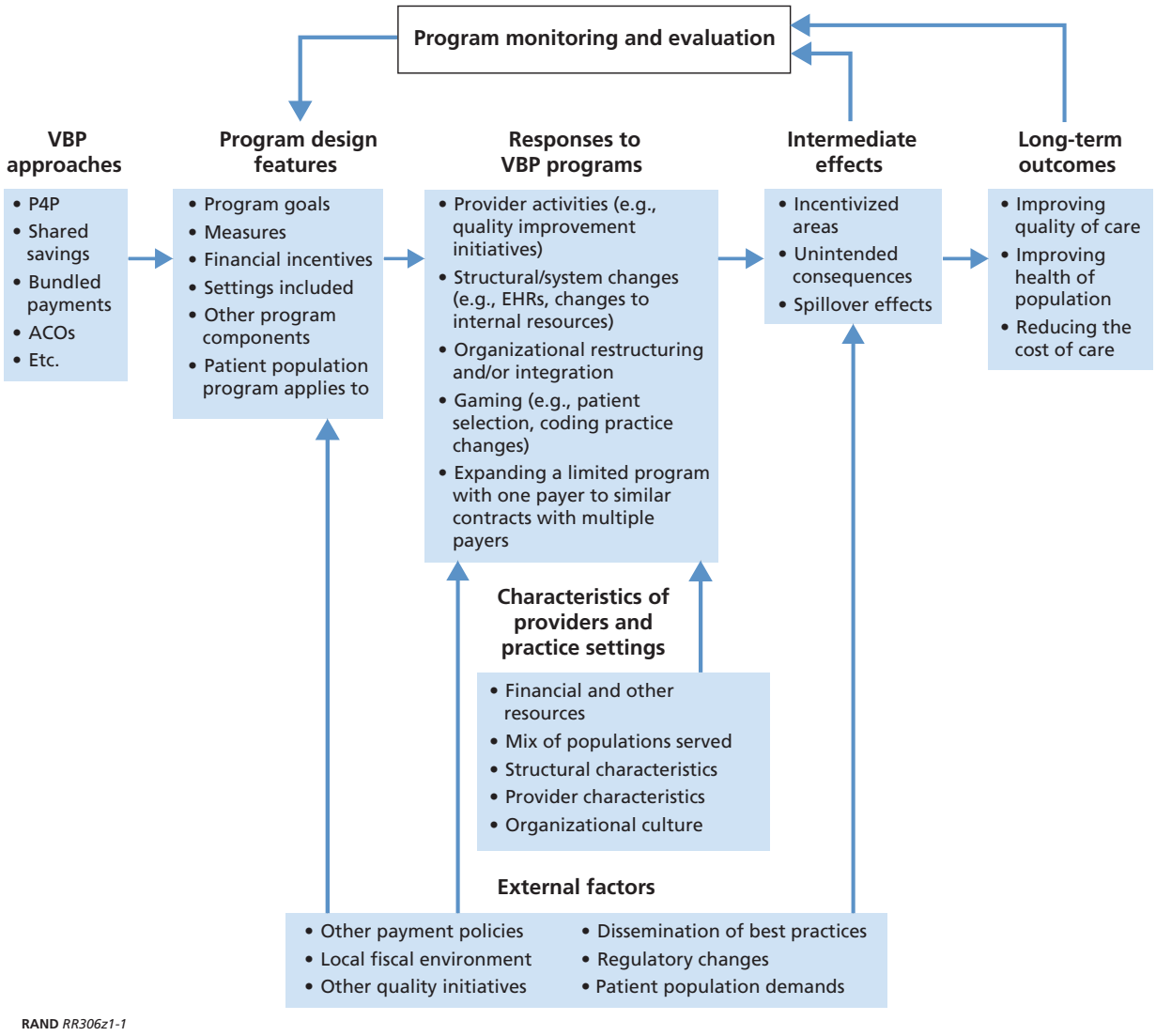
HHS is actively considering the federal government's near- and long-term strategy for how to design and implement VBP programs to achieve the three aims set forth in the National Quality Strategy,[115, 116] which focus on improving the overall quality of care, improving the health of the U.S. population, and making care affordable by reducing the cost of quality health care. In designing their VBP strategy, HHS seeks to apply the best available evidence to guide policymaking regarding the expansion of VBP across a range of Medicare program settings. To inform HHS's strategy development, RAND reviewed the published evidence and consulted with experts on whether VBP programs have been successful in meeting goals, to identify features of successful VBP programs, and to identify knowledge gaps to inform the focus of future evaluation and monitoring efforts.

## Conceptual Framework for Assessing the Effects of Value-Based Purchasing Programs

Several useful frameworks have been developed regarding how to evaluate VBP programs. These include a framework by Dudley et al. focused on assessing P4P experiments,[12] one by McHugh and Joshi to guide VBP program evaluation,[117] and another by Fisher et al. specifically focused on evaluating ACOs.[118] The three frameworks have similar elements. For example, the Dudley framework describes three core elements—the incentive, predisposing factors, and enabling factors—that influence or mediate the response to provider incentives for quality improvement, while the framework by McHugh and Joshi defines program design, the participants, and contextual factors as elements associated with impacts and implementation. The Fisher ACO evaluation framework similarly considers program components (i.e., ACO contract characteristics, implementation activities), provider characteristics (ACO structure and capabilities), and external factors (i.e., environmental context) that influence intermediate outcomes and impacts. Because ACOs represent a new model of VBP that is just starting to be tested by both public and private payers, Fisher and colleagues emphasize that both formative and summative research will be important for guiding policymaking. The McHugh and Joshi framework also emphasizes the importance of implementation (i.e., formative evaluation) research.

Figure 1.1 displays the framework we developed for this project, which was adapted from the three existing frameworks. For example, in our VBP framework, program design features, characteristics of providers and practice settings, and external factors correspond to the incentive, predisposing factors, and enabling factors described in the Dudley model. Characteristics of providers and practice settings and external factors are important contextual elements that influence the response of providers to the incentives. Our framework attempts to expand on the previously developed frameworks by detailing some of the specific factors that should be considered within each of the elements of the framework.

**Figure 1.1. Value-Based Purchasing Conceptual Framework**




RAND *RR306z1-1*

Because VBP programs are natural experiments and the associated research is observational in nature, Dudley (2005) underscores that it is critical that evaluators select theory-driven hypotheses about how incentives affect behavior so as to identify potential confounding factors that could explain observed effects.[13] A framework is a useful construct to help develop theory-driven hypotheses.

The conceptual framework offers a foundation for considering the design features of the incentive program, as well as other mediating factors that influence whether and how providers may respond to the incentives and whether programs are successful in reaching stated goals. We refer the reader to Appendix B, which lists a set of factors that our panel of technical experts (discussed below) recommended be systematically collected for all VBP programs to facilitate efforts to evaluate VBP programs and to compare and contrast observed impacts across

programs. The framework also can be used to guide discussions about the design and implementation of existing VBP programs and those in development and to define a structured agenda for monitoring and evaluating VBP programs, with the explicit goal of developing knowledge to improve the functioning of these programs.

## Methods and Research Questions

For this study, we defined VBP programs as private or public programs that link financial reimbursement to performance on measures of quality (i.e., structure, process, outcomes, access, and patient experience) and cost or resource use. Three broad categories of VBP models were the focus of our review: (1) P4P, which includes both "pay for quality" and "pay for quality and resource use, efficiency, or costs"; (2) shared savings models that typically, but not exclusively, are being deployed in the context of ACOs; and (3) bundled payments for episodes of care (only when paired with holding providers accountable for performance on quality measures). We excluded pay-for-reporting and demand-side programs (e.g., tiered networks and consumer incentives).

We define each of the three broad types of VBP models as follows:

- **Pay-for-performance** refers to a payment arrangement in which providers are rewarded (bonuses) or penalized (reductions in payments) based on meeting pre-established targets or benchmarks for measures of quality and/or efficiency. These financial incentives are intended to change provider behavior to achieve a set of objectives specified by the payer.
- **Accountable care organization** refers to a health care organization composed of doctors, hospitals, and other health care providers who voluntarily come together to provide coordinated care and agree to be held accountable for the overall costs and quality of care for an assigned population of patients. The ACO payment model ties provider reimbursements to performance on quality measures and reductions in the total cost of care. Under an ACO arrangement, providers in the ACO agree to take financial risk and are eligible for a share of the savings achieved through improved care delivery provided they achieve quality and spending targets negotiated between the ACO and the payer.
- **Bundled payments**[*] are a method in which payments to health care providers are based on the expected costs for a clinically defined episode or bundle of related health care services. The payment arrangement includes financial and quality performance accountability for the episode of care. Episodes can be defined in different ways, cover varying periods of time (e.g., one year for a chronic condition, the period of the hospital

[*] Other common terms used for bundled payment arrangements are *episode-based payment*, *episode payment*, *episode-of-care payment*, *case rate*, *evidence-based case rate*, *global bundled payment*, and *global payment*.

stay and 30 days post-discharge), and include single or multiple health care providers of different types (e.g., hospital only, hospital and ambulatory provider).[1, 104, 119]

Table 1.2 lists the research questions that ASPE asked RAND to address in its review. The questions address three broad areas of inquiry: (1) measuring the performance of VBP programs; (2) the results of performance in VBP programs; and (3) improving the performance of VBP programs.

**Table 1.2. Research Questions**

| |
|---|
| **Measuring Performance in Value-Based Purchasing Programs** |
| **1.** What goals should be set? (How should success be defined for VBP programs?) |
| **2.** What are the metrics by which VBP programs can and should be evaluated? |
| **3.** Which aspects of VBP are measurable and which are not? |
| **4.** What is the relationship between health outcomes and what is measured in VBP programs? |
| **Results of Performance in Value-Based Purchasing Programs** |
| **5.** Based on the metrics used to date, have VBP programs facilitated improvements in quality and value?<br>5a. What improvements in health outcomes attributable to VBP can we expect, and over what time horizon?<br>5b. What cost savings attributable to VBP can we expect, and over what time horizon? |
| **6.** Does performance on unmeasured aspects of quality of care suffer when providers focus on improving performance on what is being measured ("teaching to the test")? Conversely, are there "spillover effects" whereby quality improvement efforts improve care more broadly? |
| **7.** If a provider/institution performs highly on all the VBP metrics but has average performance on everything that is not measured, which proportion of total potential improvement in health will be achieved?<br>In other words, if we imagine that a high-performing health system produces "X" amount more "quality-adjusted life years" than an average-performing system, what fraction of that X would be produced by a health system that was higher-performing on metrics commonly included in VBP programs currently, but was average-performing in unmeasured areas? |
| **8.** How likely is it that improvements in our ability to measure what is important will change enough over the next five to ten years to significantly affect the answer to (7)? |
| **9.** Are there unexpected effects of VBP programs, including impacts on racial/ethnic and socioeconomic disparities, and access to care? |
| **10.** What are the features of the highest-performing providers/institutions and their adaptations to VBP? |
| **11.** What are the characteristics of the lowest-performing providers/institutions and their behaviors in response to VBP? |
| **12.** How much does it cost a provider/institution to improve on the measured performance areas?<br>12a. Are the incentive levels of VBP programs sufficient to cover the costs of investing in quality improvement?<br>12b. How do organizations weight these factors related to VBP and decide on quality improvement investments? |
| **Improving the Performance of Value-Based Purchasing Programs** |
| **13.** What are the critical gaps in knowledge about VBP, and how can these gaps be addressed? |
| **14.** What are the structural and implementation features of the most successful VBP programs? |
| **15.** Within VBP programs, how can practices from the highest-performing providers/institutions be disseminated? |
| **16.** To what extent can VBP programs that have a positive impact in health care be improved and expanded? |

We used three approaches to gather information to address the research questions:

- **Environmental scan of existing value-based purchasing programs:** We reviewed information that was publicly available for both publicly and privately sponsored VBP programs. We extracted information on program characteristics, program effects, and study designs when evaluations were available.

- **A review of the published evaluation literature on value-based purchasing:** We examined the peer-reviewed published literature for studies that specifically evaluated the impact of P4P, ACO, or bundled payment programs. We drew heavily from existing review articles where available.
- **Input from a technical expert panel:** Recognizing that many of the design issues and implementation lessons have not found their way into the published literature and likely never will, we convened a technical expert panel (TEP) to provide input on the study questions. The TEP was composed of VBP program sponsors (i.e., private plans and a regional multi-stakeholder collaborative), providers from health systems who have been the target of VBP programs, health services researchers with expertise in examining the effects of VBP programs, and federal participants who represented public payers. The TEP met twice in person, in May and June of 2013, for an all-day meeting facilitated by RAND staff. We provided the TEP with the findings from the environmental scan of programs and the literature review as background to inform the panelists' discussions.

At the start of the chapters that focus on the environmental scan and literature review, we detail the specific methods we used. For the literature review, we assessed the methodological quality of each study and the strength of the evidence as a whole for each research question that examined the impact or effect of VBP. Our approach followed the methodology outlined in the Agency for Healthcare Research and Quality's (AHRQ's) *Methods Reference Guide for Effectiveness and Comparative Effectiveness Reviews*.[120] We did not grade the evidence for those questions with descriptive results (e.g., goals used to assess success).

We rated the methodological quality of each study using three categories of rating (poor, fair, and good). This is a grade of the absolute strength of the evidence presented in each study. **Good** indicates a low risk of bias (i.e., the study has strong methods to guard against bias), **fair** indicates a medium risk of bias, and **poor** indicates a high risk of bias. In the evidence tables, we provide short explanation for each good and poor rating. Factors that contributed to our assessment of methodological quality included study design (randomization, matching), analytic techniques (attempts to control for confounding explanations), intervention characteristics (size of intervention and ability to generalize results more broadly), and conflict of interest/independence of the evaluator. Countries other than the United States have implemented VBP programs, particularly P4P programs, and we have included selected studies from non-U.S. programs in our review; however, the findings from these studies may not readily generalize to the United States because of differences in health system organization and structure, financing, and delivery, as well as substantial differences in the design of the VBP intervention.

We graded the strength of the evidence as a whole for each research question using four grade levels:

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

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

Key considerations that affect the classification according to this scheme are the risk of bias, consistency across studies, directness, precision, coherence, residual confounding, and strength of association. The three senior researchers on the team (Cheryl Damberg, Melony Sorbero, and Grant Martsolf) independently graded the collective evidence, discussed differences in ratings, and together generated the final rating. Because two of the VBP categories (ACOs and bundled payments) are quite new in their development and implementation, there is currently insufficient evidence on the effects of these VBP models.

## Organization of This Report

The remainder of the report addresses the findings from our three data collection approaches. Chapter Two focuses on the results of the VBP environmental scan, while Chapters Three through Five focus on the findings from the literature review for P4P (Chapter Three), ACOs (Chapter Four), and bundled payments (Chapter Five), and Chapter Six summarizes the key points of the TEP's discussion. We give our concluding thoughts in Chapter Seven.

# 2. Environmental Scan of Existing Value-Based Purchasing Programs

The purpose of the environmental scan of public and private VBP programs was to describe the current VBP landscape and provide information to address selected research questions. The review focused solely on publicly available documentation; within the scope of this contract, we were unable to conduct interviews with VBP program sponsors to gather additional information.

## Methods

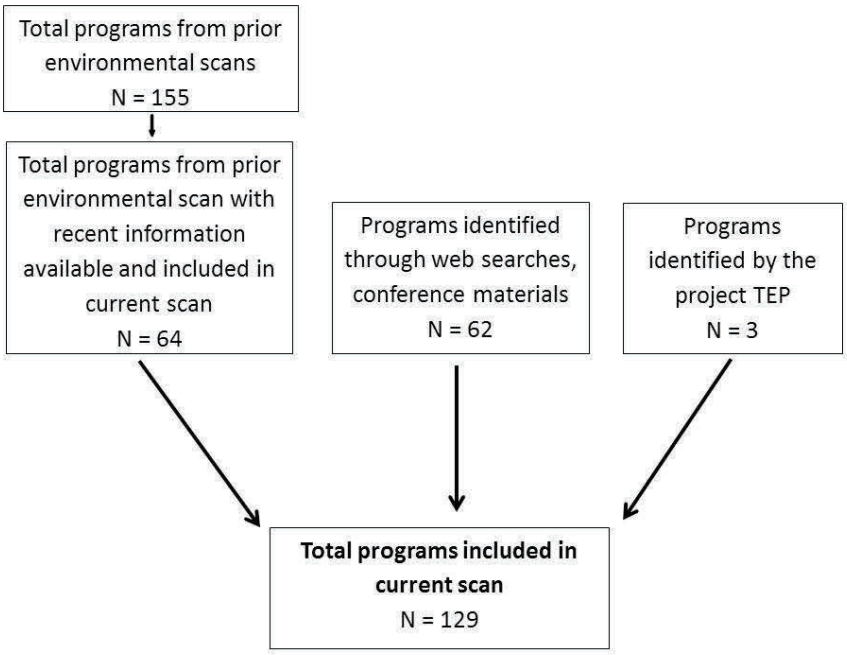
We compiled a list of VBP programs for review and stratified the list by the type of VBP model (i.e., P4P, ACO, bundled/episode-based payments). To develop the list, we began by drawing on lists of P4P program sponsors that were generated for prior physician and hospital P4P environmental scan projects conducted by RAND on behalf of ASPE.[19, 20] The list included programs sponsored by CMS, commercial health plans, regional multi-stakeholder coalitions, and Medicaid.* Additionally, we drew from a more recent RAND review of performance-based incentive programs.[121] Some of the programs identified in the earlier environmental scans were no longer in existence, had evolved into distinctly different programs, or had no recent information available. As a result, we winnowed the initial list of 155 programs down to those programs for which current information (2009 or later) could be found (n=64). We supplemented the list of 64 with an additional 62 programs that were newly added (e.g., ACOs, bundled payments) or where program sponsors had replaced prior programs. We identified the 62 programs from the following sources:

- materials from the 2012 and 2013 National P4P Summit sponsored by the Integrated Healthcare Association (IHA)
- the CMS website and press releases from CMS
- a Google search that focused on identifying bundled payment programs
- recent reports on P4P activities, including a 2010 report based on a survey of health plan P4P sponsors[122] and a report on VBP in skilled nursing.[123]

* Sources used to generate the list of programs included (1) a review of P4P programs by Rosenthal and colleagues (2004), (2) a 2004 Med-Vantage study of P4P programs by Baker and Carter, (3) a 2005 Med-Vantage survey of P4P programs, (4) the Leapfrog Compendium of incentive and reward programs, (5) review of the CMS website, (6) a Lexis/Nexis search of major U.S. newspapers, (7) a broad Google-based Internet search, (8) a search of relevant trade journals, and (9) input from RAND staff and TEP members.

During the introductory call with the TEP, members of the TEP identified three additional programs for inclusion. With these additions, we reviewed 129 VBP programs (Figure 2.1).

**Figure 2.1. Process Used to Identify Value-Based Purchasing Programs Included in Environmental Scan, Public Document Review**




For the 129 VBP programs, we gathered information by searching the program sponsor websites and conducting Google searches using the program sponsor or VBP program name. For each VBP program, we documented contextual information and key attributes of the program to the extent possible based on the contents of public documents. We extracted information on

- program goals
- types of metrics used
- type of incentives employed
- target of the incentives
- program effects
- type of support provided by sponsors
- changes anticipated for the program.

Only programs for which at least a portion of this information was publicly available were included in our scan. We compiled the data in Microsoft Excel for analysis. Appendix A contains a list of the programs included in our review.

## Findings from the Scan of Public Documents

The 129 VBP programs reviewed included 91 P4P programs, 27 ACOs, and 11 bundled payment programs. Nearly all of the program sponsors identified in RAND's earlier environmental scans continue to offer VBP programs, although many of the programs have since evolved to include new measures, new types of providers, and different forms of incentives. While we did not track the different health plan product lines covered by the programs, Med-Vantage reports an expansion in the percentage of health plans that include all covered lives in their VBP programs, from 11 percent of plans in 2008 to 55 percent of plans in 2010.[122] Many sponsors had expanded their portfolios to include multiple VBP program types, some of which involve new payment models, such as shared savings, and new delivery models, such as medical homes, ACOs, and bundled payment.

The availability of public information varied considerably across the design features and types of VBP programs. For example, few details are publicly available about the specific measures involved or the payment arrangements for many of the ACOs that have newly formed or that are in development. The programs we reviewed do not represent the universe of all VBP programs in current operation in the United States, and the documentation for some programs we reviewed was not complete; the results should be considered in light of these limitations. Our review was limited by the propriety nature of much of the information about these programs, which are frequently sponsored by private entities (e.g., commercial health plans).

## Program Sponsors

Fifty-six discrete entities sponsored the 129 VBP programs, and they fell into four categories of sponsors: (1) CMS, (2) private-sector commercial health plans, (3) regional collaboratives of stakeholders, and (4) states through their Medicaid programs (Table 2.1).

- **CMS programs:** The 16 identified CMS programs include completed demonstrations (e.g., Nursing Home VBP), current programs (e.g., Hospital Acquired Condition [HAC] Payment Policy, Hospital VBP programs), programs in the early stages of testing (e.g., ACO Shared Savings Program, Bundled Payment for Care Improvement) and a program still in the planning stages (Physician Value Based Payment Modifier).
- **Private health plan sector:** Private health plans are the most common sponsors of VBP programs overall, and many plans offer multiple VBP program types. For example, we report on six programs for Aetna (a physician P4P program, a program targeting physician groups in California conducted under the IHA's value-based P4P collaborative, a hospital P4P program, a Medicare Advantage ACO program, three separate commercial ACO agreements, and an agreement with Hoag Hospital in California as part of the IHA bundled payment initiative).
- **Regional collaboratives:** These groups sponsor seven P4P programs in our scan, but not other types of VBP models. Two are participants in the Bridges to Excellence program, and two have P4P programs that involve medical home pilots (i.e., the Oregon Health Leadership Council and the Puget Sound Health Alliance). The IHA sponsors two

programs, while the New York Department of Health is involved in two separate collaboratives, each with its own P4P program.

- **State Medicaid programs:** The 20 programs included consist of 10 skilled nursing facility initiatives, several ACO and coordinated care programs, a bundled payment program in Arkansas, five traditional physician or hospital P4P programs, and a P4P program in Vermont that included a medical home initiative.

**Table 2.1. Sponsors of Value-Based Purchasing Programs**

| | Number of Examined Programs | | |
|---|---|---|---|
| **Sponsor Types** | **Pay-for-Performance N=91** | **Shared Savings/ ACO N=27** | **Bundled Payment N=11** |
| CMS | 10 | 4 | 2 |
| Private health plans | 56 | 21 | 8 |
| Regional collaboratives | 8 | 0 | 0 |
| States/Medicaid programs | 17 | 2 | 1 |

NOTE: The "N" in each column refers to the number of programs for which we found publicly available information.

## Program Goals

We reviewed program documentation to inform research question #1 regarding what goals *should* be set for VBP programs. We were able to identify in public documents goals or objectives for approximately half (n=63) of the 129 VBP programs. With only a few exceptions, VBP sponsors stated goals at a high level (e.g., "improved health," "bend the cost curve") which are rarely quantifiable or easily measured to determine success or achievement. The exceptions were five programs that established quantifiable goals related to desired cost savings, two of which are new Medicaid programs:

- **Blue Cross Blue Shield of Massachusetts Alternative Quality Contract (AQC):** Reduce the medical expense trend of participating physician organizations by half over a five-year contract term.
- **Blue Shield of California California Public Employees' Retirement System (CalPERS) ACO:** Keep 2010 health care premium costs flat (zero growth in premium).
- **Cox Health Plan Episodes of Care Pilot:** Reduce potentially avoidable complications by 25 percent.
- **Colorado Medicaid Accountable Care Demonstration:** Reduce the annual increase in the cost of care by two percentage points.
- **Oregon Medicaid Coordinated Care:** A five percentage point reduction in emergency department utilization, hospital readmissions, and high-cost imaging to achieve overall savings to offset the $20 per member per month (PMPM) the program is investing.

Many VBP sponsors' goals encompass multiple dimensions or domains, such as "Improve the quality of health care delivery for Medicare beneficiaries while reducing program

expenditures." In Table 2.2, we summarize the frequency with which sponsor goals addressed specific goal domains.

**Table 2.2. Stated Goals of Value-Based Purchasing Programs**

| | Number of Examined Programs | | |
|---|---|---|---|
| **Goal Domains** | **P4P N=33 of 91** | **Shared Savings/ ACO N=21 of 27** | **Bundled Payment N=10 of 11** |
| Clinical quality | 25 (76%) | 15 (71%) | 8 (80%) |
| Cost/affordability | 15 (46%) | 13 (57%) | 6 (60%) |
| Patient outcomes | 8 (24%) | 8 (38%) | 6 (60%) |
| Coordination of care | 1 (3%) | 5 (24%) | 4 (40%) |
| Patient experience | 1 (3%) | 8 (38%) | 2 (20%) |
| Appropriate utilization | 3 (9%) | 1 (5%) | 2 (20%) |
| Collaboration | 3 (9%) | 1 (5%) | 0 |
| Safety | 5 (15%) | 0 | 0 |
| Infrastructure/health information technology | 3 (9%) | 1 (5%) | 0 |
| Access | 0 | 2 (10%) | 0 |
| Patient-centered care | 1 (3%) | 0 (0%) | 0 |
| Recognize/reward providers | 5 (15%) | 1 (5%) | 0 |

NOTE: The "N" in each column refers to the number of programs for which we found publicly available information.

**Pay-for-performance program goals:** Improving clinical quality (e.g., "evidence-based care," "meaningful quality improvement," "break through improvement") was included among the goals of three-quarters of the 33 programs for which we could obtain documentation. Half of the programs cited cost reduction/affordability goals. Cost goals were more common among newer P4P programs. Less commonly mentioned goals were patient outcomes, safety, patient experience, and recognizing and rewarding physicians.

**Shared savings/accountable care organization goals:** Similar to P4P programs, ACO sponsors emphasized clinical quality and cost as primary goals. Patient experience, patient outcomes, and coordination of care were less frequently cited.

**Bundled payment:** Program sponsors most frequently cited clinical quality as a goal, followed by cost and patient outcomes.

## Types of Providers Who Are the Target of Incentives

We identified the type(s) of providers that are the target of the financial incentives (Table 2.3) as well as the form of the financial incentive (Table 2.4).

### *Pay-for-Performance*

In our sample of programs, physician groups were most frequently the target of P4P incentives, though some programs include more than one type of provider, such as individual physicians and physician groups. One program, the Tufts Health Plan Coordinated Care Model, contracts with multiple provider types, including ACOs and integrated delivery systems.* Individual physicians, most commonly primary care physicians (PCPs), are the second-largest target of incentives. Med-Vantage reported that 98 percent of the commercial health plans responding to its survey had a PCP program either in operation or in development and 61 percent had specialist programs in operation (32 percent) or in development (29 percent). The report also indicated that the most common specialties included in the P4P programs were obstetrics and gynecology (OB/GYN), cardiology, orthopedics, and endocrinology. As in our sample, Med-Vantage reports a lower percentage of health plans with hospital P4P programs than physician P4P programs, and that portion (40 percent) was unchanged since Med-Vantage's 2008 survey. There are 19 hospital programs, 10 skilled nursing facility programs, and one nursing home program (i.e., the CMS nursing home P4P demonstration) in our sample of P4P programs.

### *Shared Savings/Accountable Care Organizations*

The ACO programs involve agreements between payer sponsors and health care providers, typically physician group practices with hospitals, wherein the providers assume financial and quality accountability for defined patient populations. CMS currently has three distinct ACO programs: (1) the Pioneer ACO Model; (2) the Medicare Shared Savings Program (MSSP); and (3) the Advanced Payment Initiative. Each of the three CMS ACO programs has different criteria for provider eligibility. For example, the Pioneer program is open to group practices, networks of individual practices, partnerships or joint ventures between hospitals and physicians, hospitals, or federally qualified health centers.[124] The MSSP is open to each of these entities plus critical access hospitals and rural clinics. Additionally, ACOs must agree to accept responsibility for at least 5,000 Medicare FFS beneficiaries to be eligible for the MSSP. The Advanced Payment Initiative is open only to ACOs that do not include any inpatient facilities and have less than $50 million in annual revenue or ACOs in which the only inpatient facilities are critical access

* An integrated delivery system (IDS) refers to an organization composed of a network of physicians and hospitals or physicians only which provide a continuum of health care services to patients who are enrolled in the system.

hospitals and/or Medicare low-volume rural hospitals with less than $80 million in annual revenue.[124] The commercial health plan ACOs typically involve integrated delivery systems, physician/hospital organizations, and medical groups. The Medicaid program in Oregon contracts with entities known as coordinated care organizations that are responsible for members' mental, physical, and dental care.

### *Bundled Payments*

The bundled payment initiatives in our sample address both chronic and acute episodes and therefore target multiple provider types. The CMS bundled payment initiative allows for participation and gain sharing by physician groups, hospitals, ACO-type providers, and post-acute providers.* Some bundled payment initiatives are targeting specialists who perform procedures, such as orthopedic surgeons for hip and knee replacement or cardiac surgeons for coronary artery bypass graft (CABG) surgery. In the CMS Acute Care Episode (ACE) Demonstration, the savings were shared by the participating providers and the Medicare beneficiaries who received care from the participating providers.

**Table 2.3. Health Care Provider Type(s) That Are the Target of Value-Based Purchasing Programs**

| | Number of Examined Programs | | |
|---|---|---|---|
| **Provider Type** | **P4P N=91** | **Shared Savings/ ACO N=27** | **Bundled Payment N=11** |
| Physician groups | 36 (40%) | 0 | 4 (36%) |
| Individual physicians | 33 (36%) | 0 | 4 (36%) |
| Hospitals | 19 (21%) | 0 | 3 (27%) |
| Skilled nursing facilities/nursing homes | 11 (12%) | 0 | 1 (9%) |
| ACO/integrated delivery system | 1* (2%) | 27 (100%) | 3 (27%) |
| Dialysis facilities | 1 (1%) | 0 | 0 |
| Health plans | 3 (3%) | 0 | 0 |

NOTE: The "N" in each column refers to the number of programs for which we found publicly available information. In addition, percentages do not sum to 100 because some programs include multiple provider types.

* Gain sharing refers to financial arrangements between the payer and the providers to share a portion of savings that are generated through reduced health care utilization or provision of less expensive care (e.g., use of generic drugs), typically only if the provider maintains a level of quality.

## Types of Incentives

The types of financial incentives offered to providers have expanded well beyond bonuses, the most common form of payment among P4P programs, to include new types and combinations of incentives. We were able to characterize the incentive structure for about 68 percent of the programs in our scan (see Table 2.4).

**Table 2.4. Types of Financial Incentives Used in Value-Based Purchasing Programs**

| | Number of Examined Programs | | |
|---|---|---|---|
| **Incentive Structure** | **P4P N=58 of 91** | **Shared Savings/ ACO N=23 of 27** | **Episodes of Care N=8 of 11** |
| Bonus | 35 (60%) | 0 | 1 (13%) |
| Change in fee schedule or diagnosis-related group (DRG) | 12* (21%) | 3 (13%) | 0 |
| Shared savings | 5 (9%) | 13 (56%) | 5 (62%) |
| Shared savings and shared risk | 1 (2%) | 6 (26%) | 0 |
| Bonus and shared savings | 5 (9%) | 0 | 0 |
| Bonus and shared savings/shared risk | 0 | 1 (4%) | |
| Episode fee adjusted for quality | 0 | 0 | 2 (25%) |

NOTE: The "N" in each column refers to the number of programs for which we found publicly available information.
*Includes the CMS HAC Payment Policy, which prevents payment for selected hospital-acquired conditions at the higher DRG rate, and the CMS Hospital Readmission Reduction Program, which adjusts the DRG payment rate downward.

### *Pay-for-Performance*

P4P programs have historically used bonuses as the type of incentive; however, shared savings incentives have become increasingly common, particularly related to performance on spending and utilization measures. Some health plans, such as Tufts, offer different types of shared savings incentive structures based on the provider's level of experience with managing risk. Most of the health plans participating in the IHA P4P program paid bonuses for performance on clinical quality, health information technology, and patient experience measures and offered shared savings based on performance on a set of resource use measures (e.g., generic prescribing, readmissions). The newly emerging IHA Value-Based P4P program will reward physicians organization performance on a total cost of care measure as a basis for shared savings, with the amount modified up or down by the physicians organization's performance on a composite measure of quality.

### *Shared Savings/Accountable Care Organizations*

Most of the ACOs in our environmental scan sample have shared savings arrangements, and a few have shared risk. For example, the CMS MSSP and Pioneer Models offer either a one-sided or a two-sided approach. Additionally, in the Pioneer Model, ACOs that have shown savings over the first two years are eligible to move to capitation in year 3. The Blue Cross Blue Shield of Massachusetts AQC allows for shared savings and shared risk, and offers a bonus up to 10 percent above the global budget based on performance on quality measures.

### *Bundled Payment Programs*

Among the bundled payment programs for which we have information, offering shared savings to providers was most common, including the CMS ACE demonstration and the Bundled Payments for Care Improvement initiative. Two programs adjust the episode fee for quality. In the United Healthcare Oncology Episodes of Care pilot, any future increases in the episode fee require the practices to achieve improved outcomes, a reduction in the total cost of care, or both. The Geisinger ProvenCare for CABG initiative tied adherence to the ProvenCare process measures to surgeons' individual compensation.

## Measures

We were able to catalog information on the performance measure domains for approximately 92 percent of the VBP programs in our scan, but detailed information on the exact measures program sponsors used was available for a minority of programs. Therefore, we summarized the measures being utilized for the different types of programs, by setting, at the domain level.

### *Pay-for-Performance*

Among the ambulatory-setting P4P programs for which we found measure information (n=57), clinical quality (i.e., process-of-care and intermediate outcome measures) was the most commonly measured domain. VBP programs typically use the National Committee for Quality Assurance (NCQA) Healthcare Effectiveness Data and Information Set (HEDIS) preventive and chronic care measures. However, for VBP programs that use HEDIS chronic care measures, we found it difficult to determine from their public documentation whether they were measuring intermediate outcome measures (e.g., glycolated hemoglobin [HbA1c]/blood sugar control or blood pressure control), process measures (e.g., testing HbA1c levels or blood pressure), or both. Structural measures were the next most common measurement domain, typically addressing the adoption or use of health information technology or rewarding physicians and/or groups for obtaining NCQA certification for the chronic care or patient-centered medical home (PCMH) programs. Roughly, half of the ambulatory P4P programs used patient experience measures, and nearly half are measuring cost and hospital/ emergency department utilization. Where access is measured, P4P programs typically used the ambulatory Consumer Assessment of Healthcare

Providers and Systems survey, although some of programs involving PCMHs were also measuring same-day appointment scheduling and the availability of other forms of access such as email and phone visits.

The hospital P4P programs where we found some measure information (n=17) typically used the CMS/Joint Commission clinical process measures. Several used measures of readmission, mortality, and patient safety (e.g., Leapfrog or the AHRQ patient safety indicators) in their programs. Patient experience, typically measured by the Hospital Consumer Assessment of Healthcare Providers and Systems survey, is included in slightly more than half of the hospital P4P programs. A small number rewarded hospitals for participation in quality improvement (QI) initiatives.

Information about the measures utilized by 10 Medicaid skilled nursing facility P4P programs was documented in a 2011 report from the National Research Corporation.[123] The most common types of measures reported were staff levels, training, and retention. Eight of the 10 programs measured customer satisfaction and regulatory compliance. The next most common measures were clinical care, employee satisfaction, and culture change/ person-centered care, each measured by five of the 10 programs. In the cost/resource use domain, one program is measuring operating costs and one is measuring Medicare utilization. The report also notes a trend toward incenting a culture of person-centered care. The CMS Nursing Home VBP program included measures of staffing and turnover, hospital readmissions, and outcome measures from the Minimum Data Set.

### *Shared Savings/Accountable Care Organizations*

We were able to identify some of the measures used by 23 of the 27 ACO programs included in our scan. Each of the CMS programs is using the same set of 33 measures, which includes HEDIS clinical process (preventive and chronic care) and intermediate outcome measures, Consumer Assessment of Healthcare Providers and Systems survey results about patient experience, all-cause hospital readmissions, ambulatory sensitive care hospital admissions, patient safety measures (e.g., screening for fall risk, medication reconciliation), and a measure of electronic health record (EHR) functionality. CMS is phasing in the measures over three years, with the first year as pay-for-reporting only. Additionally, each of the three CMS ACO programs is measuring cost as a basis to determine shared savings.

The two ACOs in our scan participating in the Brookings-Dartmouth Accountable Care Initiative reported using a common set of measures that are being phased in over time. The first set of measures the ACOs implemented were claims-based measures. These included four measures of overuse (appropriate imaging studies for low back pain, avoidance of antibiotic treatment for adults with acute bronchitis, etc.), seven population health measures (breast cancer screening, HbA1c blood sugar testing, use of appropriate medications for asthma, persistence of beta blocker treatment after a heart attack, etc.), one safety measure (annual monitoring for patients on persistent medications), all-cause 30-day readmissions, and eight utilization measures

(hospital days per 1,000, emergency room visits per 1,000, use of generic drugs, doctor visit within seven days of discharge, imaging rates, etc.). The next set of measures, implemented in early 2012, included 11 clinically enriched measures for coronary artery disease, diabetes, hypertension, pediatric immunizations, and colorectal cancer screening. The third phase adds patient-reported measures, including patient experience (2012) and patient-reported outcomes (2015).

The Blue Cross Blue Shield of Massachusetts AQC includes 32 ambulatory measures and 32 hospital measures. The ambulatory measures include HEDIS clinical process and intermediate outcome measures and eight adult and pediatric patient experience (Consumer Assessment of Healthcare Providers and Systems survey) measures. The hospital measures include process measures for acute myocardial infarction (AMI), congestive heart failure (CHF), pneumonia, and surgical care plus eight AHRQ Patient Safety Indicators and four Hospital Consumer Assessment of Healthcare Providers and Systems survey measures.

### *Bundled Payment Programs*

We were able to determine a minimal level of measure information for nine of the 11 bundled payment programs in our scan. The programs are targeting a diverse set of conditions, and the most common measure domain is cost. In the hospital setting, the CMS ACE Demonstration utilized a broad set of clinical process, outcome, and patient safety measures for each of the six procedures. Process measures were largely drawn from the Surgical Care Improvement Project (SCIP), and outcomes included readmissions, inpatient and 30-day mortality, and average and median length of stay (LOS). Conversely, the Geisinger ProvenCare program for CABG used a large set of clinical process measures and avoided tying physician compensation to outcome measures so that physicians would not hesitate to treat patients that are more complicated. The Blue Cross Blue Shield of Tennessee Orthopedic Bundled Payment Program ties reimbursement to performance on quality and efficiency measures, and the Horizon Blue Cross Blue Shield of New Jersey program for hip and knee replacement measures patient functional status, readmissions, and patient safety. In the IHA bundled payment pilots for hip and knee replacement, plans and providers determine the measures in contrast to the IHA P4P programs in which common quality measures are used and reported across the plans. IHA expects that gain-sharing agreements will include both quality and efficiency measures but does not provide a menu of options.

Little information was publicly available regarding measures used in ambulatory care bundled payment programs. The Arkansas Medicaid program targets six conditions, and provider gain-sharing is dependent on achievement of "must pass" quality indicators, which differ for each episode type. The United Healthcare Oncology Episodes of Care pilot ties future increases in the episode fee to improved outcomes, reduction in the total cost of care, or both.

## Benchmarks

Benchmarks refer to the performance threshold the provider must meet (either absolute or relative) to achieve the incentive payment. Information about the types of benchmarks used was available for only 34 percent (n= 44) of the programs in our scan. We found no information about the benchmarks used for the bundled payment programs.

### *Pay-for-Performance*

The most common type of benchmark among the P4P programs is an absolute threshold only (n=15). Ten P4P programs in our scan used relative thresholds only, which may be based on the performance of peers in the market, the state, or nationally. Other programs, such as the CMS Hospital VBP program, have two paths to earning incentives: achieving an absolute threshold or showing improvement over time (11 of the 39 P4P programs had this combination).

### *Shared Savings/Accountable Care Organizations*

Very little information was publicly available about the types of benchmarks used for ACO models, with the exception of the three CMS ACO models. In the shared savings programs, CMS is establishing the cost benchmark for each agreement period, for each ACO, using three-years-prior expenditure data. Quality benchmarks are based on national percentile rankings from the year prior, and points are assigned on a sliding scale based on the ACO's performance. The Pioneer ACO program originally had absolute benchmarks to encourage very high performance. Participating ACOs have expressed concern that the standards are higher than those that "best-in-class" providers have achieved to date and will be costly to meet. In response, CMS will measure and reward improvement on the quality metrics for 2013. The CMS PGP demonstration utilized absolute thresholds for quality measures.

**Table 2.5. Type of Benchmarks Used in Value-Based Purchasing Programs**

| | Number of Examined Programs | | |
|---|---|---|---|
| **Benchmark** | **P4P N=39 of 91** | **Shared Savings/ ACO N=6of 27** | **Episodes of Care N=0 of 11** |
| Absolute threshold | 15 (38%) | 5* (83%) | Not available in public documents |
| Relative threshold | 10 (26%) | 0 | Not available in public documents |
| Absolute threshold and improvement | 11 (28%) | 1 (17%) | Not available in public documents |
| Relative threshold and improvement | 3 (8%) | 0 | Not available in public documents |

NOTE: The "N" in each column refers to the number of programs for which we found information.
*For the three CMS ACO models, CMS assigns points for each quality measure on a sliding scale.

# 3. Review of the Pay-for-Performance Literature

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

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

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

## Methods

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

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

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

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

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

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

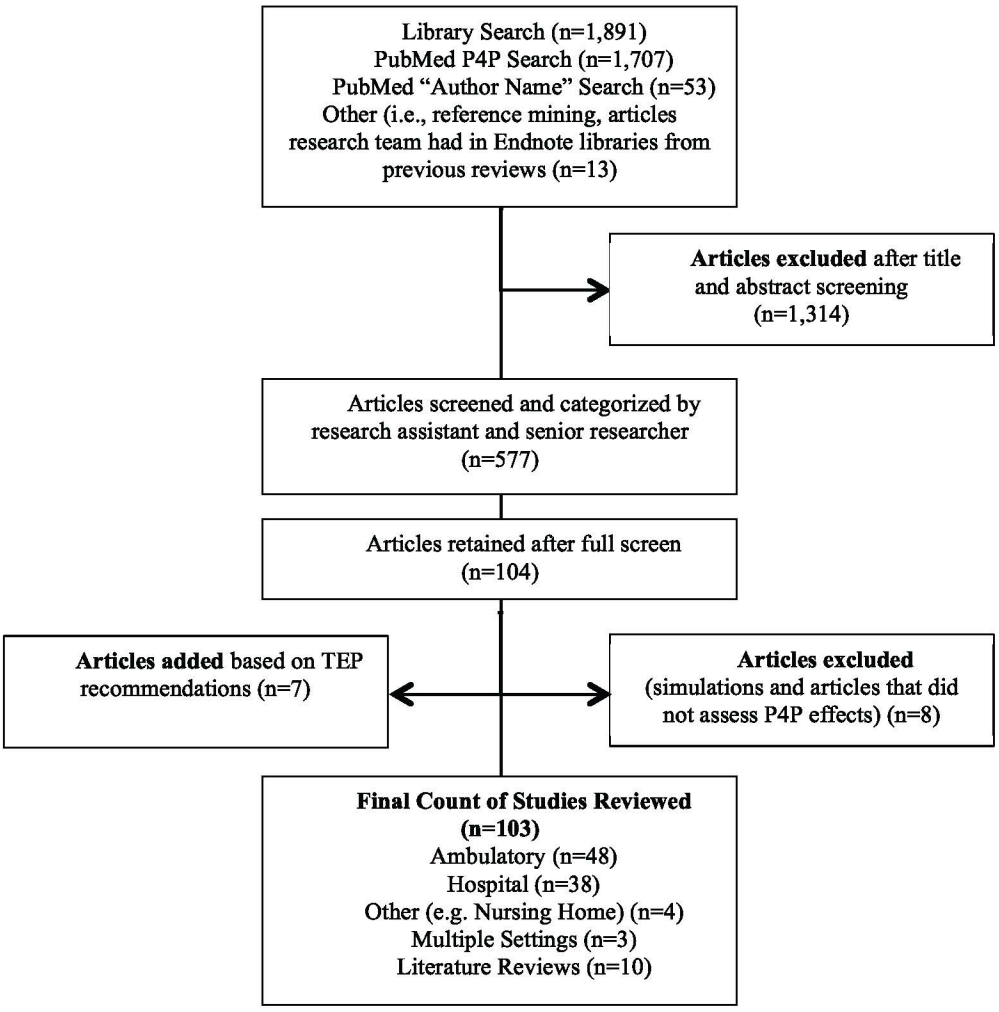
The research assistant on the team (Laura Raaen ) conducted an initial screening of titles and abstracts for relevance and content. If there was indecision about whether or not an article was relevant, it was included. A senior researcher (Cheryl Damberg) on the team reviewed the final set of titles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

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

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

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

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




## Research Questions

### *Measuring Performance in Value-Based Purchasing Programs*

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

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

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

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

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

**Strength of Evidence: Not applicable, descriptive only.**

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

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

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

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

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

*Hospital Measures*

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

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

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

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

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

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

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

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

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

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

*Ambulatory Measures*

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

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

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

*Nursing Home Measures*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## *Results of Performance in Value-Based Purchasing Programs*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

#### *Pay-for-Performance Programs Focused on Hospitals*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

*P4P Programs in Other Settings*

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

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

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

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

*P4P Programs Focused on Physicians or Physician Groups*

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

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

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

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

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

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

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

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

We rated the other three studies as fair or poor:

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

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

*P4P Programs Focused on Hospitals*

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

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

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

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

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

*P4P Programs Focused on Other Settings*

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

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

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

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

*Conclusion*

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

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

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

*P4P Programs Focused on Physicians or Physician Groups*

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

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

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

Two studies were of good methodological quality:

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

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

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

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

#### *P4P Programs Focused on Hospital Groups*

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

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

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

#### *Conclusion*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Overall, the studies show small to no unintended effects.

*Unintended Effects*

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

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

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

#### *Spillover Effects*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**Table 3.9. Factors Associated with Performance on Incentivized Measures**

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**Strength of Evidence: Not applicable, descriptive only.**

### *Improving the Performance of Value-Based Purchasing Programs*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**Strength of Evidence: Insufficient.**

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

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

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

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

# 4. Review of the Accountable Care Organization Literature

ACOs represent the latest performance-based payment innovation designed to reduce health care spending and improve care delivery. ACOs combine much stronger financial incentives in the form of shared savings and shared risk models among the ACO players (i.e., physician group, hospital, and health plan) than many previous VBP efforts. ACOs are focused on improving care coordination and increasing the sense of responsibility that providers feel to improve patient care regardless of where it occurs.

The most common ACO arrangement includes both quality performance benchmarks and spending targets. If the quality benchmarks are met or exceeded, then the ACO is eligible to receive a portion of the "savings" generated by health spending that is below the projected spending targets. Some ACO arrangements also include a shared-risk component wherein the ACO is at risk for absorbing at least a portion of costs if actual spending is in excess of the spending target.

Although ACOs are still in their nascent stage of development, their growth has been rapid, with both public and private payers testing ACO models. CMS has embarked on testing three ACO models: (1) MSSP (n=220), (2) Pioneer ACOs (n=32), and (3) Advance Payment ACOs (n=35). There are few articles on ACOs in the literature that were published prior to 2010, and they tend to focus on outlining the rationale for moving to an ACO-type model of care delivery and payment and the general framework of these models. By the end of September 2013, there were at least 493 organizations that had an ACO contract.[118, 166]

Because of the rapid experimentation occurring around ACOs, relatively little is known about what features, organizational structure, or contract design are most important to guarantee their long-term success. Much of the existing ACO literature focuses on case studies of design, development and implementation of individual ACOs or the design features or implementation experiences across a group of ACOs. The few studies that report results generally report how performance improved compared with projected trends rather than comparing results against a control group of providers and/or patients, and the results represent the experience over one to two years of implementation. It is difficult to know whether the results are generalizable or sustainable, or whether they are the result of the ACO or broader changes occurring in the health care environment. Consequently, we caution readers regarding the interpretation of these studies.

## Methods

We reviewed the nascent literature on ACOs and shared savings arrangements, first searching the PubMed and WorldCat databases for literature published from January 1, 2007, through November 14, 2012. We used multiple search terms (Table 4.1) to identify relevant titles. We

included only articles published in English. We supplemented the results of this search with reference mining, cross-referencing of relevant articles, and ad hoc Google Scholar searches. For programs evaluated by articles through this search, we also included articles published after November 2012. In addition to the literature search, the TEP recommended a few additional studies for inclusion. Although there is prior experience and a literature related to captiation, an aspect of some ACO arrangements, we did not consider that previous work as applicable to the current ACO structures; as such, we excluded it from our review. We catalogued the search results in Endnote Software.

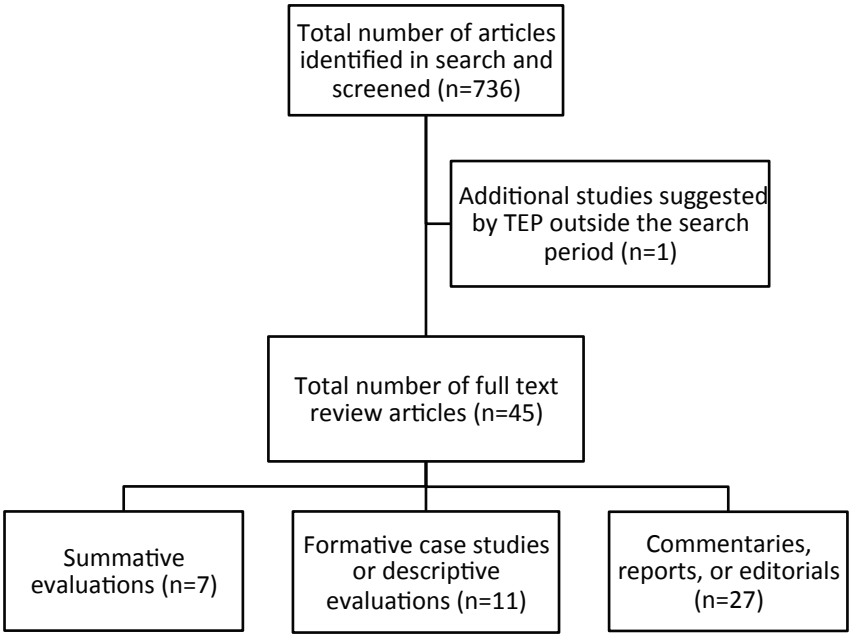
**Table 4.1. Search Terms Used in Accountable Care Organization Literature Review**

| Search Terms | Search Engine | Search Dates |
|---|---|---|
| (accountable care organization* OR ACO OR ACOS) AND (Quality[tiab] OR quality improvement OR quality indicators, health care OR "quality of care" OR "quality of health care") | PubMed | January 1, 2007–November 6, 2012 |
| (share$ adj3 savings).mp. [mp=title, abstract, original title, name of substance word, subject heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] | Medline on OVID | January 1, 2007–November 6, 2012 |
| ((kw: accountable and kw: care and kw: organization* OR kw: aco OR kw: acos) ) or ((kw: shared and kw: saving*)) AND (kw: health* OR kw: medical OR kw: patient* OR kw: physician* OR kw: doctor* OR kw: hospital* OR kw: nurs*) | WorldCat | January 1, 2007–November 6, 2012 |
| (accountable adj2 care adj2 organization$).mp. or ACO or ACOS NOT (gene or genetic$).mp. OR algorithm$.mp. or Algorithms/ | Medline on OVID | January 1, 2007–November 14, 2012 |

Titles and abstracts were screened for relevance and content by a research assistant and a senior researcher. If there was indecision about whether or not an article was relevant, it was included. Abstracts and titles were then reviewed by a senior researcher who identified the relevant articles for full-text review. U.S. program evaluations, case studies, and empirical studies with quantitative or qualitative data on shared savings payment components were reviewed in full text. To avoid accumulation of articles pertaining to the conceptual or theoretical components of shared savings payment models, we excluded commentaries and editorials from full-text review when abstracts did not mention learning from existing ACOs.

Our review of the shared savings payment literature resulted in a total of 737 articles after removal of nonrelevant duplicates (Figure 4.1). A total of 45 articles went through full-text review. Of the 45 articles we reviewed, only seven were actual program evaluations of ACOs or other shared savings with six of these examining the effect on performance measures included in the program and one assessing whether there were spillover effects. The remainder were commentaries or formative evaluations assessing the implementation experiences of ACOs. We synthesized the evidence qualitatively to address the research questions that were the focus of this study.

**Figure 4.1. Process Used to Identify Articles for Review, Accountable Care Organizations**




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

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

## Research Questions

### *Measuring Performance in Value-Based Purchasing Programs*

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

Our review of public documents for 20 ACOs found that clinical quality (n=14 of 20 programs) and cost/affordability (n=12 of 20 programs) were most commonly cited as the goal. Less frequently mentioned were patient experience (n=8 of 20), patient outcomes (n=7 of 20), and care coordination (n=5 of 20). We could only find public documentation on two of the ACOs regarding explicit measurable goals. The Blue Shield of California California Public Employee Retirement System (CalPERS) ACO sought to have zero growth in premiums for 2010. The Colorado Medicaid Accountable Care Demonstration sought to reduce the annual increase in the cost of care by two percentage points. We recognize that all ACOs are setting explicit targets with payers as part of their negotiations, but this information is confidential and not available for our review.

In the published studies we reviewed, the goals that were the focus of the ACO included a combination of reducing costs, improving quality, improving access and patient experience, and aligning incentives across payers, providers, and patients. Multiple ACO case studies and commentaries stated that shared savings models should strive to achieve systematized goals in an effort to align incentives across payers and providers.

**Strength of Evidence: Not applicable, descriptive only.**

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

The literature review found no information on this question. Please refer to Chapter Six for the summary of the TEP's discussion.

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

The literature review found no information on this question. Please refer to Chapter Six for the summary of the TEP's discussion.

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

The literature review identified no evaluations that formally assessed the relationship between health outcomes and shared savings arrangements within the context of ACOs. The measures used by many ACOs mirror those in P4P programs in the ambulatory setting and we refer the reader to the relevant section of Chapter Three.

**Strength of Evidence**: **Insufficient.**

*Results of Performance in Value-Based Purchasing Programs*

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

We identified six evaluations examining the effect of implementing an ACO or ACO-like model on quality or health care costs (Table 4.2). All used quasi-experimental designs and included a mix of simple pre-post implementation comparisons with multivariate difference in difference models using some form of control group. Two of the evaluations focus on the PGP demonstration that ran between 2005 and 2010, which CMS identified as a precursor to ACOs. The 2009 Report to Congress on the PGP demonstration[167] reported results for the first two years of the demonstration. Subsequent press releases by CMS highlight overall demonstration results, but we were unable to identify a final evaluation report. The article by Colla and colleagues[62] further examined the effect of PGP demonstration on health care costs over four years of the demonstration, overall for all beneficiaries and for dually eligible beneficiaries. Markovich[63] reported the effects during the first two years of Blue Shield of California CalPERS Sacramento Pilot ACO Program launched January 2010. The ACO represents a partnership between the Dignity Health hospital system, Hill Physicians Medical Group, and the Blue Shield HMO for 41,000 CalPERS members. Salmon and colleagues[64] report 2010 results from three practices participating in the Cigna Collaborative Accountable Care initiative located in New Hampshire, Texas, and Arizona. Two studies focused on the AQC developed and implemented by Blue Cross Blue Shield of Massachusetts.[65, 66, 87] This program began in 2009 and included seven provider groups in 2009, with four added in 2010.

The incentive structure of the ACO programs differed somewhat. The PGP demonstration set targeted risk-adjusted expenditures based on the rate of growth of costs in the local community. In years 1 and 2 of the demonstration, participating physician groups could earn up to 80 percent of the savings they generated that were greater than two percentage points lower than their target. If per capita spending was less than two percentage points lower than the target, the savings were retained by CMS. In years 3 through 5 of the demonstration, half of the savings could be earned only if the physician group achieved national benchmarks or demonstrated improvements on selected quality measures as well. These measures focused on management of diabetes, CHF, coronary artery disease, and hypertension, as well as cancer and hypertension screening measures. The CalPERS pilot also had a shared savings component based on zero growth in health care costs in the first year. Partners in the ACO shared savings if health care costs were below the targeted amount. Providers also shared risk with the health plan wherein expenses in excess of the target were equally absorbed by the partners. The CIGNA model incorporates a care coordination fee that is paid to practices at the beginning of the year. In the first year, it is based on the activities planned to improve care or reduce cost. At year's end, if the trend in a practice's total medical cost has improved at least two percentage points relative to a comparison group and quality has also improved, the fee is increased for the next year. The AQC model is

based on a unique design wherein provider groups are eligible to receive a portion of the difference between their global budget and total medical spending. They are also eligible for an additional bonus of up to 10 percent PMPM for meeting standards on a set of 64 quality measures. Blue Cross Blue Shield of Massachusetts also provides technical assistance including the provision of quality and cost data to the 11 AQC provider groups.

In addition to the above evaluations, CMS issued a press release on the early experiences of the Medicare Pioneer ACO on July 16, 2013.[67] In the first performance year, the ACO performed better overall than the Medicare FFS comparison population on the 15 measures for which comparable data are published. The extent to which performance improved among the Pioneer ACOs was not reported in this press release.

**Strength of Evidence: Not applicable, descriptive only.**

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

Pre-post analyses indicated that only two of the physician groups in the PGP demonstration achieved benchmark performance on all 10 measures used in year 1 of the demonstration. By the end of year 2, however, all 10 participating PGPs achieved benchmark performance on at least 25 of the 27 quality indicators (10 diabetes measures, 10 CHF measures, and seven coronary artery disease measures) and five achieved benchmark performance on all 27 measures. On average, physician groups increased performance on diabetes mellitus measures by nine percentage points, CHF measures by 11 percentage points, and coronary artery disease measures by five percentage points. However, much of this increase was likely due to national trends in improvement. When PGP demonstration performance was compared with local controls for seven measures that could be generated with claims data, significant improvement by year 2 was observed for only four of the measures. Colla and colleagues also found that spending reductions in the PGP demonstration did not appear to be associated with lower performance on quality measures, as reflected by performance on measures of readmission and emergency department visits. The Sacramento pilot ACO did not publish data on quality measures other than hospital readmissions, which declined during the first two years of the ACO pilot. Early results from the CIGNA initiative showed that all three practices scored higher than their comparison groups on all five process-of-care measures except the New Hampshire practice's HbA1c screening measure. The New Hampshire and Texas practices had small improvements in quality between 2009 and 2010 (0.6 percent and 0.7 percent), while the Arizona practice had a slight decline (–0.3 percent) but the differences were not significant. For each program, at least some changes in quality were observed within the first two years of implementation. None of the evaluations reported on patient outcomes.

Song et al.[65] found that the Blue Cross Blue Shield of Massachusetts AQC model contributed to improvements in the proportion of eligible enrollees receiving recommended care for chronic disease care (2.6 percent) and pediatric care (0.7 percent) during the first year of the program.

However, the study observed no effect on adult preventative care measures during the first year. Song et al.[66] found that these results persisted (and even grew) through year 2 of the program. The AQC model contributed to improvements in the proportion of eligible enrollees receiving recommended care for chronic care management (3.7 percent), pediatric care (2.3 percent), and adult preventive care (0.4 percent) across the two intervention years.

**Strength of Evidence: Insufficient to Low.** While the number of patients covered by these three programs is substantial, the study findings represent only 14 communities and three different financial arrangements. A very small number of quality measures were assessed relative to a control group and these showed inconsistent improvements. Further research is likely to change our estimate of the effects.

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

All four program evaluations reviewed from the literature attribute various degrees of costs savings to the shared savings payment model. The second year of the BSCA-CalPERS ACO resulted in an estimated $37 million in savings and a compound annual growth rate of PMPM fees of ~3 percent after the first two years of the program. The pilot also saw a substantial decrease in utilization, as average inpatient length of stay decreased 15 percent in year 1 of the program and 12.1 percent in year 2. Still, it is important to note that this pilot utilized a global budget approach for a group of enrollees with a predetermined provider network and is not directly generalizable to the broader Medicare population covered under a FFS payment model. Furthermore, the article does not include information regarding which subsets of CalPERS enrollees (e.g., high-risk groups) were most influential in driving down costs.

The PGP demonstration evaluation did not show conclusive costs savings in the first two years that PGPs were eligible to receive shared savings payments. Only four PGPs achieved total costs savings that exceeded the 98 percent costs savings target, while four PGPs were within the 98–102 percent cost target and did not achieve cost savings, and two PGPs experienced increased costs of approximately $2.2 million. Across the 10 PGPs, the program achieved a reduction of actual expenditures of $120 per beneficiary, which was 1.2 percent lower than the target expenditure rate ($p<.01$) by year 2 of the demonstration. However, Colla and colleagues found, between 2005 and 2009 of the PGP demonstration, $532 in average annual savings per beneficiary in the dually eligible population ($p<.001$), while non-dually-eligible beneficiaries in this same period saw only $59 in average annual savings per beneficiary ($p<.28$). Furthermore, when Colla et al. adjusted dual-eligibles' cost savings by program year, they found that costs savings were achieved primarily in the first few years of the program and later years saw more rapid increase in spending compared with controls. Furthermore, according to results posted on the CMS website, only two of the 10 group practices earned savings in all five years of the demonstration, and three of the group practices did not achieve savings in any of the five years. The other five group practices were inconsistent in their ability to generate savings, with some earning savings in the middle years of the evaluation but not in the first or last year.

The Salmon Cigna ACO study reported that total medical costs for the Arizona practice were $27.04 PMPM more favorable than costs in its comparison group ($p<0.10$) and that, compared with expected costs, the New Hampshire and Texas practices had modest improvements in PMPM costs of $1.78 and $6.56, but results were not significant. Both studies were of poor methodological quality.

Song et al.[65] found that provider groups in the AQC experienced smaller increases in PMPQ total medical spending compared with control groups, with a magnitude of $15.51 PMPQ (1.9 percent) during the first year of the program. These savings were derived primarily from referring patients to lower cost facilities. Song et al.[66] found that these results persisted and even grew across the first two years of the program. Practices in the AQC experienced smaller increases in PMPQ total medical spending compared with the control groups, generating an estimated savings of $22.58 PMPQ (the equivalent of 2.8 percent) during the first two years of the program.

CMS[67] reported that the costs for the Pioneer ACO beneficiaries increased 0.3 percent in 2012, compared with 0.8 percent growth for similar Medicare FFS beneficiaries. Thirteen of the 32 ACOs shared savings with CMS producing an estimated $33 million in savings for Medicare. Two Pioneer ACOs had shared losses, two Pioneer ACOs were leaving the ACO program, and an additional seven were switching to the MSSP ACO model, which involved less risk to providers.

**Strength of Evidence: Insufficient.**

**Table 4.2. Evidence on Effectiveness of Accountable Care Organization Value-Based Purchasing Programs**

| Reference | Program Description | Study Design | Incentive Structure | Measures Examined | Findings | Assessment of Methodological Quality |
|---|---|---|---|---|---|---|
| Sebelius, 2009[167] | Physicians Group Practice Demonstration (PGP demo) included 10 physician groups ranging in size from 232 to 1291 physicians.<br>(Apr 2005–Mar 2010)<br>The number of beneficiaries assigned to the PGPs in PY2 ranged from 9,715 to 38,743 and the average number of beneficiaries per site was 21,958. | Quasi-experimental pre-post design with difference in difference analyses using with local comparison groups for 7 claims based measures and expenditures. Compared baseline year (2004) to program years 1 and 2. Data for later years not available at time report was published. | Physician groups were eligible to receive up to 80% of savings generated in excess of a 2% savings threshold conditional on improved performance on 27 quality measures in year 2 focused on diabetes (10 measures), CHF (10 measures), coronary artery disease (7 measures) and 32 measures in years 3–5 (added 3 hypertension and 2 cancer screening measures). | **Cost:** Total and per person expenditures compared with target; inpatient expenditures; outpatient expenditures<br>**Utilization:** E&M visits<br>**Process:** Quality measures included in program | **1.** The number of E&M visits per year was stable (5.4 per beneficiary in baseline year, 5.5 per beneficiary in year 2);<br>**2.** 2 practices achieved benchmark performance on all 10 measures in year 1.<br>**3.** 10 group practices reached benchmark performance on 25 of 27 quality indicators in year 2; 5 reached benchmark performance on all 27 measures in year 2.<br>**4.** Between baseline year and year 2, group practices increased their performance on diabetes measures by 9 percentage points, CHF measures by 11 percentage points and coronary artery disease by 5 percentage points.<br>**5.** Between baseline and year 2, PGP physician groups showed greater improvements in performance than local controls on 4 of 7 claims-based measures.<br>**6.** 4 PGPs had combined savings in PY1 and PY2 of $ 26,907,000 (less than 98% of target); 4 PGPs achieved neither savings not had increased costs (between 98–102% of target). 2 PGPs had negative savings of $3,484,000 (greater than target of 102%).<br>**7.** Net Savings to Medicare Trust Funds for years 1 and 2: $2,260,000 (estimated total expenditure savings minus performance payments).<br>**8.** Actual expenditure: $120 per person less or 1.2% less than Target Expenditures per beneficiary for the combined 10 PGPs in PY2 ($p< .01$).<br>**9.** On average, outpatient expenditures were $83 per person year less than expected, while inpatient expenditures were $25 per person year less than expected and not statistically significant. | **Good:** solid study design, but small number of provider organizations and unclear how generalizable to broader group of ACOs |

| Reference | Program Description | Study Design | Incentive Structure | Measures Examined | Findings | Assessment of Methodological Quality |
|---|---|---|---|---|---|---|
| Colla et al., 2012[62] | Physicians Group Practice Demonstration (PGP demo) included 10 participating physician groups. See above for greater detail | Quasi-experimental, pre-post design with difference in difference analyses for costs, 30-day readmissions, and emergency department use between 2001 and 2004 (pre period) and 2005–2009 (post period). The study compared 990,177 Medicare beneficiaries receiving care from physician groups participating in PGP demo to 7,514, 453 control Medicare beneficiaries. Analyses adjusted for beneficiary age; sex; race; federal disability and Medicaid eligibility status; race-specific zip code level measures income; Clinical risk adjustment using low-variation conditions | See above | **Cost:** Total annual Medicare payments per person; major categories of costs (e.g., acute care hospital, skilled nursing, professional services);<br>**Utilization:** physician services within Berenson-Eggers Type of Service (BETOS) categories;; emergency department use<br>**Outcomes**: probability of 30-day readmissions | **1.** Overall, adjusted average annual Medicare payments per beneficiary in PGP demo participating sites increased by $114 than among control beneficiaries (95% CI, $12–$216, P=.03)<br>**2.** Average annual savings were significant among the dually eligible beneficiaries ($532, 95% CI, $277–$786, p= .001) but not the non-dually eligible beneficiaries ($59, 95% CI, $166 in savings to $47 in additional spending, p=.28).<br>**3.** Only 4 sites saved a significant amount across all beneficiaries, while 3 sites had no significant change and 3 sites increased expenditures relative to controls during the PGP demo. Only two of the ten groups exhibited savings under the Hierarchical condition category (HCC) risk adjustment approach; both had relatively large increases in HCC scores relative to their control group<br>**4.** No observed reductions in Medicare skilled nursing spending in the dually eligible.<br>**5.** No overall association between the PGP demo and the probability of emergency department visits either in the full PGP demo population or among dually eligible beneficiaries. These averages, however, mask significant reductions (no p-value given) in emergency department visits in the sites that produced the largest savings in dually eligible beneficiaries.<br>**6.** The PGP demo was associated with lower medical 30-day readmissions on average across the 10 sites (no p-value given)<br>**7.** lower readmissions for both medical and surgical admissions in the dually eligible beneficiaries (no p-value given) | **Good:** solid study design, but small number of provider organizations and unclear how generalizable to broader group of ACOs**:** |

| Reference | Program Description | Study Design | Incentive Structure | Measures Examined | Findings | Assessment of Methodological Quality |
|---|---|---|---|---|---|---|
| Markovich, 2012[63] | Blue Shield of California Sacramento Pilot ACO Program launched January 2010. Focus on shared risk and shared savings between Dignity Health hospital system, Hill Physicians Medical Group, and the HMO for 41,000 California Public Employees' Retirement System (CalPERS) members. | Pre-post comparing 2009 to 2010 and 2011. Difference in difference comparing member costs to CalPERS member in Northern California not in ACO pilot. Details of analytic approach not reported. | Goal is a target cost per patient per month. If costs exceed target, partners share in excess expense. If expenses below target partners share in savings. Target was 0% growth in health care costs in first year. | **Cost:** Per member per month costs; estimated total savings;<br>Utilization: inpatient days per thousand members; hospital stays 20 days or greater; average LOS<br>**Outcomes**: 30 day readmission rates | First year:<br>**1.** ACO patient costs decrease 1.6%, while non-ACO patient costs increase 9.9%<br>**2.** Projected $15.5 million in savings<br>**3.** Estimated half of savings from reduced utilization; half from slowing increases in reimbursement rates<br>**4.** Inpatient days per 1,000 members reduced by 15% compared with 5.9% reduction in controls<br>**5.** ACO population reduced hospital 30 day readmissions 15%<br>**6.** ACO population reduced extended hospital stays (≥ 20 days) by 50% (reversed in year 2)<br>**7.** Greater reductions in average LOS than other areas in Northern California<br>**8.** Increased in emergency department utilization among ACO members<br>Combined first and second year:<br>**1.** Projected $37 million savings for ACO patients compared with non-ACO patients<br>**2.** 30-day readmission rate declined to 4.1% in 2011 (year 2) from 4.3% in 2010<br>**3.** Average length of stay increased by 0.21 days from year 1 results, but remained lower than CalPERS members not in ACO | **Poor:** Case study of a single ACO. inadequate description of analytic methods; unable to determine comparability of comparison group. |

| Reference | Program Description | Study Design | Incentive Structure | Measures Examined | Findings | Assessment of Methodological Quality |
|---|---|---|---|---|---|---|
| Salmon et al., 2012[64] | Launched in 2008 and now implemented in 42 practices, Cigna Collaborative Accountable Care initiative is a shared savings program to improve the quality and efficiency of care in open access benefit plans. | Quasi-experimental pre-post design with difference in difference analyses using local comparison groups for 3 practices in Arizona, New Hampshire and Texas | A care coordination fee is paid to practices at the beginning of the year. In the first year, it is based on the activities planned to improve care or reduce cost. At year's end, if the trend in a practice's total medical cost has improved at least 2% relative to a comparison group and quality has also improved, the fee is increased for the next year. | **Cost:** Absolute total medical costs (risk-adjusted PMPM), improvement in total medical costs<br>**Process:** 5 process of care measures | 2010 results:<br>Total medical costs for the Arizona practice were $27.04 PMPM more favorable than costs in its comparison group ($p<0.10$)<br>Compared with expected costs, the New Hampshire and Texas practices had modest improvements in PMPM costs $1.78 and $6.56, but results were not significant.<br>The NH practice improved 0.6% over 2009 and was 0.7 % better than its comparison group for the 5 quality measures (81.1 % compliance rate).<br>All 3 practices scored higher than their comparison groups on all 5 quality measures except the NH practice's HbA1c screening measure. NH and Texas practices had small improvements in between 2009 and 2010 (0.6% and 0.7%) while the AZ practice had a slight decline (–0.3%) but the differences were not significant. | **Fair:** focused on a limited number of practices in three geographic areas. |

| Reference | Program Description | Study Design | Incentive Structure | Measures Examined | Findings | Assessment of Methodological Quality |
|---|---|---|---|---|---|---|
| Song, Safran et al., 2011[65] | In 2009, Blue Cross Blue Shield of Massachusetts implemented a global payment system called the AQC. Provider groups assume responsibility for spending and groups were eligible for quality-based bonus payments. | Quasi-experimental pre-post design with difference in difference analyses using local comparison groups for seven provider groups one year after program implementation | Provider groups eligible to receive a portion of the difference between global budget and actual spending, as well as an additional 10% of their global budget in bonus for meeting a set of 64 quality measures. | **Cost:** Total medical costs (PMPQ)<br>**Process:** 7 chronic care measures, 5 adult preventive care measures, 6 pediatric care measures, and composite measures for each category | **Cost:** Increase in PMPQ was smaller for intervention group compared with control group at a magnitude of $15.51 PMPQ (1.9%). Savings derived primarily from moving patients to lower cost facilities.<br>Process: The proportion of eligible enrollees who received recommended care increased faster in intervention groups compared with controls for chronic care management (2.6%), pediatric care (0.7%,). No effect on adult preventative Care quality measures. | **Good:** Regional, but strong study design |
| Song, Safran, et al., 2012[66] | In 2009, Blue Cross Blue Shield of Massachusetts implemented a global payment system called the AQC. Provider groups assume responsibility for spending and groups were eligible for quality-based bonus payments. | Quasi-experimental pre-post design with difference in difference analyses using local comparison groups for seven provider groups starting the program in 2009 and four provider groups starting the program in 2010 | Provider groups eligible to receive a portion of the difference between global budget and actual spending, as well as an additional 10% of their global budget in bonus for meeting a set of 64 quality measures. | **Cost:** Total medical costs (PMPQ)<br>**Process:** 7 chronic care measures, 5 adult preventive care measures, 6 pediatric care measures, and composite measures for each category | **Cost:** Increase in PMPQ was smaller for intervention group compared with control group at a magnitude of $22.58 (2.8%); $15.51 (1.9%) in year 1 and $26.72 (3.3%) in year 2.<br>**Process:** The proportion of eligible enrollees who received recommended care increased faster in intervention groups compared with controls for chronic care management (3.7%), pediatric care (2.3%), and adult preventive care (0.4%). All measures improved in year 2 while adult preventive care did not improve in year 1. | **Good:** Regional, but strong study design |

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

We found one recently published study of good quality (quasi-experimental comparisons from 2007–2010), by McWilliams and colleagues,[87] which assessed spillover effects under an ACO-type arrangement. The authors of this study compared FFS Medicare beneficiaries served by the 11 provider organizations in the AQC with beneficiaries served by other providers (control group), and used a difference-in-differences approach to estimate changes in spending. This study found spillover effects to the Medicare population among providers exposed to the Blue Cross Blue Shield of Massachusetts AQC. After two years, statistically significant costs savings were observed among Medicare beneficiaries treated by AQC providers compared with beneficiaries in the control group, for a 3.4 percent savings relative to an expected quarterly mean of $2,895. Consistent with the findings for the Blue Cross Blue Shield of Massachusetts commercial enrollees, the cost savings were greater for patients with more medical conditions (≥5 conditions) and savings accrued largely from lower spending on outpatient care (procedures, tests, and imaging). The study did not find spillover effects on quality performance for five process measures and two outcomes (30-day readmissions and potentially avoidable hospitalizations).

**Strength of Evidence: Insufficient.** There are no studies at this point in time documenting unintended consequences related to the implementation of ACOs and shared savings arrangements. There is only one study that examined spillover effects, and while positive effects were found related to costs, it is unclear whether the results from this single study would generalize to the experiences of other ACOs.

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

**Strength of Evidence: Insufficient.** No studies have examined this topic because ACO models are new and in the process of being tested.

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

**Strength of Evidence: Insufficient.** No studies have examined this topic because ACO models are new and in the process of being tested.

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

We identified no studies describing the impact of ACOs on racial/ethnic disparities. However, in their evaluation of the PGP demonstration, Colla and colleagues[62] focused on dually eligible Medicare beneficiaries out of concern that physician practices could try to reduce Medicare-paid hospital or skilled nursing facility days in order to meet expenditure targets, thus shifting a higher proportion of health care costs from Medicare, which covers acute care services for the dually eligible, to Medicaid, which covers Medicare premiums, cost sharing, and long-term (custodial) nursing home services. However, the authors did not find any reductions in the amount of spending for Medicare skilled nursing or an increase in the number of dually eligible institutionalizations and thus could not conclude that PGPs were shifting costs over to Medicaid. The PGP demonstration evaluation showed that excluding Indirect Medical Education (IME) and DSH payments from costs savings calculations could influence costs and performance payments. However, effects on costs and payments varied across physician groups, and no conclusive effect on disparities was observed.

Colla and colleagues reported that emergency department use decreased among dual-eligibles at the group practices that generated the greatest savings, which could suggest improved access to or improved quality of care. In addition, the PGP demonstration Report to Congress[167] reported the number of evaluation and management (E&M) visits per person per year was stable in the first two years of the demonstration indicating that access did not decrease. The Sacramento Pilot ACO witnessed an unexplained increase in emergency department utilization during the two years of the program, which could suggest reduced access to care, but no additional data were reported to determine this.[63]

Some case studies and commentaries identified consolidation of market power in ACOs as a possible barrier to improved access to care.[168–170] Commentaries speculate that ACO arrangements in academic medical centers may fail to function because of cultural organizational factors that limit the involvement of providers in these types of payment models. Such cultural factors include an institutional emphasis on autonomous departments; tenure systems that focus on support, publications, and scholarly reputation; external grant funding that may mitigate the need for these institutions to look for costs savings; the predominance of part-time physician staff due to research; and the influence of junior trainees that lack experience in efficient medical practices.[171, 172] However, others have found that the ACO concept is easily adapted to the research- and primary-care-dominated environment of certain academic medical centers.[173] Rural hospitals, many of which receive disproportionate share payments, also face barriers to implementing ACO arrangements due to low patient volumes and decentralized physician practices.[169] Since academic medical centers and rural hospitals are vital in supporting vulnerable populations' access to care, it will be important to track how these systems fair in ACO arrangements.

**Strength of Evidence: Insufficient.**

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

Only the PGP demonstration evaluation highlighted organizational characteristics that were potentially influential in achieving standards for the four top-performing physician group practices (i.e., practices that exceed their costs savings targets and thus were rewarded with performance payments, though only two of these remained top performers across all five years of the demonstration). Top performers were characterized as being affiliated with an academic medical center or a freestanding physician group practice. Furthermore, the academic medical center physician practices had integrated hospitals.

In the study of the CIGNA ACO model, leaders from the three participating practices reported that the initial care coordination fee and the patient-specific reporting from CIGNA were critical factors for improvements in costs and quality. Additionally, all three practices had some experience with either P4P or capitation prior to participating in this initiative.

The predominant recommendation from the case studies and commentaries was centered on strong physician leadership and physician representation in the governance structure. Case studies found that physician leadership with a clear strategy and vision was necessary to change practice culture to one that is comfortable with sharing the risk of a predetermined patient population.

**Strength of Evidence: Insufficient.**

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

The PGP demonstration analysis indicated that the lowest-performing physician groups were all affiliated with a non-academic hospital and did not receive performance payments. The report speculated that hospitals would not be able to maintain inpatient revenue while simultaneously reducing avoidable admissions or using lower cost care practices and thus consistently failed to meet performance targets.

**Strength of Evidence: Insufficient.**

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

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

None of the ACO evaluations included in our literature review addressed whether the program incentive levels of VBP programs were sufficient to cover the costs of care. However, multiple case studies of nascent ACOs examined the costs of investing in quality improvement and provide insight on the appropriate levels of incentives.

Audet and colleagues[168] found that of the hospitals responding to the Health Research and Educational Trust's National Survey of Hospital Readiness for Population-Based Accountable Care in September of 2011, only 49.7 percent of respondents believed they had the necessary financial strength to take on risks. Seventy percent of respondents had processes in place to continuously monitor the use of services and costs compared with revenue. Responses from hospitals already participating in an ACO model (n=47) showed that reducing costs was the second greatest barrier to implementation of the ACO model, behind reducing clinical variation. In a case study of eight ACOs, health plans commented that incentives for quality improvement alone were not sufficient to reward provider groups that are already top performers, and that additional incentives should be implemented that focus on attainment of high standards of performance.[174] Bailit and Hughes[175] conducted interviews with 32 payer and provider organizations incorporating shared savings arrangements at various organizational levels and found that payers were concerned that small provider organizations would not be able to cover the costs of investing in quality improvement with the given incentive levels. Harris Meyer[176] heard similar remarks during an interview with a relatively small MSSP ACO (35 small physician groups with 16,000 assigned beneficiaries) that is in the early stages of development. Smaller ACOs may lack the financial backing needed to employ care managers or implement a sophisticated EHR infrastructure. After performing case studies of four organizations trying out shared savings arrangements, three of which included a hospital partner, Moore and Coddington[100] found that all four of the organizations agreed that incentives (1) need to be significant enough to change behavior early, (2) need to evolve with time as improvements in data and reporting become available, and (3) need to be catered internally to different organizational units (e.g., individual physicians, group/business unit incentives, inter-organizational incentives).

**Strength of Evidence: Insufficient.**

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

None of studies we reviewed described how organizations weight factors related to shared savings arrangements and how they decide on QI investments.

**Strength of Evidence: Insufficient.**

### *Improving the Performance of Value-Based Purchasing Programs*

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

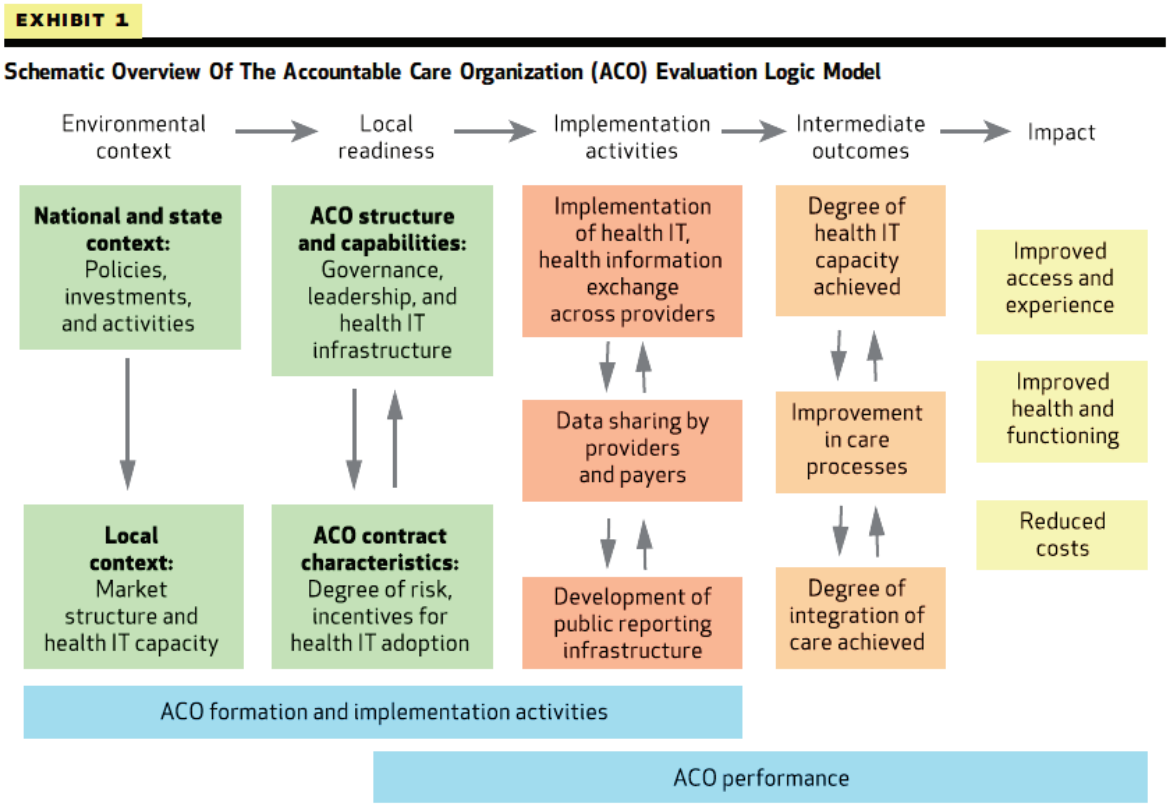
Since the shared savings payment model and ACOs are relatively new in comparison to other VBP models, there remain many critical gaps in knowledge. There are many challenges related to the design of ACOs and their contractual arrangement with payers, including whether a minimum level of integration is required for ACO formation,[177] whether the potential benefits

from ACOs offset the potential negative effect of too much consolidation and market power in a specific market, whether particular governance models are more conducive to successful ACO creation and implementation,[167] how specialists are best incorporated into ACOs,[176] and whether there are regulatory and legal barriers to ACO formation and how this varies by state.[178] There are also simple foundational questions, such as "What distinguishes an ACO from a primary PCMH, and are these distinctions meaningful?"[118, 170] Formative evaluation is critical to understand issues that must be addressed to facilitate the successful implementation of ACOs and shared savings programs.

Additionally, there is little published information on the effect of ACOs on health care costs and quality of care. Future studies should assess what characteristics and features of ACOs (such as care management programs and redesigned care processes) are necessary for cost savings and quality improvement to occur.[62, 167] They should also explore the extent to which ACO savings are influenced by preexisting expenditure trends or the financial incentives of a given program;[167] the optimum size of an ACO to maximize organizational performance or efficiencies of scope or scale;[99] and how differences in case-mix adjustment methods affect ACO performance and score calculations.[62] Future studies should also address the effect of ACOs on key outcomes. There is also a need to understand the extent to which unintended effects or spillover effects are occurring.[167] Multiple authors have noted that ACOs will need to know how many of their patients are receiving care outside of the ACO network, and how often. More information on how ACOs perform with respect to these patients and those receiving care only from the ACO will also be critical for ACOs' QI programs.[118, 176, 179]

To help consider how ACOs and shared savings models should be evaluated, Fisher et al.[118] developed a framework and logic model regarding ACO formation, implementation, and performance. Based on the experiences from the Brookings-Dartmouth ACO Learning Network pilot sites, feedback from national experts and ACO participants, and observations of other efforts to evaluate delivery system reforms, they identified five main domains for evaluation (environmental context, local readiness, implementation activities, intermediate outcomes, and impact) that fall within two overlapping stages for evaluation: (1) ACO formation and implementation activities and (2) ACO performance. Evaluations of ACO formation and implementation activities would focus on the environmental context, local readiness, and implementation activities, while evaluations of ACO performance might start with assessment of local readiness and span implementation activities, intermediate outcomes, and actual impacts. Figure 4.2 displays the features identified by Fisher and colleagues within five domains. They note that distinguishing between ACO formation/implementation and ACO performance may be difficult because ACO performance could improve in areas where data collection is not feasible, such as organizational anticipation of and preparation for future ACO contracts.

**Figure 4.2. Elements That Should Be Addressed in Evaluations of Accountable Care Organizations, as Identified by Fischer et al., 2012**




SOURCE: Fisher et al., 2012

Other researchers have developed frameworks focusing on more narrow aspects of ACO evaluation. Kroch and colleagues[180] attempted to objectively evaluate ACO readiness using a distinct framework and a scoring methodology based on six core capability components (people-centered foundation, health home, high-value network, payer-partnership, population health data management, and ACO leadership) that measures 154 operating activities necessary for ACO implementation. They scored these operating activities for their applicability to ACOs using a 0–4 Likert scale, and then weighted them by relevance to ACO formation. The resulting score reflects the degree to which the population targeted for accountable care had the potential to be affected by the ACO operations. The researchers used this framework to evaluate 59 organizations across the country that are part of the Premier Partnership for Care Transformation Readiness Collaborative (organizations considering ACO contracts) with various hospital arrangements between August 2010 and June 2011. A forthcoming study from The Commonwealth Fund will use these operating activities to evaluate organizations in the Premier Partnership for Care Transformation Implementation Collaborative.

Relatedly, Audet and colleagues[168] administered the Health Research and Educational Trust's 2012 National Survey of Hospital Readiness for Population-Based Accountable Care to 1,672

hospitals. The survey assessed contract characteristics (e.g., hospital-payer partnership and payment arrangements), capabilities (e.g., medication reconciliation as part of an established care plan, shared clinical information across settings of care, identification of patients who transition between settings of care), and ACO context (e.g., multihospital health system, urban status, teaching status). Another study of four nascent ACO organizations by Kreindler and colleagues[177] used interviews to gather information for measuring the impact of social-identity management strategies on the outcomes of ACOs (reducing costs and improving quality). The authors commented that structural measurement that assesses the levels of integration, physician- versus hospital-led governance, and evolution of social identity within ACOs will require mixed-methods evaluations.

Multiple case studies of existing ACOs have looked at various other formative components including shared savings payment arrangements and patient attribution, risk adjustment, and benchmarking methodology.[100, 151, 174, 175, 179, 181] Information from these kinds of surveys and case studies can provide a picture of the status of ACO formation and implementation nationally and regionally, and provide information that can be used to develop structural measures similar to those recommended by Fisher and colleagues.

Because the shared savings payment model and ACOs are relatively new in comparison to other VBP models, there remain many critical gaps in knowledge. Our review of the literature identified several areas where information about ACOs is missing. Table 4.3 highlights approaches identified in the literature to fill these gaps.

**Table 4.3. Approaches to Fill Information Gaps Identified in the Accountable Care Organization Literature**

| Areas Identified for Future Research and Evaluation | Approaches to Fill Information Gaps |
|---|---|
| Contextual factors | • Conduct research on differences in contractual arrangements between ACO payers and providers.<br>• Examine whether a minimum level of integration is required for ACO formation.<br>• Evaluate particular governance structures for their impact on ACO creation and performance.<br>• Examine best practices among high- and low-performing ACOs (e.g., care management programs).<br>• Assess the optimum size of an ACO to maximize organizational performance or efficiencies of scope or scale.<br>• Evaluate the impact of patients who seek care outside the ACO network on ACO performance. |

| Areas Identified for Future Research and Evaluation | Approaches to Fill Information Gaps |
|---|---|
| Program design | • Evaluate the differences in performance between ACOs participating in programs with different incentive designs (advanced payment model versus Pioneer ACO).<br>• Assess the distinctions between PCMH and ACO program designs and their impact on successful formation and achievement of goals.<br>• Examine the regulatory, legal, and financial barriers to successful formation and implementation of ACO and other shared savings programs.<br>• Determine the extent to which measures are aligned between public and private shared savings programs.<br>• Evaluate the impact of benchmarking methodology (minimum achievement versus improvement) on performance across ACOs. |
| Measures | • Analyze the effectiveness of individual measures and types of measures within shared savings programs (clinical quality measures versus patient experience and outcomes measures),<br>• Examine performance on indicators across ACOs,<br>• Determine which of the measures contributed most to the increase in savings,<br>• Determine whether the current set of measures in use by the Medicare ACO programs is optimal for driving desired behavior changes and achievement of goals. |
| Disparities | • Examine whether diminished variation in quality of care will lead to reduction in inequalities.<br>• Assess the reach and feasibility of implementation of ACO programs among minority and disadvantaged populations. |
| Outcomes | • Assess the extent to which ACOs improve quality and reduce costs, including which areas improvements are observed in.<br>• Explore whether the magnitude of shared savings is influenced by preexisting expenditure trends or program financial incentives. |
| Unintended/spillover effects | • Monitor the ACO programs for effects on supply-side market consolidation.<br>• Examine whether there changes to unmeasured areas (either positive spillover effects or negative undesired effects).<br>• Examine what types of care areas FFS-model ACO members seek care for outside the ACO, as well as the frequency with which ACO members obtain care outside the ACO. |
| Other gaps | • Identify driving force(s) of the improvement or lack of improvement across incentivized measures.<br>• Assess provider satisfaction under the ACO program. |

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

ACOs are still very new, and there are insufficient numbers with performance evaluations to identify variation in performance and the structural and implementation features associated with performance. Further, we do not have adequate variation in the shared savings models to identify aspects of the contractual arrangements that likely lead to success.

Case studies of nascent ACOs highlighted successful structural and implementation features that were in use or deemed necessary for success by ACO leaders. Introduction of care management programs, including care navigators and disease-specific case management, was the primary implementation feature highlighted in these articles. More systematic structural

recommendations included flexibility in program designs and savings arrangements, alignment and coordination of the ACO incentives with similar payment models such as the PCMH, timely and trended data updates for providers multiple times during the performance year, clarification of antitrust monitoring and evaluation at both the state and federal level, ACO accreditation process, and encouragement of multi-payer contracts.

Commentaries also provided useful recommendations for improving and expanding the ACO model. Shields and colleagues[101] suggest that physician performance payments be made based on performance at multiple levels of the organization (e.g., physician-level and organizational-level set of metrics). They also argue for physician leadership in all governance bodies and physician hospital boards. Others have argued for more flexible and long-term (three to five year) ACO arrangements and contracts to encourage long-term commitments to the ACO model.[182, 183] Davis and Schoenbaum[184] note that many of the lessons learned from the failure of the managed care boom of the 1990's can be applied to the broader structural components of ACOs. The resistance to the managed care model demonstrated that payment models and arrangements should align across the broader health system in order to avoid excessive administrative burden, that patients preferred getting care from smaller practices that allowed them to see the same physicians over time and that public backlash from patients who did not feel they were receiving proper care or services was a significant barrier. Guterman et al.[185] have suggested several structural features that are related to the lessons learned from managed care that can lead to improving and expanding the effects of ACOs. Some of them include:

- ACOs guarantee patients have access to a constant source of coordinated and primary care
- Shared savings payments made from payers to ACO providers account for total savings and not just savings for individual payers
- ACO patients and beneficiaries are informed and educated about which of their providers are part of an ACO
- ACOs make an explicit commitment to serving their community including disadvantaged and underinsured patients
- Criteria are developed that identify requirements for entry and continuation of ACO participation that is contingent on performance and accountability of care (e.g., public reporting) and not structural characteristics.

**Strength of Evidence: Insufficient.**

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

Issues in implementing ACO infrastructure can be addressed through technical assistance programs. According to Shortell and Casalino,[102] loosely organized small practices, independent practice associations, and physician-hospital organizations (PHOs) will be in most need of technical assistance because they face the largest financial and organizational barriers to implementing EHR systems, which are vital for providing coordinated care. Technical assistance

can be provided by the private sector and the Medicare quality improvement organizations for two distinct ACO components. The first is the development of organizational, financial, legal, and budgeting capabilities necessary to comply with performance reporting and new payment arrangements. The second component includes the activities ACOs undertake to improve quality and performance (e.g., implementation of EHR, promotion of physician leadership, practice redesign, etc.). A Commonwealth Fund Report[185] suggests that CMS work with payers to create information exchanges and standardized reports that provide timely and actionable feedback while HHS provides technical assistance for EHR implementation through the Office of the National Coordinator for Health Information Technology (ONC). Others have pointed out that CMS's online training seminars and expert learning collaboratives (e.g., the Brookings-Dartmouth ACO Learning Network, the NCQA ACO Accreditation platform, the Veteran's Health Administration collaborative, the American Medical Group Association collaborative) are important mechanisms for transferring knowledge on best practices and technical assistance.[101, 169, 176, 182] Furthermore, it is important to track community level costs and quality (e.g., expansion of ACO partnerships with local health departments, schools, or community-based organizations) in order to understand ACO effects on market power consolidation, cost-shifting practices, and insurer's selective pursuit of better-risk/healthier patients.[186, 187]

**Strength of Evidence: Not applicable, descriptive only.**

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

The literature review found no information regarding this question. Please refer to Chapter Six for the summary of the TEP's discussion.

# 5. Review of the Bundled Payment Literature

In this project, we were asked to examine bundled payment as a form of VBP that specifically includes both cost and quality components in the VBP design, which differs from prior bundled payment models (e.g., DRGs used in hospital setting) that focused solely on reducing costs.

Bundled payment refers to a form of payment to providers that is based on predetermined expected costs for a group of related health care services.[1] The bundles can be constructed in many different ways, covering different periods of time (e.g., a one-year episode of diabetes or a hospital admission) and different provider types providing services in single setting (i.e., hospital) or multiple settings (e.g., hospital, ambulatory, and skilled nursing). The expressed goals of bundled approaches to payment are to improve coordination across the providers engaged in caring for a patient during an episode of care and, in turn, improve cost efficiencies or savings. By putting all providers whose services are included within the bundle jointly at financial risk for management of the patient, bundled payment creates incentives to reduce the number and cost of services within the bundle.[188] In VBP bundled-payment applications with an explicit quality performance component, providers also are accountable for ensuring the production of high-quality care and/or outcomes.

A recent AHRQ systematic review of the published evidence by Hussey et al. evaluated the effects of bundled payment on spending, utilization, and quality.[1] This review examined 58 studies that examined 20 different bundled payment interventions plus four review articles that summarized studies of the Medicare Inpatient Prospective Payment System. Sixteen of the 20 interventions examined in the AHRQ systematic review addressed bundling of services by single institutional providers (i.e., hospital, skilled nursing facility, home health provider), 17 included public payers only, and most examined public or international insurance prospective payment systems. Only one of the 20 bundled payment interventions in this review contained quality performance incentives as part of the design of the bundled payment program (Geisinger "ProvenCareSM"), and thus met our definition of VBP. Many of the programs reviewed were also more limited in scope than current VBP-based bundled payment programs. Although most of what the systematic review summarizes is not relevant to our exploration of evaluating the impact of VBP-type bundled payment models, Hussey and colleagues concluded that implementation of these bundled payment interventions resulted in reductions in health spending and use of services.

The authors of the systematic review state that bundled payment programs can explicitly incorporate quality measurement in various ways, such as through eligibility thresholds that determine participation in the bundled payment program (i.e., condition of contracting), to assess potential negative consequences of bundling, to inform provider performance improvement initiatives, and for use as a P4P payment adjustment (e.g., shared savings or bonus paid). The

authors of the systematic review found inconsistent and generally small effects on quality of care, which was assessed in the context of determining whether bundled payments might lead to worsening of quality of care in efforts to reduce spending within the episode. Although the findings were consistent across different programs and settings, the authors rated the body of evidence as low due to concerns about bias and residual confounding. They also noted there was insufficient evidence to identify the influence of design and contextual factors on bundled payment effects.

At the time of our review, other VBP-type bundled payment demonstrations were under development or in the process of being tested, including the Medicare Bundled Payment for Care Improvement demonstration and the IHA's Bundled Episode Payment and Gainsharing demonstration. These programs were not sufficiently far along in their development and implementation to have generated published results to include in this review. Additionally, several private payers have been experimenting with implementation of bundled payment/episode-based payment approaches, such as Horizon Blue Cross Blue Shield of New Jersey's implementation of PROMETHEUS Payment. Again, results were not available to us for review because the programs have not been fully implemented. Furthermore, private health plans tend to view these new payment models as opportunities to gain new business by demonstrating lower costs and higher quality to purchasers; as such, they are less open to disclosing design features and results that would place them at competitive disadvantage.

## Methods

We augmented the Hussey et al.[1] review with a search of the published literature between January 1, 2011, and January 30, 2013 (Table 5.1). A librarian performed this search. The Hussey et al. review included studies—both from the peer-reviewed and grey literature—that addressed the three research questions of focus for their study (i.e., impact of bundled payment on health care spending, utilization, and quality measures; differential effects by key design features; and differential effects by key contextual factors). They excluded studies that did not report any of the outcomes of interest, did not report on a bundled payment intervention, or were limited to describing theoretical models. Studies of bundled payment interventions done outside the United States were included if they met criteria for generalizability to the United States.

One trained reviewer (Daniel Mandel), with input from a second senior researcher on the project (Cheryl Damberg), scanned the titles and abstracts and selected studies for full-text screening. For each of the selected studies, we performed reference mining to identify additional studies for potential inclusion. Two reviewers independently abstracted the data per study (Daniel Mandel, Cheryl Damberg). Studies frequently included incomplete descriptions of intervention design and context elements as well as evaluation methods and results.

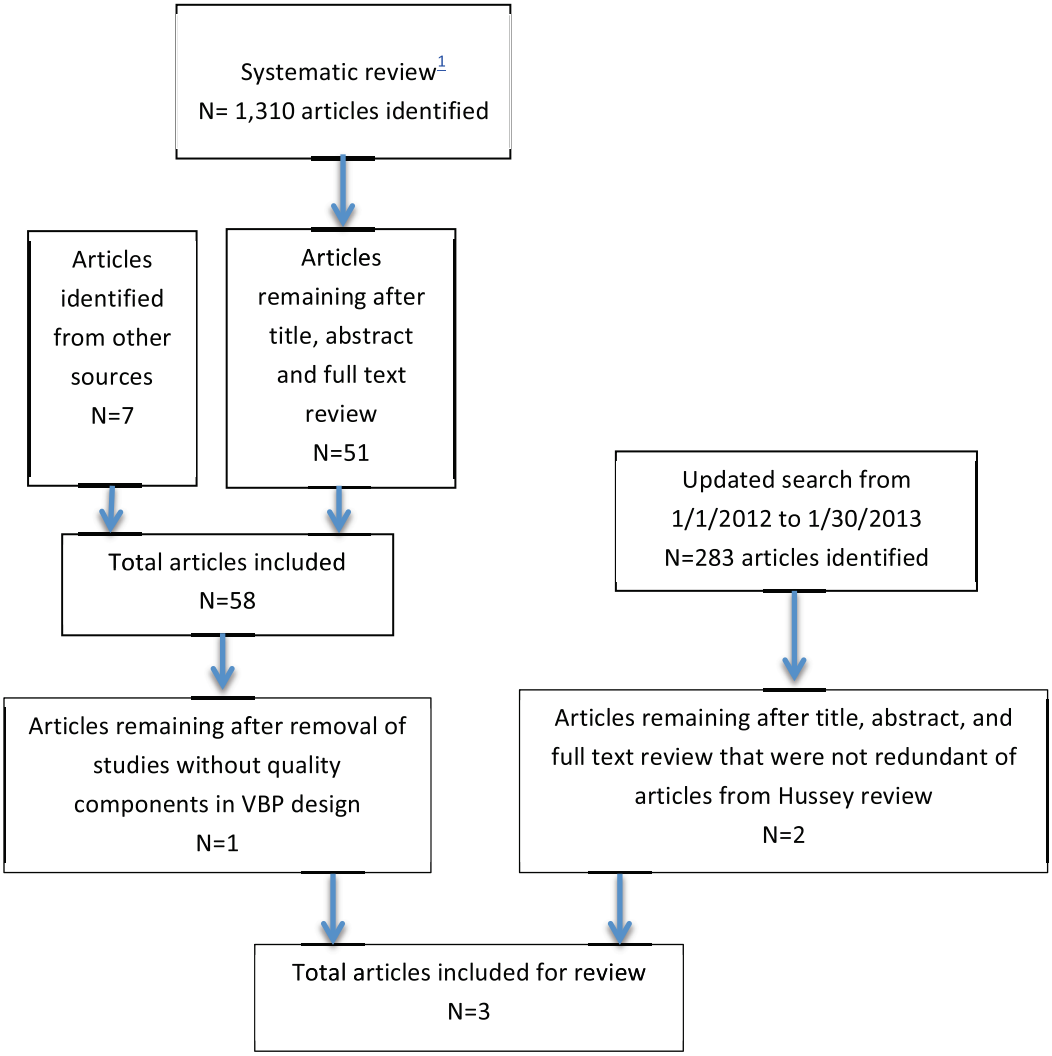
We found 238 additional bundled payment articles, reports, and commentaries between January 1, 2011, and January 30, 2013. Beyond the Geisinger ProvenCareSM bundled payment

program identified in the Hussey review (Figure 5.1), our search yielded two additional bundled payment studies that included both cost and quality elements in their design and were evaluations of bundled payment programs. The three published studies that we include in our review are (1) Geisinger ProvenCareSM, (2) the Medicare ACE demonstration, and (3) PROMETHEUS Bundled Payment Model.[104] All interventions occurred in the United States. None of the studies involved randomization: All were observational studies. All three programs addressed bundles that paid for and measured care provided in the hospital, while one of the three (i.e., PROMETHEUS) also addressed care provided in ambulatory care settings. The Geisinger quantitative study was of fair quality, the ACE demonstration case study evaluation was of poor quality, and the PROMETHEUS case study evaluation was of good quality.

**Table 5.1. Search Terms Used in Bundled Payment Literature Review**

| Search Terms used in Hussey et al., 2012 Systematic Review | Search Engine | Search Dates |
|---|---|---|
| (bundl*[tiab] OR episode[tiab] OR "prospective payment"[tiab] OR warranty[tiab] OR warranti*[tiab] OR global[tiab]) AND (payment[tiab] OR finance*[tiab] OR reimburse*[tiab] OR incentive*[tiab] OR fees[tiab]) AND (trial[tiab] OR compare*[tiab] OR effect*[tiab] OR impact[tiab] OR outcome*[tiab] OR result*[tiab]) | PubMed | January 18, 2012–January 30, 2013 |
| (bundl*:ti or episode:ti or prospective:ti or warranty:ti or warranti*:ti or global:ti) and (payment*:ti or finance*:ti or reimburse*:ti or incentive*:ti or fees:ti) and (trial:ti or compare*:ti or effect*:ti or impact:ti or outcome*:ti or result*:ti) | Cochrane Library | January 18, 2012–January 30, 2013 |
| **Terms Used to Augment Hussey Review** | | |
| (bundl* OR episod*) AND (payment* OR pay OR paying OR pays) | PubMed | January 1, 2007–November 14, 2012 |
| ((bundl$ or episod$) adj5 (payment$ or pay or paying or pays)).mp. | Medline on OVID | January 1, 2007–November 14, 2012 |

**Figure 5.1. Process Used to Identify Articles for Review, Bundled Payments**




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

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

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

## Research Questions

Our review of the literature on bundled payment programs focused on addressing the questions listed below. Because bundled payment VBP models are new, we frequently could not find much information in the published literature to inform many of the questions.

### *Measuring Performance in Value-Based Purchasing Programs*

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

In our review of public documents from bundled payment VBP programs, program sponsors most often cited improving clinical quality (8 of 10 programs), cost/affordability (6 of 10 programs), and patient outcomes (6 of 10 programs). Additional goals that program sponsors mentioned were improved coordination of care/cooperation among providers, improved patient experience, and appropriate utilization of services. The Cox Health Plan Episodes of Care Pilot, the Aetna/Hoag IHA bundled payment program, and the Arkansas Medicaid program all cite the goal of reduced potentially avoidable complications. The two CMS-sponsored bundled payment initiatives stated that their goals were to simultaneously improve quality, reduce expenditures, and increase cooperation among providers (ACE Demonstration) or align provider incentives (Bundled Payments for Care Improvement). Similarly, the Blue Cross Blue Shield Tennessee Orthopedic Bundled Payment aims to improve quality, patient outcomes, patient experience, and care coordination while decreasing cost. One of the programs, the Cox Health Plan Episodes of Care Pilot, specifically indicated that it wanted to reduce potentially avoidable complications by 25 percent ($510,000).

The three published studies we reviewed indicated the following goals for the bundled payment programs that were the focus of evaluation:

- Establish best practices for CABG survey, develop risk-based pricing, and create a mechanism for patient engagement.
- Contain costs.
- Improve quality and outcomes, achieve cost savings, improve decisionmaking for patients, improve coordination among providers, provide high-quality, low-cost care, and increase market share.

Test implementation of bundles with goals of encouraging coordination of care through joint accountability, decreasing spending while improving quality, and creating incentives to eliminate services that are clinically ineffective or duplicative (transfer financial responsibility to providers for events related to technical risk, such as avoidable complications).

**Strength of Evidence: Not applicable, descriptive only.**

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

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

3. Which aspects of bundled payments are measurable and which are not?

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

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

Because VBP-type bundled payment programs are very new, there is little published evidence on the effects of these interventions on health outcomes (e.g., mortality, complications) as associated with the clinical measures used in these programs. Among the three studies we reviewed, two articles identified that they were tracking outcomes (Table 5.2), but only one of the two reported results.[68] This study showed that surgeons closed the gap in performance on the clinical process-of-care measures, achieving 100 percent compliance for those undergoing elective CABG surgery; however, no statistically significant differences were found between patients in the intervention group and historical controls across 19 health outcome measures (complications, operative mortality, readmissions). While many of the outcome measures showed lower rates in the intervention group (directionally correct), they did not achieve statistical significance. This finding is not surprising, because the study was not adequately powered to be able to detect changes in outcomes due to limited samples sizes in both the intervention and control group of patients.

Research studies that have attempted to test the relationship between process-of-care measures and outcomes have been challenged in demonstrating results for a number of reasons, including small effect sizes that are difficult to disentangle from other factors that might explain the result, little variation in performance among providers, and poor methods for controlling confounds.

**Strength of Evidence: Insufficient.**

**Table 5.2. Articles Examining the Relationship Between Performance on Bundled-Payment Value-Based Purchasing Measures and Patient Outcomes**

| Reference | Program Description | Study Design | Program Measure(s) | Patient Outcome(s) | Findings | Assessment of Methodological Quality |
|---|---|---|---|---|---|---|
| Casale et al., 2007[68] | Geisinger ProvenCareSM (2006–2007) | 3 hospitals; 8 surgeons within the 3 hospitals<br>Single observational retrospective pre-post study design.<br>254 elective CABG cases occurring in 2006–2007 period (117 intervention patients) compared with 137 historical controls (treated in 2005) | 40 clinical process measures developed at Geisinger based on American Hospital Association/American College of Cardiology 2004 Guideline Update for CABG Surgery | 19 outcomes (e.g., operative mortality, complications, readmissions) | Despite increased adherence in process measures, no statistically significant difference across 19 health outcome measures (complications, operative mortality, readmissions). No deterioration in outcomes.<br>Many outcomes showed lower rates (directionally correct, not statistically significant) in intervention population. | **Fair:** Single observational study, 3 sites, 254 CABG cases, short time period.<br>Inability to find relationship between improved processes of care and outcomes may have been due to too small of sample to provide adequate statistical power for analyses of these outcomes.<br>Generalizability to other non-integrated health systems might be limited, as the integrated EHR was a key component in the redesign of care processes |
| Zucker, 2011[79] | Medicare ACE Demonstration (2009–2010) | Baptist Health System (TX) in three hospital sites within system, with surgeons<br>Case study, methods not reported. Describes results for 2009–2010 period | Prophylactic antibiotic prior to surgery and after surgery. anti-platelet medication at discharge, venous thromboemolism prophylaxis | 30-day mortality<br>30 day readmission<br>Post-op sepsis<br>Revascularization rates<br>% of CABG performed off-pump<br>Post-op stroke<br>% of CABG patients returned to the OR<br>Inpatient mortality | Not reported | **Poor:** One ACE site's self-reported results and site self-selected into the ACE demonstration. Unclear whether results generalize or would be confirmed by independent evaluator. Compare pre-post. No data showing whether case mix changed over period or whether steering of high risk cases to other facilities. |

*Results of Performance in Value-Based Purchasing Programs*

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

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

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

Of the three studies we reviewed (Table 5.3), two describe the impact of bundled payments. The third study focused on formative aspects of implementation of the bundled payment program in three sites that never implemented the bundled payments by the close of the study period.

Of the two studies reporting on impact, the Zucker[79] case study is of poor quality, as the site self-selected into the ACE demonstration and therefore may differ from other potential bundled payment site, was not an independent evaluation of the ACE demonstration, and is limited to one study site showing pre-posts results with no comparison group. The other study[68] also was conducted in a single integrated health system with very unique characteristics that makes generalization of the Geisinger experience difficult. The authors of this study noted that expansion to larger systems with more hospitals or those without a provider-owned insurance company would add significant logistical complications that have not yet been assessed based on this study's findings. The Geisinger ProvenCareSM experience was also unique in that Geisinger leveraged its integrated EHR platform to build 40 processes of care into clinical decision support protocols (e.g., reminders, prompts) to ensure the desired processes occurred. This forcing function could only be overridden with physician justification as to why the physician was not following the recommended American College of Cardiology/American Hospital Association guidelines.

For the two studies we reviewed,[68, 79] there were changes in costs, utilization, and in care processes, but there were no documented effects on outcomes. In the Geisinger study (a two-year examination), the authors felt the impact of reengineered processes could be measured over a reasonably short period. The bundled payment resulted in a 5 percent reduction in hospital charges for ProvenCareSM patients compared with patients under previous payment model, incorporation of 40 care processes from the American Hospital Association/American College of Cardiology guidelines into the EHR, an increase in discharge to home from 81.0 percent to 90.6 percent (p=.033), increased adherence to the 40 clinical process measures from 59 percent to 100 percent (p=0.001), no change in post-op LOS, a 16 percent reduction in total LOS (from 6.3 to 5.3 days), and a 15.5 percent reduction in 30-days readmission rate (from 7.1 percent to 6.0 percent) (no p-value given). There were no differences between intervention and comparison patients across 19 health outcome measures (complications, operative mortality, readmissions), suggesting no deterioration in outcomes—a potential concern of controlling spending and

utilization. The absence of a finding between improved process performance and outcomes is a function of too small of a study sample to detect an effect.

The case study report regarding the Baptist Health System's participation in the ACE demonstration did not report any results for clinical process, outcomes, or utilization. They did report savings, including lower device costs (saving $2 million initially, and $800,000 in year 2) and a reduction in spending of $4.3 million between 2009 and 2011 (~$2,000 per case, $500,000 total). They also reported that physicians earned ~$280 in gain-sharing payments per episode and that beneficiaries earned ~$320 per person through reductions in Part B premiums.

In the Hussey et al. systematic review assessing the impact of implementation of 19 non-VBP type bundled payment programs,[1] all of the bundled payment programs showed declines in spending and utilization, with minor and inconsistent effects on quality measures. Cost reductions typically were approximately 10 percent or less. Utilization reductions, such as LOS, saw between a 5 and 15 percent reduction compared with historical experience or controls. Once implemented, reductions in spending occurred immediately due to the revised payment rate. Overall, the findings were inconsistent on the quality measures regarding both the direction and magnitude of effects on different quality measures within a single study and for similar quality measures between studies. Studies of the Medicare IPPS found that most quality measures (not part of the program) improved post-implementation, but it was unclear whether the changes could be attributed to the IPPS. A number of studies demonstrated either no change or a decline in mortality rates following implementation of the IPPS (both in-hospital and mortality up to one year post-discharge), although again, the overall population mortality rate declined during the same period. There is less evidence that exists on nonmortality health outcomes. The evidence did not indicate that the IPPS lead to increases in hospital readmissions, there is insufficient evidence to assess impact on emergency department admissions, and one study found an increase in the level of instability of patients at discharge.

**Strength of Evidence: Insufficient.**

**Table 5.3. Evidence on Effectiveness of Bundled Payment Programs**

| Reference | Program Description | Study Design | Incentive Structure | Measures Examined | Findings | Assessment of Methodological Quality |
|---|---|---|---|---|---|---|
| Casale et al., 2007[68] | Geisinger ProvenCareSM Integrated delivery system (hospital and 8 surgeons)<br>Fixed-price bundled payment for elective CABG surgery that included pre-op evaluation and work-up, all hospital and professional fees, all routine post-discharge care (rehab), and management of all related complications. Warranty for follow-up preventive care.<br>40 care processes from American Hospital Association/American College of Cardiology guidelines were incorporated into EHR, including order sets, templates, and time outs. Patients engaged as partners in their treatment.<br>(2006–2007) | Single observational retrospective pre-post study design.<br>254 elective CABG cases occurring in 2006–2007 period (117 intervention patients) compared with 137 historical controls (treated in 2005) | Bundled payment with P4P for quality<br>Contains P4P incentives for quality care elements (up to 20% of compensation), | Adherence to 40 clinical process measures<br>Clinical outcomes measures (operative mortality, complications, 30-day readmission)<br>LOS, reduction in hospital charges. | Reduction in hospital charges of 5% for ProvenCare patients vs. patients under previous payment model<br>Increase in discharge to home from 81% to 90.6% (p=.033)<br>Increase adherence from 59% to 100% for 40 clinical process measures (p=0.001)<br>No change in post-op LOS. 16% reduction in total LOS (from 6.3 to 5.3 days). 15.5% reduction in 30-days readmission rate (from 7.1% to 6.0%) (no p-value given).<br>No statistically significant difference for 19 health outcome measures (complications, operative mortality, readmissions), although many were directionally lower in the ProvenCare group vs. comparison population. | **Fair:** Single observational study, 3 sites, 254 CABG cases, short time period.<br>Inability to find relationship between improved processes of care and outcomes may have been due to too small of sample to provide adequate statistical power for analyses of these outcomes.<br>Generalizability to other non-integrated health systems might be limited, as the integrated EHR was a key component in the redesign of care processes |

| Reference | Program Description | Study Design | Incentive Structure | Measures Examined | Findings | Assessment of Methodological Quality |
|---|---|---|---|---|---|---|
| Hussey et al., 2011[104] | PROMETHEUS Bundled Payment Pilot Experiment covers all care to treat defined clinical episode, particularly services recommended by guidelines or experts. 21 bundles. Payment rates ("evidence-informed case rates) set on basis of historical service use and cost patterns; warranty that the costs of potentially avoidable complications will not exceed an agreed-upon amount; Includes a Scorecard to measure and reward quality. 3 sites (Crozer Keystone Health System-Independence Blue Cross (PA), Employers' Coalition on Health (II), and Priority Health-Spectrum Health (MI)).<br>(2008–2011) | Qualitative formative evaluation of 3 sites attempting to implement the bundled payment model (2009–2011 | Bundled payment with shared savings based on scorecard measures | Implementation issues: defining bundle, defining payment arrangement, performance measurement and systems to support, and care redesign | None of pilot sites used PROMETHEUS payment method or executed bundled payment contracts by third year of pilot. Many implementation challenges: problems with defining bundles (problems with FFS claims information used to determine services that are part of bundle), complexity of building on existing systems, problems defining the payment method (payer hesitancy to allocate shared savings payments, provider hesitancy to accept withhold, how to allocate payments and accountability), challenges implementing quality measurement (EHR is critical but eMeasures implementation is time and resource intensive and lack of data exchange), engaging frontline physicians, and delivery redesign.<br>Strong organization commitment and support from senior management is required for successful implementation<br>Bundled payment viewed as easier to implement for procedures than chronic conditions because more clearly defined clinical pathway and fewer providers involved<br>Chronic conditions viewed as having more improvement opportunities, particularly for reducing avoidable complications | **Fair:** Qualitative descriptive study of three sites |

| Reference | Program Description | Study Design | Incentive Structure | Measures Examined | Findings | Assessment of Methodological Quality |
|---|---|---|---|---|---|---|
| Zucker, 2011[79] | Medicare ACE Demonstration (hospitals and surgeons). 5 sites are participating (this study reports on one site's results)<br>3-year demo provides single bundled payment for both hospital (Part A ) and physician (Part B) services for an inpatient stay for orthopedic (hip, knee and lower extremity joint replacement) and cardiac procedures (CABG surgery, valve replacement, pacemaker implantation, defibrillator implantation and coronary artery angioplasty). FFS beneficiaries. Entities competitively bid for each service, and price reflected a discount off the base DRG payment.<br>CMS shares 50% of the savings with beneficiary up to $1,259 (maximum annual Part B premium)—exact amount varies by site and procedure. Optional use of incentives/gain sharing with physicians permitted.<br>Baptist Health System (TX) in three hospital sites and surgeons within the Baptist system<br>(2009–2010) | Case study, methods not reported.<br>Self-report on results for 2009–2010 | Shared savings with beneficiaries<br>Shared savings with physicians<br>Bundled payment | Cost savings<br>Utilization (LOS)<br>Health outcomes (30-day mortality, 30-day readmission, post-op sepsis, post-op stroke, % of CABG patients returned to the OR, in+E2patient mortality, revascularization rates)<br>Clinical processes/safety (Prophylactic antibiotic sue and venous thromboemolism prophylaxis, anti-platelet medication at discharge) | Deployed a single uniform order set across entire hospital system. No results reported on clinical process, outcomes, or utilization.<br>Health system negotiated lower device costs (saving $2 million initially, and $800,000 in year 2)<br>Baptist reports reductions in spending of $4.3 million between 2009–2011 (~$2,000 per case, $500,000 total)<br>Physicians earning ~$280 in gain-sharing payments per episode<br>Health system negotiated lower device costs (saving $2 million initially, and $800,000 in year 2)<br>Beneficiaries earned ~$320 per person through reductions in Part B premiums | **Poor:** One ACE site's self-reported results and site self-selected into the ACE demonstration. Unclear whether results generalize or would be confirmed by independent evaluator. Compare pre-post. No data showing whether case mix changed over period or whether steering of high risk cases to other facilities. |

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

None of the three studies we reviewed found unintended consequences in the context of VBP bundled payment arrangements. The sparse literature does not provide information on whether bundled payments will lead to undesired effects or spillover effects. However, some researchers have cautioned that without the correct preventive mechanisms in place, providers may skimp on care, lowering the quality and costs of care.[189, 190] Quality measures built into the structure of bundled payment approaches are one potential safeguard against skimping on care. Also, without mechanisms in place to ensure that an episode of care is appropriate, payers may end up paying for more care if providers start to unbundle services.

The Hussey et al. review[1] examined the evidence on various hypothesized undesired effects that might occur with the implementation of bundled payment arrangements (e.g., increasing the number of bundles [volume], underuse of appropriate care services that may lead to poorer outcomes for patients, selection of low-risk patients into the bundles and avoidance of high-risk [potentially more expensive] patients, upcoding to maximize payment for the bundle, and moving services in time or location to quality for separate reimbursement). This review found limited evidence on unbundling services and upcoding, but consistent evidence of shifting services to other provider types. There was little evidence that there were major effects on quality; rather, the findings were mixed, with some measures improving while other worsened. In some cases, patient risk increased, but it was unclear whether this was real or due to coding changes. Studies of the Medicare IPPS did not indicate that the IPPS led to increases in hospital readmissions, while one study found an increase in the level of instability of patients at discharge.

**Strength of Evidence: Insufficient.** (among studies with cost and quality). The Hussey et al.[1] review graded the strength of the evidence as low because only two of the 58 studies were rated as good methodologically and 19 were rated poor. They were concerned that many of the studies used a pre-post design that may have been subject to bias from secular trends.

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

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

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

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

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

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

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

The three studies we reviewed did not contain information that contrasted the features of high and low performers. One program indicated that strong organization commitment and support from senior management is required for successful implementation. Another program noted that bundled payment was seen as easier to implement for procedures than for chronic conditions because procedures have more clearly defined clinical pathways and fewer providers involved in the care process. Another program indicated that when the health system took the financial risk and protected the physicians from any downside, this made it easier to gain their buy-in. Transparency about costs with physicians was also seen as important to manage overall spending within an episode.

**Strength of Evidence: Insufficient.**

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

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

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

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

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

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

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

The implementation of bundled payment programs has highlighted challenges related to the design and implementation of these programs and overlaying the bundled payment mechanism on existing payment systems. Formative evaluation at this early stage is vitally important, as it can highlight issues that must be addressed to facilitate the successful implementation of bundled payment programs. Among the various implementation challenges identified across these three studies (as well as those in the Hussey et al. review[1]) are the following:

- problems associated with defining bundles (problems with FFS claims information used to determine services that are part of bundle)
- complexity of building on existing payment and billing systems and having to create manual workarounds to ensure payment
- problems defining the payment method (payer hesitancy to allocate shared savings payments)
- provider hesitancy to accept withhold
- how to allocated payments and accountability
- difficulty implementing quality measurement (EHRs are critical, but eMeasures implementation is time- and resource-intensive, and there is a lack of data exchange)
- difficulty in engaging frontline physicians, and delivery redesign.

A variety of bundled payment evaluations, including the one by Casale,[68] indicate that the resources used to develop processes to manage implementation of the bundle are substantial.

The formative evaluation of three sites attempting to implement bundled payments using the PROMETHEUS approach[104] found that the recommended services within the bundled represented only a portion of cost of an episode, and there was variability in patterns of care for a single condition. They also noted feasibility issues related to independent providers having to collaborate and share payment and providers being able to manage increased financial risk.

There is little published evidence at this stage about the impact these programs are having on quality performance, particularly when providers are dually incentivized to maintain or improve quality while controlling cost and utilization. Future studies should assess the impact of bundled payments on key outcomes such as improved coordination, health functioning, and reduced morbidity and mortality, and they will require sufficient power to detect any effects. There is also a need to understand whether unintended effects are occurring or whether there are spillover effects (such as reductions in disparities in care).

Commentaries and reports have raised important gaps in knowledge in the current bundled payment literature. Sood and colleagues[191] raised four critical gaps in knowledge relating to the Medicare National Pilot Program on Payment Bundling signed into law by the Affordable Care Act: (1) Which diagnoses to include for bundled payment? (2) How long should an episode of care be? (3) Which entities will be eligible to receive bundled payments? (4) How will the

quality of care be measured? Grabowski et al.[189] raise several important gaps in knowledge in relation to the current post-acute bundled payment environment: How will hospitals receiving bundled payment react to decreases in nominal Medicare payments rates and how will access to care be affected? How do different models of bundled payment for acute care hospitalization affect costs and quality of care?

Other authors have noted gaps in knowledge about the impact on quality of care and costs associated with implementing bundled payment in other clinical settings, such as emergency departments or dialysis centers.[192–195] To move forward with the Bundled Payments for Care Improvement program, the Center for Medicare and Medicaid Innovation should test-pilot programs with providers at different stages of readiness, with incentives tied to readiness. Offering technical and best practice support to those groups that are not ready for large-scale bundled payment programs can also speed up payment reform.[190]

**Strength of Evidence: Not applicable, descriptive only.**

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

There is little information at this stage, as few VBP-type bundled payments have actually been implemented. However, given the requirements of these models that bundle care provided across providers within a single setting or across settings, an integrated health information technology system that supports the exchange of clinical information is essential to fully manage care in a coordinated manner. Additionally, enhanced functionalities in EHRs, such as clinical decision support, can help with prompting physicians to deliver appropriate care and better manage the use of services. Because bundled payments alter the financial risk landscape, they also require a substantial amount of trust between hospital and physicians as well as the payers who implement these programs to find ways to improve efficiencies.

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

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

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

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

# 6. Summary of Technical Expert Panel Discussion

As summarized in Chapters Three, Four, and Five, the literature on VBP provides an incomplete picture regarding whether these programs are successful, what defines success, and what elements need to be present for VBP programs to succeed. Many studies have examined whether P4P has improved performance on the incentivized measures, and this evidence shows modest effects from the initial P4P program designs. P4P programs are evolving both in terms of expanding the types of measures included (i.e., cost and resource use measures, outcomes) and modifying incentive structure (i.e., shared savings), and these types of changes may lead to different responses by providers and different results. Newer VBP models—ACOs and bundled payment programs—have not been in use long enough to accrue much evidence, and the models differ in important ways from the earlier P4P experiments that may lead to very different effects. While ACOs are being formed and tested, bundled payment programs have experienced more difficulties with implementation so that there is a functioning program to evaluate. Substantial gaps remain in what we know about the elements of the VBP program and the environment (e.g., provider, patient, and market factors) in which a program is implemented and how these factors influence the impact of these programs. How and why VBP programs work or do not work are very complicated questions.

The published literature and public document review could not address a number of the research questions that were the focus of this project. Because many of the design and implementation lessons have not found their way into the published literature and likely never will, we asked our TEP to provide input on questions where no or limited published information exists. These areas included what characterizes successful VBP programs (i.e., design and implementation elements), what should be measured in VBP, the costs of provider compliance with improving performance on the measures, and provider responses to VBP. We also asked the TEP to comment on our assessment of the questions that we were able to address from the literature review, to provide advice on what HHS and other policymakers should do to advance our knowledge about how best to design VBP programs to achieve stated goals, and what future evaluation and monitoring activities should be considered to strengthen the VBP knowledge base.

The TEP was composed of 14 individuals with substantial knowledge of VBP, representing VBP program sponsors (i.e., health plans, community collaboratives, and public payers), providers and health systems that have been the target of VBP programs, and academic researchers with VBP evaluation expertise. In addition to the TEP members, senior staff members from ASPE, CMS, and AHRQ who are engaged in VBP work for the federal government also participated in these discussions. The TEP met twice, for full-day meetings, in Washington, D.C., on May 6, 2013, and June 6, 2013, to review and comment on each of the 16

research questions. To frame the discussion at each meeting, RAND provided a high-level summary of the findings for each research question and a set of questions for the TEP to consider. The two senior RAND project leaders led the TEP discussions. The panelists based their input on professional experience as sponsors, evaluators, and the targets of incentives and anecdotal information. Below, we summarize the insights offered by this panel, organized by key themes that emerged in the discussions.

## Value-Based Purchasing Program Design and Implementation

### *Setting Goals and Measuring Success*

After reviewing the environmental scan results that showed that program sponsors typically set high-level goals related to improving quality and containing costs, the TEP agreed that VBP programs should establish these types of aspirational goals. However, the TEP commented that the larger goal of VBP is to transform the way that care is delivered to enhance performance. TEP members outlined the following additional aspirational goals that they believed would be important to establish and potentially use to assess VBP program success:

- **Stimulate organizational nimbleness to take a new performance target and rapidly learn and improve against the target.** TEP members indicated that a key goal of VBP is improving the functional capacity of providers to be able to learn and improve. Therefore, it is important to understand whether there is institutional capacity in health systems and provider organizations to improve quality against a moving target, and whether providers can maintain performance levels once targets are achieved. Panelists commented that VBP programs should affect providers' willingness to change, their measurement capacity to identify problems, and their ability to respond to correct quality defects.
- **Promote innovation.** The TEP commented that part of the value of VBP is the innovation that occurs to fix the fundamental problems that lead to poor quality/outcomes within provider organizations, ideally improving quality across providers in response to the incentive scheme. Examples they cited were creating more integrated data systems to improve communication between providers and care management protocols that span care settings to improve transitions in care between the hospital and ambulatory settings, investments in registries that allow physicians to track and better manage high-risk populations, development and use of risk assessment tools, and provision of clinical decision support. There was interest among the panelists in capturing whether and how VBP initiatives are stimulating innovation.

The ability to use these types of goals to assess VBP program success is predicated on defining what the concepts mean and how to operationalize the concepts for measurement. For example, what defines nimbleness and how would one measure it? What constitutes transformational change? Substantial work would be required to transform these aspirational goals into measurable goals, and policymakers would need to determine whether expending resources to do so would generate useful information to inform VBP programs.

While the TEP supported the use of aspirational goals, the panelists recognized that it is difficult to determine success in meeting these goals due to the lack of specific targets. TEP members indicated that it is important for individual VBP programs to establish measurable "intermediate" goals for what constitutes success. The TEP suggested that VBP sponsors use external benchmarks—drawn from the HEDIS, Consumer Assessment of Healthcare Providers and Systems survey, or Healthy People 2010 performance data—to set national or regional goals for what constitutes success and then measure whether providers were successful in achieving established benchmarks. For example, CMS could peg success for physicians or Medicare Advantage plans to achievement of the 90th percentile of performance for HEDIS measures. Performance benchmarks could similarly be established for total medical expense, by pegging the growth in risk-adjusted total cost of care year-to-year to no more than "X" percent higher than the overall rate of inflation (i.e., the Consumer Price Index).

### *Design Issues*

#### Use of Consumer Incentives and Alignment with Provider Incentives

TEP members discussed the potential value of consumer-oriented incentives that would direct consumers to higher-performing providers. The TEP observed that consumer incentives that drive market share toward high-quality providers would help align incentives in the system and potentially have stronger effects than the current P4P-style bonuses, which work at the margin to change provider behavior. In effect, this pits providers against each other by providing incentives for patients to choose better performing providers, a strategy that is at odds with the TEP's guidance about not establishing "relative" performance targets that promote competition among providers. The expectation is that overall quality would improve for two reasons: (1) the portion of people using high quality providers will increase, and (2) the prospect of losing patients, among those who are not high performers, will provide an additional incentive for providers to improve their performance. CMS already has begun to use consumer incentives. For example, beneficiaries are precluded from enrolling online in a Medicare Advantage plan that has consistently low performance for three years. Starting in the fall of 2012, CMS began sending letters to beneficiaries enrolled in low-performing Medicare Advantage plans to encourage them to shift to high-performing plans; CMS allows beneficiaries in low-performing plans to change plans any time during the year to facilitate plan switching. In 2015, CMS will have the authority to terminate consistently low-performing contracts. The TEP encouraged CMS to continue to expand the use of tools like these to push quality improvement in a strategic way.

Additionally, some plans have provided beneficiaries with direct financial incentives to obtain certain types of care by either reducing their financial burden by varying co-pays (i.e., value-based insurance design) or actually paying people to do something, such as obtaining screenings or prenatal care services. For example, some Medicaid plans have experimented with paying patients to obtain immunizations. Some panelists expressed particular concern about

paying patients to receive care and that doing so could increase costs to the system and work against efforts to improve affordability.

### Measure Alignment

A number of panelists discussed the importance of measure alignment across VBP programs to give providers a clear signal of what is important and not have inconsistent measure definitions. Other TEP members stressed that the downside of measure alignment is that it can result in trimming the measure list to the least common denominator set. Furthermore, different health plans manage vastly different patient populations, and it is more important for measures to align with the payer's population than with other VBP programs that may focus on a different population. Health plans should target and adapt VBP programs to the unique population with which they work. There was consensus, though, that if programs are measuring an area where established measures exist, they should use the measures as defined and not tweak the measures.

### Performance Targets

Panelists discussed the importance of the methodology used to measure and reward performance. A number of panelists stressed the importance of rewarding both achievement and improvement. Multiple panelists felt very strongly that the programs should not be designed as a "tournament" wherein relative thresholds are used and providers are pitted against each other. Relative thresholds can result in some high-performing providers being rewarded while other high-performing providers are not when the differences in their performance are very small and not clinically meaningful (e.g., delivering a service 97 percent versus 98 percent of the time). Instead, TEP members offered that the reward should be based on objective targets that are defined in absolute terms. Providers can strive for a number of targets along a continuum and compete against themselves rather than setting up a competition between providers. This approach provides motivation to move up the scale. Some TEP members felt that the targets should be developed with provider input to ensure they are clinically meaningful. One TEP member commented that having recognition awards for the "most improved" has served as a strong motivator for improvement and has resulted in a narrowing of the performance gap between higher- and lower-performing providers.

### Measures

The TEP highlighted problems with the narrow set of measures typically being used in VBP programs, commenting that the measures evaluate only a small fraction of the total care delivered (~<20 percent in their estimation) and that some conditions have a large number of measures, while others have few or none. It was their opinion that the current, more narrowly focused set of measures tends to encourage "teaching to test" (that is, focusing only on improving areas that are measured and incentivized by the P4P program and ignoring clinically important areas that are not) rather than wholesale improvement. Moreover, the TEP expressed

concern that it is hard to show that VBP programs lead to performance improvements when the incentivized measures are the same set of measures that have been used for nearly a decade (i.e., Joint Commission, HEDIS), many of which have topped out. They commented that shifting measurement focus to areas where performance is lagging would better address the question of whether VBP can improve the delivery of care in areas not previously the focus of reporting and incentives.

#### *Expand the Measures in Value-Based Purchasing to Incentivize Broad Improvement*

Many members of the TEP thought that a broad and more comprehensive set of measures in VBP programs would create incentives for providers to perform well across the board, rather than focus narrowly on a small number areas, which promotes focusing only on improving those areas and ignoring others. However, neither the literature nor the TEP addressed how many measures are reasonable or practical to implement or when the data collection burden on providers becomes excessive.

Expanding the set of measures included in VBP programs to more comprehensively assess care delivered and include infrequently captured measure domains will require the development of additional measures and even new types of measures. The TEP noted that it will be important to develop a framework to guide future directions about what to measure and, in turn, what measures need to be developed. They also commented that the framework should address the multiple levels where behavior change needs to occur and where interventions should be directed (i.e., health system, institution, and individual provider). Developing new measures is a time- and resource-intensive activity. Measurement concepts must be defined, specifications developed, data collection processes piloted and data validated, among other steps. Measurement areas identified by TEP members as particularly in need of work include the following.

#### *Increase Measurement of Patient Outcomes and Functional Status*

The TEP members agreed that the ultimate objective of VBP is to hold providers accountable for and financially incentivize provider performance primarily based on measures of health outcomes. CMS expressed that it is moving toward this in its hospital and physician VBP programs, but is struggling to determine the right mix of structure, process, and outcomes in its programs. An example of this transition to outcomes is illustrated in the hospital VBP program. In the first year of hospital VBP, 70 percent of the measures were process measures, whereas in the second year the percentage dropped to 30 percent as currently outlined in CMS's proposed Notice of Rule Making. Questions remain about the pace at which CMS should push toward outcomes measurement, the types of outcomes to use, and the consequences of those actions. There was consensus among the TEP that functional status/health status is an important, feasible measure and shifts programs toward outcomes. There are settings and providers that are already measuring functional status on a regular basis. One TEP member noted there are examples where functional status is collected: Medicare ACO programs are paid for reporting patient-reported

functional limitations and experience of care and CMS collects health status information in nursing home and home health arenas. The Dartmouth Institute is measuring quality-adjusted life years and has built functional status (this is considered a vital sign) into a provider order for life-sustaining care for those at or near end of life. Other provider representatives noted that they are also measuring health status for some conditions. One TEP member suggested that CMS could implement the Patient Reported Outcome Measures (PROMs) questions, as the UK has done, to assess functioning.

#### *Advance the Ability to Measure Cost and Value*

Some VBP program sponsors are starting to measure and hold providers accountable for the total cost of care to the payer in addition to quality performance, in an effort to create an incentive to improve the coordination of care between hospital and ambulatory providers and to begin to move towards rewarding value in health care. One of the difficulties cited in defining value is that the various stakeholders perceive it differently—so perspective is important (i.e., value to whom). The TEP thought it important to achieve consensus on an overarching view of what value means and then leave it to local entities (e.g., payers, multi-stakeholder organizations) to develop measures to determine how best to assess value. Many organizations have struggled with how best to measure and convey value to providers and consumers, highlighting the need to consider how to construct measures of value.

#### *Include Measures That Assess the Appropriateness of Care*

TEP panelists were supportive of including measures of appropriateness (i.e., overuse) in VBP programs, but they recognized additional work is required to develop the definitions and engage providers in use of these measures. They noted, however, that without an external impetus, providers have little incentive to use practice guidelines or protocols that might withhold care due to the current FFS and malpractice systems, which instead provide an incentive to increase the use of diagnostics and procedures. The TEP also suggested that providers under risk-sharing arrangements will be more likely to implement appropriateness guidelines. However, one provider TEP member stated that, based on her hospital's experience, implementing appropriateness criteria measures in a health system can take years. TEP members suggested that measurement of shared decisionmaking is one of the keys to implementing appropriateness of care. A TEP representative of one health system noted that it is piloting a process of "patient appropriate order entry" where the specialist has to attest that a discussion with the patient about the appropriateness of care occurred. Another TEP member recognized the challenge that physicians could face if appropriateness of care measures conflict with patient preferences.

#### *Enhance the Ability of Electronic Health Records to Support Performance Measurement and Improvement*

There was widespread agreement that it is important to incentivize and help providers build the infrastructure for quality improvement. EHRs may facilitate measurement and improvement, but

TEP members felt that is unlikely in the near term. Based on their experiences to date, the TEP expressed concern that most EHRs are far from including a comprehensive set of standardized data in data fields that can readily produce meaningful data for performance measurement, in part because providers are not demanding it. Currently, meaningful use under CMS only requires that EHR vendors generate a handful of quality measures, which is not the same as freeing up the data for use in other ways. ASPE noted that it is working with the Office of the National Coordination for Health IT to make EHRs function to facilitate automated capture and reporting of quality measures, but this will be a long process.

#### *Explore the Capability of Data Registries to Support Performance Measurement*

The TEP also discussed the potential utility of specialty-society data registries (e.g., the American College of Cardiology), which are more likely to be trusted data sources by providers, to support performance measurement, particularly outcome measures. However, some TEP members were skeptical about the use of specialty society registry data precisely because the data are controlled by professional societies, which make governance decisions that may be different than what may be needed for VBP programs. Professional societies may be reluctant to release data for VBP programs that show significant room for improvement. In addition, in most cases these data are not verified for comparability across the different institutions submitting them and require audits and data quality improvement prior to use for accountability and payment.

### *Implementation Issues*

Program sponsor characteristics, provider characteristics, and the interactions between programs and providers affect whether VBP programs are implemented successfully. The TEP provided important insights on this issue based on their experience implementing or participating in VBP programs.

#### Provider Engagement and Support

The TEP viewed provider engagement as critical to garnering the desired response to the incentives. Additionally, the panelists commented that VBP sponsors can support providers in their efforts to succeed.

#### *Involve Providers in Measure Selection*

The TEP emphasized that for there to be buy-in, providers need to feel comfortable that there is a relationship between measures that are the basis for payment in the VBP program and what physicians believe represents good care that will positively impact patient outcomes (i.e., evidence-based measures that are clinically compelling). The TEP also indicated that measures need to be feasible from the provider's perspective, and the actions needed to influence the

measure need to be within the provider's locus of control. Moreover, provider involvement in measure selection may help identify potential unintended consequences early on in the process.

#### *Help Providers Succeed by Providing Support*

There was an extensive discussion among the TEP of the importance of support to help providers improve, particularly by using health information technology, data registries, and the provision of technical assistance. Examples of technical assistance mentioned by TEP members included providing comparative benchmarking data on variations in practice and factors contributing to differences (e.g., greater use of name brand drugs, higher use of costly imaging), infrastructure support, relevant and timely patient clinical data to facilitate care management, QI support and coaching, and additional staffing support such as care managers. CMS commented that they received numerous requests for data in the PGP demonstration and now in the ACO demonstrations; providers want more data delivered in a timely manner to help improve care for patients. TEP members acknowledged that the challenge is identifying which data are most useful for providers. Supporting providers through the sharing of best practices and consultative support were also identified as important. Data transparency was also seen as a strong motivator of change for providers.

#### *Align the Incentives That Front-Line Providers Face*

The TEP expressed concern that in many VBP programs the financial incentives do not "trickle down" to the individual providers. Many VBP programs target the organization or practice-level rather than individual physicians; therefore, it can be difficult to get full engagement from front-line staff if they are unaware of incentives and/or have incentives that are unaligned with the VBP program. The TEP felt it was important that VBP sponsors work with providers to ensure that incentives are aligned at all levels of the system.

### Elements Likely to Affect the Success Implementation of Value-Based Purchasing Programs

Others factors that affect VBP implementation relate to the sophistication of providers and their ability to successfully operate in a VBP environment. The panelists offered insights based on their experience about what elements contribute to successful program implementation. First, the size of the patient population—if the patient population covered by a VBP program is not large enough, it is hard to get the attention of providers (e.g., such as in bundled payment programs). Second, the complexity of the program—if the program is too complex, providers will lose the line of sight. Third, data and analytic capabilities to manage risk—if you give providers capitation, they need the full claims detail (encounter data are not sufficient) to be able to manage the risk. Fourth, it is vital to build the infrastructure for quality improvement, such as training providers on panel management. Fifth, systems improvement is important, such as building registries to track population health and team-based care. Additional features that the TEP identified from their own experiences as VBP sponsors or as providers exposed to VBP

incentives that characterize high performers include provider engagement and the provision of data and other quality improvement support, performance targets that reward both achievement and improvement, and sufficiently sized incentives to garner the attention of providers. The American Medical Group Association also has developed a framework and defined the characteristics of a high-performing health system.

#### Dissemination of Best Practices from the Highest-Performing Providers in Value-Based Purchasing

TEP members reported that the sharing of best practices is occurring through trade conferences and regional QI activities; however, this information is not routinely published. Panelists said that it would be useful to extract and compile lessons learned from providers about best practices they have implemented and to widely disseminate this information. They encourage HHS to conduct case studies of high-performing providers to see what factors they identify as contributing to producing positive results. However, to fully understand what best practice is also requires examining what poor performers are doing, as these providers may be doing many of the same things as the high performers. By looking at both sets of performers and what differentiates high and low performers, one can start to winnow the set of elements that contribute to the desired outcome.

Alternative approaches to disseminating best practices were discussed by the TEP. Some panelists felt that for dissemination to be effective, awareness is needed of how low-performing organizations/providers with different resources and capabilities than the high performers will interpret and use the information that is being disseminated. Some providers may be more receptive to the information if the provider is "like them," meaning that providers may need peer-to-peer coaching by providers located in their own community who have similar characteristics to overcome resistance to adoption of certain practices. Other providers who are willing to innovate may look to other organizations for their "good ideas" as a way to continue to improve, regardless of where they are located or their characteristics, and will embrace best practices from dissimilar organizations or practices.

## Monitoring and Evaluation of Value-Based Purchasing Programs

### *Framework for Assessing Value-Based Purchasing Programs*

The TEP strongly endorsed the need for a conceptual framework to guide how to assess VBP programs (see as examples Chapter One, Figure 1.1). Given the complex interplay between incentive program designs, provider characteristics, and other external factors, evaluations of VBP programs need to consider an array of factors that potentially contribute to observed effects. Additionally, because VBP programs are largely natural experiments and the associated research is observational in nature, several members of the TEP stressed the importance of selecting theory-driven hypotheses about how incentives affect behavior, so as to identify potential

confounding factors that could explain observed effects.[13] The TEP believed a framework would be a useful construct to help develop theory-driven hypotheses.

### *Qualitative Evaluation Work Can Inform Value-Based Purchasing Design and Implementation*

The TEP broadly agreed that more qualitative research is needed to understand what has been learned by those who design and sponsor VBP programs and by the providers who are targets of the VBP programs. There has been a lot of iterative work by VBP program sponsors, and case studies could shed light on the lessons that have been learned that are not making their way into the published literature. Qualitative research focused on understanding what does and does not work regarding design and implementation would be useful to those designing VBP programs. One TEP member suggested qualitative comparative analysis as one qualitative analytic methodology that might be a good fit for VBP evaluations, as it attempts to isolate key factors that are necessary conditions and those that are sufficient conditions to achieve the outcome. This approach acknowledges that there are a number of possible paths or combinations of elements (e.g., alternative designs) that may lead to the desired outcome. Some TEP members also supported looking at programs that are successful (i.e., the positive deviants) to isolate what they are doing that seems to be contributing to success. However, we note that similar to the need to have a comparison group in an impact study, it is important to look at programs that are not successful as they may be doing many of the same things as unsuccessful programs. The other area flagged by the TEP where qualitative work would be beneficial is understanding what changes providers are making in response to VBP programs.

### *Quantitative Assessment of Impacts*

The TEP supported the need to evaluate the impact of VBP programs, and that having a common set of variables that potentially influence outcomes, such as program characteristics (e.g., size and type of incentives), market characteristics (e.g., extent of monopoly power among providers in the market), provider characteristics, and other facilitators/enablers, would facilitate this work. The TEP highlighted some of the challenges with evaluations conducted over the past decade: (1) the measures included in a VBP program are often also included in national performance measurement programs (e.g., CMS) and the VBP programs by other private sponsors, making it difficult to tease out the effect of any individual VBP program; (2) presence of other incentives (e.g., public reporting/transparency of performance results) that make it difficult to isolate the effects of the financial incentives; (3) absence of a comparison population when a VBP program is implemented statewide or nationally; (4) small size of payment incentives; (5) VBP programs typically have used the same core measures (i.e., HEDIS, Joint Commission) that have been used for more than a decade and which are largely "topped out"; and (6) a substantial lag for the data required to assess impact, such as data on avoiding admissions and readmissions.

The TEP discussed whether success should be defined by levels (i.e., absolute performance achieved) or the counterfactual (i.e., the extent of improvement in performance compared with what it would have been absent the VBP program). Quality is improving broadly over time, and, as such, provider practices may reach stated goals without the VBP program. CMS may consider the program a success if a certain level of performance is met, whereas researchers would consider success by measuring whether improvements in performance occurred for those exposed as compared with controls. This is a key disconnect between program administrators and academic evaluators in what is of interest.

TEP members did not feel that assessing the relationship between the process measures included in the VBP program and health outcomes was a priority for VBP evaluation work, given an array of challenges in trying to demonstrate the associations in practice outside an RCT. Challenges include the fact that the patient population that is receiving the process measure may be different in important ways from those included in the RCT, there is less variation to detect effects, and that the RCT effect sizes were small to begin with, making it harder to observe in practice after accounting for potential confounds. Instead, the TEP thought that if VBP programs shift the focus of measurement and accountability more toward outcomes, it would be better to measure this directly. Doing so would give providers the license to figure out how to achieve good outcomes. It would also provide disincentives to focus only on improving processes that are measured and incentivized by the P4P program and provide flexibility for innovation of best practice models and program characteristics.

### *Collecting a Common Set of Factors Across Value-Based Purchasing Programs*

A significant challenge in trying to make sense of the findings from various VBP studies is the lack of information to compare the features of the different programs—including design elements, provider characteristics, and other external factors—and their influence on observed outcomes. The TEP strongly supported the development of a common catalog of VBP program characteristics (design features, contextual factors) to determine whether we are studying the same thing or different things when comparing across studies. The list of elements contained in Appendix B represent the type of elements the panelists recommend be systematically collected across VBP programs. They felt that having a common set of variables across all programs could facilitate synthetic approaches to assessing the impacts of VBP programs and help to elucidate the necessary and sufficient conditions under which VBP works. They encouraged the field to agree on the common elements that both public and private sector VBP programs should collect to help evaluate VBP's impacts.

### *Comparison Groups Needed for Impact Assessments*

The TEP discussion highlighted differences between academic evaluators and VBP program operators, who have different standards and expectations regarding the rigor of evaluation studies. The VBP program sponsors on the TEP felt that study designs need to be adapted to fit

with the needs for making program changes, such as more rapid but less rigorous initial evaluation cycles to generate information that could be used to adjust program design based on early experiences with implementation. For studies that focus on measuring the impact of VBP programs, the TEP agreed that including a comparison group whenever it is possible to construct one should be a minimum standard to control for possible confounding factors that may explain observed outcomes. They acknowledged that absent a comparison population, it would be easy to be misled when the trends in quality or costs were happening irrespective of the VBP program intervention. The TEP recognized that a comparison group is not feasible when a program is implemented nationally and suggested that appropriate analytic techniques be used in absence of a comparison group to control for possible confounds.

### *Spillover Effects*

When presented with evidence from the literature review of the limited and mixed evidence on spillover effects, the TEP offered additional thoughts on spillover effects based on their own experiences, which have not appeared in the published literature. In the context of the Blue Cross Blue Shield of Massachusetts AQC, the Blue Cross Blue Shield of Massachusetts TEP member stated that the Harvard evaluation team has examined the effect on quality indicators not included in AQC, particularly for children with special needs. Spillover effects were not observed on the non-incentivized measures with respect to quality of care. The AQC has also observed that within the 11 practices participating in AQC, improvements were larger for AQC members (i.e., HMO members) than for Blue Cross Blue Shield of Massachusetts PPO members, again suggesting no spillover effects to other patients within the same insurer. Based on this finding, the AQC assumes that there are no spillover improvements for PPO patients for other health plans. The AQC panelist offered two possible reasons for the absence of spillover effects. First, Blue Cross Blue Shield of Massachusetts has provided AQC providers with better data on AQC members than other plans' members, so a provider's behavior changes only for the AQC patients. Second, the AQC providers invested in the use of case managers and other resources for high-risk subgroups covered by the AQC, and these resources were not available for other patient populations the provider serves. Other TEP members agreed this is a common occurrence, as plans provide targeted resources for their members who are the focus of the VBP programs. (We note that in the literature review section of the ACO chapter of this report, the article published by McWilliams et al. describes spillover effects on utilization of services from the AQC to Medicare, but inconsistent effects on quality). The TEP supported the need to monitor spillover effects in VBP programs.

### *Disparities*

The TEP panelists recognized the importance of monitoring the effects of VBP programs on disparities in care. They also agreed that assessing the effect of VBP on disparities is difficult to monitor due to the lack of routinely collected data on the demographic and socioeconomic

characteristics of patients. TEP members indicated that they had faced challenges in capturing this information, despite their interest in capturing self-reported language, health literacy, and indicators of patient vulnerability to help improve their ability to work with patients. However, several providers on the TEP stated they were making inroads in the data they capture to be able to examine disparities. For example, one delivery system reported that it has a mandatory data gathering protocol for zip code, race, and ethnicity.

The published evidence on the effects of VBP (mainly P4P programs) provides little evidence that P4P programs have helped to close gaps in care between population subgroups. One TEP member commented that in the context of the Blue Cross Blue Shield of Massachusetts AQC, there has been evidence of reductions in racial and ethnic disparities. A number of the AQC provider groups with low-SES patient populations achieved some of the highest gains and absolute quality scores; however, this was not a universal finding among all their groups with low-SES patients. Blue Cross Blue Shield of Massachusetts discussed the challenges of working to reduce disparities with the providers, and the providers signaled to Blue Cross Blue Shield of Massachusetts what they needed to make the efforts worth their while in the incentive structure; as a result, Blue Cross Blue Shield of Massachusetts raised the weight on outcome measures so that providers would receive more in the way of financial incentives for performance on the measures. In response, the provider practices with a low-SES patient mix worked to innovate with their population and to get their doctors to improve quality.

### *Undesired Effects*

There was consensus among TEP members that as the number of VBP programs and the magnitude of the financial incentives increase and the types of measures that are the focus of the incentive (i.e., shifting from process measures to outcomes) change, the potential for undesired behaviors is likely to increase. The TEP panelists supported continued monitoring for unintended consequences, including the loss of revenue for providers caring for disadvantaged populations, the excessive use of patient exclusions, and potential overtreatment of patients based on the types of measures used. They also supported the need to evaluate patient experience and patient turnover from practices (by looking at claims and survey data) to assess whether providers are avoiding caring for more difficult patients (i.e., has the risk profile of the provider changed over time) and creating access barriers. The TEP identified a couple of VBP program design features that can help guard against undesired behaviors. First, particularly for outcome measures (i.e., clinical outcomes, LOS, and cost measures), it would be important to case-mix adjust for differences in the sickness levels of the patients treated by different providers as a way to counter the incentive for providers to seek out only healthier patients. The TEP cautioned against adjusting for race and ethnicity to avoid setting different standards for performance based on these characteristics. Second, VBP sponsors should audit the data focusing on key data elements that contribute to the performance score (upcoding of risk factors or under coding of the outcome that is being measured), to guard against gaming of the data.

There were concerns about VBP's potential (particularly the move to ACO models) for increasing market concentration and the potential adverse effects on prices. To the extent that VBP models encourage providers to become more vertically and horizontally integrated, this may mute the effects of VBP on costs. The TEP supported the need for evaluation work to determine whether consolidation is leading to adverse effects and to develop policies that hold providers accountable for cost increases (e.g., the Massachusetts law that established guardrails for the maximum rate of growth in health care spending by providers and hospitals).

### *The Composition of the Accountable Care Organization*

TEP members identified a need to understand better the composition of the ACOs, the level of integration and speed at which they are able to form, how agile the partners in ACOs are, and their ability to maneuver with the new requirements.

### *Contributors to the Cost of Care*

Finally, one TEP member suggested that a better understanding of the contributors to the cost of care should be at the top of the VBP research agenda.

# 7. Conclusion

The application of performance-based payment models represents a work in progress regarding how best to design VBP programs to achieve desired goals, the optimal conditions that support successful implementation, and provider response to the incentives. We believe that continued innovation is desired at this early stage of VBP development and implementation. Concerted efforts will be required to ensure that the lessons learned from these experiments are identified and disseminated to advance the use of VBP as a strategy for improving federal and private health care programs.

The findings of our review of the literature and public documents highlight important challenges in developing the evidence base for VBP. At this point in time, little is known about two of the three VBP models with regard to whether they can be successfully implemented and demonstrate impacts on cost and quality. Furthermore, P4P programs are evolving in their design such that the effects that are observed from new design structures may differ from the results of the experiments of the past decade. Other challenges also exist related to the measures used in VBP programs; investments are required to develop a broader set of measures and to enhance the data infrastructure to better support collection of data required to drive quality improvement and construct performance measures.

From this review, we identify three critical areas that require attention to advance progress on the federal government's use of VBP as a strategy for driving improvements in the health system:

1. **Develop a National Value-Based Purchasing Strategy.** HHS should develop a national VBP strategy for Medicare analogous to its National Quality Strategy. HHS should form a workgroup that brings together representatives from CMS, ASPE, AHRQ, and other government agencies and draws on the expertise of private-sector program sponsors and providers to develop the strategy. The strategy should outline what the federal government's goals are for VBP and thus what constitutes success, the priority areas for measurement, a timeline for increased focus on outcomes and other high-priority measurement areas, and a coordinated research agenda across CMS's VBP initiatives. The strategy will also need to consider the interplay between various CMS VBP initiatives in working to advance federal goals for VBP and how those initiatives could better align incentives to providers.
2. **Develop a Well-Defined, Coordinated Research Strategy.** Many unanswered questions remain about VBP's effectiveness and the features associated with successful VBP programs. How and why VBP programs do or do not work are very complicated questions. A well-defined, coordinated research strategy is needed to generate the information required to fill gaps in the knowledge base. Currently, federal efforts to develop, test, and evaluate VBP programs are occurring setting by setting. This presents an opportunity to coordinate the evaluation work being performed across the various

VBP initiatives within CMS to draw lessons across programs and provider settings that will inform the design and implementation of the next phase of VBP programs. As a first step, HHS could work to develop a common evaluation framework and a prioritized set of research questions, by setting and across settings, that would serve to guide CMS-sponsored evaluation studies, better align the actions of the agency to generate the desired knowledge, and coordinate use of limited evaluation resources.

The systematic collection of a core set of program design and context variables for all VBP programs would be an important step toward facilitating program evaluations and the ability to compare and contrast observed impacts across programs. Federal agencies have the ability to make collection of these variables a condition of receipt of federal funding—such as in the context of the Center for Medicare and Medicaid Innovation's grants for testing new models of care delivery and VBP, such as ACOs. HHS and CMS should leverage Medicare and Medicaid reporting requirements and HHS-sponsored experiments to learn more than we know today. Additionally, HHS could support the formation of a private/public-sector learning collaborative, with participating organizations agreeing to share design information and other data with researchers, using an agreed-upon data sharing protocol and participating in the development of the research questions.

3. **Chart a New Strategy and Process for Developing Measures to Support Federal Value-Based Purchasing Programs.** Performance measures are foundational to VBP. The heavy emphasis on performance measures in the TEP discussions underscores the importance of measures to the VBP enterprise and the inadequacy of existing performance measures to transform the delivery of health care. Progress to develop a new generation of performance measures should be accelerated and streamlined to meet the urgent and growing needs of the VBP programs to move beyond primarily assessing processes of care to also focus on evaluating patient outcomes, the appropriate use of services, and assessing quality across settings in the context of patient episodes of care. We encourage ASPE to work with measure-development experts to chart a new strategy and process for developing measures to support VBP programs.

# Appendix A: Value-Based Purchasing Programs Included in Review of Public Documents

## Pay-for-Performance Ambulatory Programs

CMS Medicare Care Management Performance Demonstration
CMS Home Health P4P Demonstration
CMS Medicare Advantage Bonus Demonstration
CMS ESRD Quality Incentive Program
WellPoint Provider P4P (12 states)
Blue Cross Blue Shield of Hawaii (HMSA) Primary Care Pay for Quality Program
Blue Cross Blue Shield of Louisiana BTE Program
Blue Cross Blue Shield of Massachusetts PCIP
Blue Cross Blue Shield of Minnesota Recognizing Excellence (discontinued 2011)
Blue Cross Blue Shield of North Carolina's Health Plan for State Employees and Teachers-BTE
Blue Cross Blue Shield of Rhode Island Quality Counts
Blue Cross Blue Shield of Texas BTE program
Geisinger Health Plan primary care physician compensation system
Geisinger Health Plan specialty physician compensation system
Highmark Quality Blue Physician Program
Horizon Blue Cross Blue Shield New Jersey Quality Recognition Program
Hudson Health Plan
Independent Health-Practice Excellence
Inland Empire Health Plan
L.A. Care
MVP Healthcare-regional primary care programs
New York City Health and Hospital Corporation (proposed)
Tufts Health Plan-Provider Network P4P (AHIP, 2009)
Western Health Advantage
IHA P4P program
New York Department of Health–Health eHeart program
New York Department of Health–CHAMPION program
Heartland Healthcare Coalition–BTE
Colorado Business Group on Health–BTE
Pennsylvania Health Choices P4P Program
Pennsylvania Access Plus P4P Program

Wisconsin Medicaid
Vermont Blueprint for Health
CMS Physician Value-Based Modifier (2015)
Anthem Blue Cross California
Aetna Provider Quality Performance-Physicians
Aetna California
Cigna California
United Healthcare Practice Rewards (84 markets in 27 states)
United Healthcare Physician Performance-Based Contracting
United Healthcare California
United Healthcare–Patient Centered Medical Home Pilots (13 states)
Dean Health Plan–Dean Value Contract (Physician)
Blue Cross Blue Shield of Massachusetts Alternative Quality Contract
Blue Cross Blue Shield of Michigan Physician Group Incentive Program
Blue Cross Blue Shield of Tennessee Physician Program
Blue Cross Blue Shield of Tennessee Medical Home
Blue Shield of California
Capital District Physician's Health Plan
CareFirst Blue Cross Blue Shield Patient Centered Medical Home
Harvard Pilgrim Plan Quality Advance Program
Health Net
HealthPartners Minnesota Quality Care Plus
Highmark Patient Centered Medical Home Pilot
Independence Blue Cross Pennsylvania PCMH P4P
Medica Choice Care Quality Improvement
Priority Health Partners in Performance
Tufts Health Plan Coordinated Care Model
IHA-Value Based P4P (2013)
Oregon Health Leadership Council
Puget Sound Health Alliance
Colorado Accountable Care Collaborative Program

## Pay-for-Performance Hospital Programs

CMS/Premier Hospital Quality Incentive Demonstration
CMS Hospital Value-Based Purchasing Program
CMS HAC Payment Policy
CMS Hospital Readmissions Reduction Program
Aetna Pathways to Excellence Hospital Incentive Program (2009)

Blue Cross Blue Shield of Massachusetts HPIP
Blue Cross Blue Shield of Michigan Hospital P4P Program (small, rural)
Blue Cross Blue Shield of Michigan Hospital P4P Program
Blue Cross Blue Shield of Rhode Island Hospital Quality Program
Blue Cross Blue Shield of Tennessee Hospital Program
Dean Health Plan Dean Value Contract (Hospital)
Excellus Hospital Performance Incentive Program
Harvard Pilgrim Plan Hospital P4P
Highmark Quality Blue Hospital Program
Horizon Blue Cross Blue Shield New Jersey–Hospital Recognition Program
MassHealth Hospital P4P Program
United Healthcare Hospital Performance Based Contracting
WellPoint Q-HIP Hospital Quality Program (14 states)

## Pay-for-Performance Skilled Nursing Facility Programs

CMS Nursing Home Value Based Purchasing Demonstration
Colorado Medicaid
Georgia Medicaid
Indiana Medicaid
Kansas Medicaid
Maryland Medicaid
Minnesota Medicaid
Ohio Medicaid
Utah Medicaid
Iowa Medicaid
Oklahoma Medicaid

## Accountable Care Organizations

CMS Medicare Shared Savings Program
CMS Advance Payment Initiative
CMS Pioneer ACO Model
CMS Physician Group Practice Demo
Aetna ACO for Medicare Advantage members
Aetna-Hunterdon Healthcare Partners ACO
Aetna Banner Health ACO
Aetna Carilion Clinic ACO
Cigna-Dartmouth-Hitchcock Medical Home (Collaborative Accountable Care)
Cigna Piedmont Physicians ACO

Cigna Partners in Care ACO
Cigna Medical Clinic of North Texas CAC
United Healthcare Arizona Connected Care
United Healthcare Medicare Advantage Shared Savings Program
WellPoint Dartmouth Hitchcock
Blue Cross Blue Shield of Illinois AdvocateCare
Blue Cross Blue Shield of Michigan Organized Systems of Care (2013)
Blue Cross Blue Shield of Minnesota–Aligned Incentive Program
Blue Shield of California CalPERS ACO
Excellus
Health Partners Minnesota–Allina ACO
Horizon Blue Cross Blue Shield New Jersey Optimus ACO
Humana-Norton Healthcare ACO
Independence Blue Cross Pennsylvania Integrated Provider Performance Incentive Plan (IPPIP)
Minnesota ACO Demonstration
Oregon Coordinated Care

## Bundled Payment Programs

CMS ACE Demonstration
CMS Bundled Payment for Care Improvement
Aetna-Hoag Bundled Payment (IHA Initiative)
United Healthcare Oncology Episodes of Care Pilot
WellPoint/Anthem Bundled Payment
Blue Cross Blue Shield of Tennessee Orthopedic Bundled Payment Program
Blue Shield of California–Sutter Health (IHA Initiative)
Cox Health Plans–Chronic Disease Episode of Care Pilot
Geisinger Proven Care for CABG
Horizon Blue Cross Blue Shield New Jersey–Episode of Care Program
Arkansas (Also includes Arkansas Blue Cross Blue Shield, Arkansas QualChoice)

# Appendix B: Program Design and Context Variables[13]

## *VBP Approaches and Program Design Features*

- Structure of incentive
  - Frequency (e.g., annual, per service)
  - Magnitude (revenue potential)
  - Type of incentive (e.g., bonus, increase on FFS/per diem/DRG payment, penalty, public reporting of performance, shared savings, upside/downside risk)
    - Use of nonfinancial incentives
    - Form of financing (e.g., withhold, new dollars, based on savings)
  - Types of benchmarks/thresholds
    - Absolute performance (percentile ranking or fixed threshold)
    - Relative performance
  - Improvement (continuous)
  - Target of the incentive
- Measures
  - Number of measures
  - Types of measures (structure, process, outcomes; cost or quality)
    - Difficulty of measure (to achieve success)—does it require patient cooperation?
    - Perceived attainability (is performance within the provider's control?)
  - Baseline performance on chosen measures
    - Use of risk/case-mix adjustment (and adjusted for what factors?)
    - Attribution method
- Sponsor of the incentive (plan, purchaser, medical group)
- Technical support provided by VBP sponsor
  - Data transparency with providers (variations analyses)/use of performance feedback
  - Case management and care coordination resources
  - Sharing of best practices/learning networks
  - Coaching/training
  - Other technical assistance
- Overall approach to paying for services (base payment model on which the VBP program operates)—FFS, capitation, global payment (hospital and physician), bundled payment
  - Consumer incentives and engagement strategies

## *Characteristics of the Providers and Practice Settings*

- Populations served (payer mix, patient characteristics including socioeconomic mix, insurance status, age, clinical conditions)
- Incentive mix at different levels of the provider organization (capitation, salary, FFS)
- Extent to which provider network is restricted

- Size of the provider (i.e., number of beds, number of patients in panel)
- Percentage of provider's patients for whom the incentive is relevant
- Organization structure/ type (e.g., integrated medical group, independent practice association, primary care practice site, medical home, etc.)
- Organizational culture, leadership
- Use of peer pressure
- Extent of provider integration within a delivery system
- Health information technology use (extent, types)
- Other incentives faced by the provider (e.g., for utilization) and magnitude of those incentives
- Cost of compliance/improving quality (versus the incentives offered)
- Clinician characteristics (e.g., specialties, age, gender)
- Use of guidelines by provider
- Participation in external QI collaboratives

### *External Factors*

- Market characteristics (e.g., market concentration/competitiveness, number of payers, market share of each payer)
- Exposure to other VBP programs and quality initiatives across payers in a market (mix of incentives in the market)
- Alignment of measures across VBP programs within a market
- Regulatory features
- Anticipation of future policy trends (momentum regarding the inevitability of VBP)

# References


1. Hussey PS, Mulcahy AW, Schnyer C, Schneider EC. Bundled payment: Effects on health care spending and quality. Rockville, MD: Agency for Healthcare Research and Quality 2012.

2. Winslow R. HMO Juggernaut: U.S. healthcare cuts costs, grows rapidly and irks some doctors—a few patients are slighted, squeezed specialists say; but firm also gets praise—how Katie avoids hospital. Wall Street Journal. Eastern Edition. 1994 September 6.

3. Curtin K, Beckman H, Pankow G, Milillo Y, Green RA. Return on investment in pay for performance: A diabetes case study. Journal of Healthcare Management. 2006 Nov–Dec;51(6):365–74; discussion 75–76.

4. Damberg CL, Raube K, Teleki SS, Dela Cruz E. Taking stock of pay-for-performance: A candid assessment from the front lines. Health Affairs (Millwood). 2009 Mar–Apr;28(2):517–525.

5. Pearson SD, Schneider EC, Kleinman KP, Coltin KL, Singer JA. The impact of pay-for-performance on health care quality in Massachusetts, 2001–2003. Health Affairs. 2008 Jul–Aug;27(4):1167–1176.

6. Stecher BM, Camm F, Damberg CL, Hamilton LS, Mullen KJ, Nelson C, Sorensen P, Wachs M, Yoh A, Zellman GL, Leuschner KJ. Toward a Culture of Consequences: Performance-Based Accountability Systems for Public Services. Santa Monica, CA: RAND Corporation. RR-306/1-ASPE. 2010.

7. Damberg CL, Shortell SM, Raube K, Gillies RR, Rittenhouse D, McCurdy RK, Casalino LP, Adams J. Relationship between quality improvement processes and clinical performance. American Journal of Managed Care. 2010 Aug;16(8):601–606.

8. Young GJ, Meterko M, Beckman H, Baker E, White B, Sautter KM, Greene R, Curtin K, Bokhour BG, Berlowitz D, Burgess JF, Jr. Effects of paying physicians based on their relative performance for quality. Journal of General Internal Medicine. 2007 Jun;22(6):872–876.

9. Christianson JB, Leatherman S, Sutherland K. Lessons from evaluations of purchaser pay-for-performance programs: A review of the evidence. Medical Care Research and Review. 2008 Dec;65(6 Suppl):5S–35S.

10. Rosenthal MB, Frank RG, Li Z, Epstein AM. Early experience with pay-for-performance: From concept to practice. JAMA. 2005 Oct 12;294(14):1788–1793.

11. 111th United States Congress. Patient Protection and Affordable Care Act Pub. L. 111-148. Washington, DC. 2010.

12. Dudley RA, Frolich A, Robinowitz DL, Talavera JA, Broadhead P, Luft HS. Strategies to support quality-based purchasing: A review of the evidence. Technical Review 10. (Prepared by the Stanford-University of California San Francisco Evidence-based Practice Center under Contract No. 290-02-0017). AHRQ Publication No. 04-0057. Rockville, MD: Agency for Healthcare Research and Quality 2004 Jul.

13. Dudley RA. Pay-for-performance research: how to learn what clinicians and policy makers need to know. JAMA. 2005 Oct 12;294(14):1821–1823.

14. Damberg CL, Sorbero ME, Lovejoy S, Martsolf GR, Raaen L, Mandel D. Measuring Success in Health Care Value-Based Purchasing Programs: Summary and Recommendations, Santa Monica, CA: RAND Corporation. RR-306/1-ASPE. 2013.

15. McGlynn EA, Asch SM, Adams J, Keesey J, Hicks J, DeCristofaro A, Kerr EA. The quality of health care delivered to adults in the United States. New England Journal of Medicine. 2003 Jun 26;348(26):2635–2645.

16. Centers for Medicare and Medicaid Services. National Provider Call: Hospital value-Based Purchasing: Fiscal Year 2015 Overview for Beneficiaries, Providers, and Stakeholders. Baltimore, MD: Centers for Medicare and Medicaid Services. 2013.

17. Health Services Advisory Group. Is Your Hospital Ready for Value-Based Purchasing? Phoenix, AZ: Health Services Advisory Group; 2013 [cited 2013 November 4]. As of November 21, 2013: http://www.hsag.com/App_Resources/Documents/VBP_factsheet_-FLCA_508.pdf.

18. Tak HJ, Ruhnke GW, Meltzer DO. Association of patient preferences for participation in decision making with length of stay and costs among hospitalized patients. JAMA Internal Medicine. 2013 Jul 8;173(13):1195–1205.

19. Damberg CL, Sorbero ME, Mehrotra A, Teleki S, Lovejoy S, Bradley L. An Environmental Scan of Pay for Performance in the Hospital Setting: Final Report. Washington, DC: Office of the Assistant Secretary for Planning and Evaluation (ASPE) 2007.

20. Sorbero MES, Damberg CL, Shaw R, Teleki S, Lovejoy S, Decristofaro A, Dembosky J, Schuster C. Assessment of Pay-for-Performance Options for Medicare Physician Services: Final Report. Santa Monica, CA: RAND Corporation. WR-391-ASPE. 2006. As of November 21, 2013: http://www.rand.org/pubs/working_papers/WR391.html

21. Doran T, Fullwood C, Reeves D, Gravelle H, Roland M. Exclusion of patients from pay-for-performance targets by English physicians. New England Journal of Medicine. 2008 Jul 17;359(3):274–284.

22. Chien AT, Li Z, Rosenthal MB. Improving timely childhood immunizations through pay for performance in Medicaid-managed care. Health Services Research. 2010 Dec;45(6 Pt 2):1934–1947.

23. Fairbrother G, Siegel MJ, Friedman S, Kory PD, Butts GC. Impact of financial incentives on documented immunization rates in the inner city: results of a randomized controlled trial. Ambulatory Pediatrics. 2001 Jul–Aug;1(4):206–212.

24. Serumaga B, Ross-Degnan D, Avery AJ, Elliott RA, Majumdar SR, Zhang F, Soumerai SB. Effect of pay for performance on the management and outcomes of hypertension in the United Kingdom: Interrupted time series study. BMJ. 2011;342:d108.

25. Gilmore AS, Zhao Y, Kang N, Ryskina KL, Legorreta AP, Taira DA, Chung RS. Patient outcomes and evidence-based medicine in a preferred provider organization setting: a six-year evaluation of a physician pay-for-performance program. Health Services Research. 2007 Dec;42(6 Pt 1):2140–2159; discussion 294–323.

26. Chung RS, Chernicoff HO, Nakao KA, Nickel RC, Legorreta AP. A quality-driven physician compensation model: Four-year follow-up study. Journal for Healthcare Quality. 2003 Nov–Dec;25(6):31–37.

27. Coleman K, Reiter KL, Fulwiler D. The impact of pay-for-performance on diabetes care in a large network of community health centers. Journal of Health Care for the Poor and Underserved. 2007 Nov;18(4):966–983.

28. Cutler TW, Palmieri J, Khalsa M, Stebbins M. Evaluation of the relationship between a chronic disease care management program and california pay-for-performance diabetes care cholesterol measures in one medical group. Journal of Managed Care Pharmacy. 2007 Sep;13(7):578–588.

29. Larsen DL, Cannon W, Towner S. Longitudinal Assessment of a Diabetes Care Management System in an Integrated Health Network. Journal of Managed Care Pharmacy. 2003;9(6):552–558.

30. Amundson G, Solberg LI, Reed M, Martini EM, Carlson R. Paying for quality improvement: compliance with tobacco cessation guidelines. Joint Commission Journal on Quality and Patient Safety. 2003 Feb;29(2):59–65.

31. Hung DY, Green LA. Paying for prevention: associations between pay for performance and cessation counseling in primary care practices. American Journal of Health Promotion. 2012 Mar–Apr;26(4):230–234.

32. Armour BS, Friedman C, Pitts MM, Wike J, Alley L, Etchason J. The influence of year-end bonuses on colorectal cancer screening. American Journal of Managed Care. 2004 Sep;10(9):617–624.

33. Chung S, Palaniappan LP, Trujillo LM, Rubin HR, Luft HS. Effect of physician-specific pay-for-performance incentives in a large group practice. American Journal of Managed Care. 2010 Feb;16(2):e35–e42.

34. Pourat N, Rice T, Tai-Seale M, Bolan G, Nihalani J. Association between physician compensation methods and delivery of guideline-concordant STD care: Is there a link? American Journal of Managed Care. 2005 Jul;11(7):426–432.

35. Greene RA, Beckman H, Chamberlain J, Partridge G, Miller M, Burden D, Kerr J. Increasing adherence to a community-based guideline for acute sinusitis through education, physician profiling, and financial incentives. American Journal of Managed Care. 2004 Oct;10(10):670–678.

36. Mandel KE, Kotagal UR. Pay for performance alone cannot drive quality. Archives of Pediatrics and Adolescent Medicine. 2007 Jul;161(7):650–655.

37. Unutzer J, Chan YF, Hafer E, Knaster J, Shields A, Powers D, Veith RC. Quality improvement with pay-for-performance incentives in integrated behavioral health care. American Journal of Public Health. 2012 Jun;102(6):e41–e45.

38. Collier VU. Use of pay for performance in a community hospital private hospitalist group: a preliminary report. Transactions of the American Clinical and Climatological Association. 2007;118:263–272.

39. Leitman IM, Levin R, Lipp MJ, Sivaprasad L, Karalakulasingam CJ, Bernard DS, Friedmann P, Shulkin DJ. Quality and financial outcomes from gainsharing for inpatient admissions: A three-year experience. Journal of Hospital Medicine. 2010 Nov–Dec;5(9):501–507.

40. Fagan PJ, Schuster AB, Boyd C, Marsteller JA, Griswold M, Murphy SM, Dunbar L, Forrest CB. Chronic care improvement in primary care: Evaluation of an integrated pay-for-performance and practice-based care coordination program among elderly patients with diabetes. Health Services Research. 2010 Dec;45(6 Pt 1):1763–1782.

41. Beaulieu ND, Horrigan DR. Putting smart money to work for quality improvement. Health Services Research. 2005 Oct;40(5 Pt 1):1318–1334.

42. Mullen KJ, Frank RG, Rosenthal MB. Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. Rand Journal of Economics. 2010 Spring;41(1):64–91.

43. 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? Journal of General Internal Medicine. 2012 May;27(5):548–554.

44. Felt-Lisk S, Gimm G, Peterson S. Making pay-for-performance work in Medicaid. Health Affairs. 2007 Jul–Aug;26(4):w516–w527.

45. Levin-Scherz J, DeVita N, Timbie J. Impact of pay-for-performance contracts and network registry on diabetes and asthma HEDIS measures in an integrated delivery network. Medical Care Research and Review. 2006 Feb;63(1 Suppl):14S–28S.

46. Lester H, Schmittdiel J, Selby J, Fireman B, Campbell S, Lee J, Whippy A, Madvig P. The impact of removing financial incentives from clinical quality indicators: Longitudinal analysis of four Kaiser Permanente indicators. BMJ. 2010;340:c1898.

47. Roski J, Jeddeloh R, An L, Lando H, Hannan P, Hall C, Zhu SH. The impact of financial incentives and a patient registry on preventive care quality: Increasing provider adherence to evidence-based smoking cessation practice guidelines. Preventive Medicine. 2003 Mar;36(3):291–299.

48. Chen JY, Tian H, Taira Juarez D, Hodges KA, Jr., Brand JC, Chung RS, Legorreta AP. The effect of a PPO pay-for-performance program on patients with diabetes. American Journal of Managed Care. 2010 Jan;16(1):e11–e19.

49. An LC, Bluhm JH, Foldes SS, Alesci NL, Klatt CM, Center BA, Nersesian WS, Larson ME, Ahluwalia JS, Manley MW. A randomized trial of a pay-for-performance program targeting clinician referral to a state tobacco quitline. Archives of Internal Medicine. 2008 Oct 13;168(18):1993–1999.

50. Chen JY, Kang N, Juarez DT, Hodges KA, Chung RS, Legorreta AP. Impact of a pay-for-performance program on low performing physicians. Journal for Healthcare Quality. 2010 Jan–Feb;32(1):13–21; quiz -2.

51. Gavagan TF, Du H, Saver BG, Adams GJ, Graham DM, McCray R, Goodrick GK. Effect of financial incentives on improvement in medical quality indicators for primary care. The Journal of the American Board of Family Medicine. 2010 Sep–Oct;23(5):622–631.

52. Rosenthal MB, de Brantes FS, Sinaiko AD, Frankel M, Robbins RD, Young S. Bridges to Excellence--recognizing high-quality care: Analysis of physician quality and resource use. American Journal of Managed Care. 2008 Oct;14(10):670–677.

53. Glickman SW, Ou FS, DeLong ER, Roe MT, Lytle BL, Mulgund J, Rumsfeld JS, Gibler WB, Ohman EM, Schulman KA, Peterson ED. Pay for performance, quality of care, and outcomes in acute myocardial infarction. JAMA. 2007 Jun 6;297(21):2373–2380.

54. Nicholas LH, Dimick JB, Iwashyna TJ. Do hospitals alter patient care effort allocations under pay-for-performance? Health Services Research. 2011 Feb;46(1 Pt 1):61–81.

55. Ryan AM, Blustein J. The effect of the MassHealth hospital pay-for-performance program on quality. Health Services Research. 2011 Jun;46(3):712–728.

56. Werner RM, Kolstad JT, Stuart EA, Polsky D. The effect of pay-for-performance in hospitals: Lessons for quality improvement. Health Affairs. 2011 Apr;30(4):690–698.

57. Calikoglu S, Murray R, Feeney D. Hospital pay-for-performance programs in Maryland produced strong results, including reduced hospital-acquired conditions. Health Affairs. 2012 Dec;31(12):2649–2658.

58. 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 Services Research. 2012 Aug;47(4):1418–1436.

59. Lindenauer PK, Remus D, Roman S, Rothberg MB, Benjamin EM, Ma A, Bratzler DW. Public reporting and pay for performance in hospital quality improvement. New England Journal of Medicine. 2007;356(5):486–496.

60. Herrin J, Nicewander D, Ballard DJ. The effect of health care system administrator pay-for-performance on quality of care. Joint Commission Journal on Quality and Patient Safety. 2008 Nov;34(11):646–654.

61. Shepard DS, Calabro JA, Love CT, McKay JR, Tetreault J, Yeom HS. Counselor incentives to improve client retention in an outpatient substance abuse aftercare program. Administration and Policy in Mental Health. 2006 Nov;33(6):629–635.

62. Colla CH, Wennberg DE, Meara E, Skinner JS, Gottlieb D, Lewis VA, Snyder CM, Fisher ES. Spending differences associated with the Medicare Physician Group Practice Demonstration. JAMA. 2012 Sep 12;308(10):1015–1023.

63. Markovich P. A global budget pilot project among provider partners and Blue Shield of California led to savings in first two years. Health Affairs. 2012 Sep;31(9):1969–1976.

64. Salmon RB, Sanderson MI, Walters BA, Kennedy K, Flores RC, Muney AM. A collaborative accountable care model in three practices showed promising early results on costs and quality of care. Health Affairs. 2012 Nov;31(11):2379–2387.

65. Song Z, Safran DG, Landon BE, He Y, Ellis RP, Mechanic RE, Day MP, Chernew ME. Health care spending and quality in year 1 of the alternative quality contract. New England Journal of Medicine. 2011 Sep 8;365(10):909–918.

66. Song Z, Safran DG, Landon BE, Landrum MB, He Y, Mechanic RE, Day MP, Chernew ME. The 'Alternative Quality Contract,' based on a global budget, lowered medical spending and improved quality. Health Affairs. 2012 Aug;31(8):1885–1894.

67. Centers for Medicare and Medicaid Services. Press release: Pioneer Accountable Care Organizations succeed in improving care, lowering costs. Baltimore, MD: Centers for Medicare and Medicaid Services; 2013 [cited 2013 August 16]. As of November 21, 2013: http://www.cms.gov/Newsroom/MediaReleaseDatabase/Press-Releases/2013-Press-Releases-Items/2013-07-16.html.

68. Casale AS, Paulus RA, Selna MJ, Doll MC, Bothe Jr AE, McKinley KE, Berry SA, Davis DE, Gilfillan RJ, Hamory BH. "ProvenCareSM": A provider-driven pay-for-performance program for acute episodic cardiac surgical care. Annals of Surgery. 2007;246(4):613–623.

69. Chien AT, Eastman D, Li Z, Rosenthal MB. Impact of a pay for performance program to improve diabetes care in the safety net. Preventive Medicine. 2012 Nov;55 Suppl:S80–S85.

70. Rosenthal MB, Li Z, Robertson AD, Milstein A. Impact of financial incentives for prenatal care on birth outcomes and spending. Health Services Research. 2009 Oct;44(5 Pt 1):1465–1479.

71. Ryan AM. Effects of the Premier Hospital Quality Incentive Demonstration on Medicare patient mortality and cost. Health Services Research. 2009 Jun;44(3):821–842.

72. Sutton M, Nikolova S, Boaden R, Lester H, McDonald R, Roland M. Reduced mortality with hospital pay for performance in England. New England Journal of Medicine. 2012 Nov 8;367(19):1821–1828.

73. Jha AK, Joynt KE, Orav EJ, Epstein AM. The long-term effect of premier pay for performance on patient outcomes. New England Journal of Medicine. 2012 Apr 26;366(17):1606–1615.

74. Werner RM, Rita T, Kim M. Quality improvement under nursing home compare: The association between changes in process and outcome measures. Medical care. 2013;51(7):582–588.

75. Hittle D, Nuccio E, Richard A. Evaluation of the Medicare Home Health Pay-for-Performance Demonstration: CY2008 Report—Volume 1: Agency Characteristics, Costs, and Quality Measure Performance among Treatment, Control, and Non-Participant Groups. 2011.

76. Shen Y. Selection incentives in a performance-based contracting system. Health Services Research. 2003;38(2):535–252.

77. Kruse GB, Polsky D, Stuart EA, Werner RM. The impact of hospital pay-for-performance on hospital and Medicare costs. Health Services Research. 2012 Oct 22.

78. Ryan AM, Burgess JF, Jr., Tompkins CP, Wallack SS. The relationship between Medicare's process of care quality measures and mortality. Inquiry. 2009 Fall;46(3):274–290.

79. Zucker M, editor. Case Study 1: Prospective Payment for Medicare Parts A and B During Hospitalization (ACE Demo). Episode Payment: Private Innovation and Opportunities for Medicare. Washington, DC: Brandeis University. 2011.

80. Beard AJ, Hofer TP, Downs JR, Lucatorto M, Klamerus ML, Holleman R, Kerr EA. Assessing appropriateness of lipid management among patients with diabetes mellitus: Moving from target to treatment. Circulation: Cardiovascular Quality and Outcomes. 2013 Jan 1;6(1):66–74.

81. Kerr EA, Hayward RA. Patient-centered performance management: Enhancing value for patients and health care systems. JAMA. 2013 Jul 10;310(2):137–138.

82. Kerr EA, Lucatorto MA, Holleman R, Hogan MM, Klamerus ML, Hofer TP. Monitoring performance for blood pressure management among patients with diabetes mellitus: Too much of a good thing? Archives of Internal Medicine. 2012;172(12):938–945.

83. Friedberg MW, Mehrotra A, Linder JA. Reporting hospitals' antibiotic timing in pneumonia: Adverse consequences for patients? American Journal of Managed Care. 2009 Feb;15(2):137–144.

84. Campbell B, Marchildon GP. Table of contents in Medicare: Facts, Myths, Problems, Promise. J. Lorimer & Co.; 2007 [cited WorldCat. ACO 20121106]. As of November 21, 2013:
http://catdir.loc.gov/catdir/toc/fy0803/2008360115.html

85. Campbell SM, Reeves D, Kontopantelis E, Sibbald B, Roland M. Effects of pay for performance on the quality of primary care in England. N Engl J Med. 2009 Jul 23;361(4):368–378.

86. Healy D, Cromwell J. Hospital-acquired conditions—present on admission: Examination of spillover effects and unintended consequences. Baltimore, MD: Centers for Medicare and Medicaid Services. 2012.

87. McWilliams JM, Landon BE, Chernew ME. Changes in health care spending and quality for Medicare beneficiaries associated with a commercial ACO contract. JAMA. 2013 Aug 28;310(8):829–836.

88. Jha AK, Orav EJ, Epstein AM. The effect of financial incentives on hospitals that serve poor patients. Annals of Internal Medicine. 2010 Sep 7;153(5):299–306.

89. Ryan AM. Has pay-for-performance decreased access for minority patients? Health Services Research. 2010 Feb;45(1):6–23.

90. 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 Affairs. 2012 Apr;31(4):797–805.

91. 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 Aug 30;372(9640):728–736.

92. American Medical Group Association. High-performing health system definition. Alexandria, VA: American Medical Group Association. 2013 [cited August 20, 2013]. As of November 21, 2013:
http://www.amga.org/Advocacy/HPHS/hphsDefinitionHandout.pdf

93. Beich J, Scanlon DP, Ulbrecht J, Ford EW, Ibrahim IA. The role of disease management in pay-for-performance programs for improving the care of chronically ill patients. Medical Care Research and Review. 2006 Feb;63(1 Suppl):96S–116S.

94. Conrad DA, Christianson JB. Penetrating the "black box": Financial incentives for enhancing the quality of physician services. Medical Care Research and Review. 2004 Sep;61(3 Suppl):37S–68S.

95. Frolich A, Talavera JA, Broadhead P, Dudley RA. A behavioral model of clinician responses to incentives to improve quality. Health Policy. 2007 Jan;80(1):179–193.

96. Rosenthal MB, Frank RG. What is the empirical basis for paying for quality in health care? Medical Care Research and Review. 2006 Apr;63(2):135–157.

97. Vina ER, Rhew DC, Weingarten SR, Weingarten JB, Chang JT. Relationship between organizational factors and performance among pay-for-performance hospitals. Journal of General Internal Medicine. 2009 Jul;24(7):833–480.

98. Yale New Haven Health Services Corporation. Medicare Hospital Quality Chartbook: Performance Report on Outcome Measures. Baltimore, MD: Centers for Medicare and Medicaid Services. 2012.

99. Lake TK, Stewart KA, Ginsburg PB. Lessons from the field making accountable care organizations real. [Book; Computer File; Internet Resource Date of Entry: 20110708]: Washington, D.C.: Center for Studying Health System Change. 2011 [cited 2011 July 8]; 7, [1] p. : digital, PDF file, ill.].

100. Moore K, Coddington D. From Volume to Value: The Transition to Accountable Care Organizations. Greenwood Village: McCannis Consulting. 2011.

101. Shields MC, Patel PH, Manning M, Sacks L. A model for integrating independent physicians into accountable care organizations. Health Affairs. 2011 Jan;30(1):161–172.

102. Shortell SM, Casalino LP. Implementing qualifications criteria and technical assistance for accountable care organizations. JAMA. 2010 May 5;303(17):1747–1748.

103. 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 Services Research. 2010 Apr;45(2):553–564.

104. Hussey PS, Ridgely MS, Rosenthal MB. The PROMETHEUS bundled payment experiment: Slow start shows problems in implementing new payment models. Health Affairs. 2011 Nov;30(11):2116–2124.

105. Arling G, Job C, Cooke V. Medicaid nursing home pay for performance: Where do we stand? Gerontologist. 2009 Oct;49(5):587–595.

106. Ragin CC. Redesigning social inquiry: Fuzzy sets and beyond. Chicago and London: University of Chicago Press; 2008.

107. Ragin CC. Using qualitative comparative analysis to study causal complexity. Health Services Research. 1999 Dec;34(5 Pt 2):1225–1239.

108. Mehrotra A, Damberg CL, Sorbero ME, Teleki SS. Pay for performance in the hospital setting: What is the state of the evidence? American Journal of Medical Quality. 2009 Jan–Feb;24(1):19–28.

109. Jackson MS. Mulling over Massachusetts health insurance mandates and entrepreneurs. [Internet Resource; Computer File Date of Entry: 20080709]: Fairfax, VA : George Mason University. 2008. x, 208 p. digital, PDF file, col. ill., col. maps. Dissertation: Thesis (Ph.D.)—George Mason University, 2008.]. As of November 21, 2013: http://hdl.handle.net/1920/3056

110. Jha AK. Measuring hospital quality: What physicians do? How patients fare? Or both? JAMA. 2006 Jul 5;296(1):95–97.

111. Jha AK, Orav EJ, Epstein AM. Low-Quality, High-Cost Hospitals, Mainly In South, Care For Sharply Higher Shares Of Elderly Black, Hispanic, And Medicaid Patients. Health Affairs. 2011 October 1, 2011;30(10):1904–1911.

112. Adams J, Mehrotra A, Thomas J, McGlynn E. Physician Cost Profiling—Reliability and Risk of Misclassification: Detailed Methodology and Sensitivity Analyses. Santa Monica, CA: RAND Corporation. TR-799-DOL. 2010. As of November 21, 2013: http://www.rand.org/pubs/technical_reports/TR799.html

113. Scholle SH, Roski J, Adams JL, Dunn DL, Kerr EA, Dugan DP, Jensen RE. Benchmarking physician performance: Reliability of individual and composite measures. American Journal of Managed Care. 2008 Dec;14(12):833–838.

114. Sequist TD, Schneider EC, Li A, Rogers WH, Safran DG. Reliability of medical group and physician performance measurement in the primary care setting. Medical Care. 2011 Feb;49(2):126–31.

115. National Quality Strategy. [cited 2013 July 15]. As of November 21, 2013: http://www.ahrq.gov/workingforquality/reports.htm.

116. U.S. Department of Health and Human Service. 2012 Annual Progress Report to Congress. National Strategy for Quality Improvement in Health Care. Washington, DC: U.S. Department of Health and Human Service April 2012 (Corrected August 2012).

117. McHugh M, Joshi M. Improving evaluations of value-based purchasing programs. Health Services Research. 2010 Oct;45(5 Pt 2):1559–1569.

118. Fisher ES, Shortell SM, Kreindler SA, Van Citters AD, Larson BK. A framework for evaluating the formation, implementation, and performance of accountable care organizations. Health Affairs. 2012 Nov;31(11):2368–2378.

119. Hussey PS, Sorbero ME, Mehrotra A, Liu H, Damberg CL. Episode-based performance measurement and payment: Making it a reality. Health Affairs. 2009 Sep–Oct;28(5):1406–1417.

120. Agency for Healthcare Research and Quality. Methods Reference Guide for Effectiveness and Comparative Effectiveness Reviews, Version 1.0. Rockville, MD: Agency for Healthcare Research and Quality 2008.

121. Schneider EC, Hussey PS, Schnyer C. Payment Reform: Analysis of Models and Performance Measurement Implications. Santa Monica, CA: RAND Corporation. TR-841-NQF. 2011. As of November 21, 2013: http://www.rand.org/pubs/technical_reports/TR841.html

122. Med-Vantage. 2010 National P4P Survey Executive Summary. Danbury, CT: IMS Health Incorporated. 2011. As of November 21, 2013: http://www.imshealth.com/deployedfiles/ims/Global/Content/Solutions/Healthcare%20Analytics%20and%20Services/Payer%20Solutions/Survey_Exec_Sum.pdf

123. OCS HomeCare. Value Based Purchasing in Skilled Nursing: A Discussion of Current Trends and Initiatives. Seattle, WA: National Research Corporation. 2011. As of November 21, 2013: http://www.ocshomecare.com/Resources/White-Papers/Value-Based-Purchasing-in-Skilled-Nursing.aspx

124. Centers for Medicare and Medicaid Services. National Health Care Expenditures Data. Baltimore, MD: Office of the Actuary, National Health Statistics Group; 2012 [cited 2013 March 30]. As of November 21, 2013: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/downloads/tables.pdf

125. Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S. Does pay-for-performance improve the quality of health care? Annals of Internal Medicine. 2006 Aug 15;145(4):265–272.

126. Armour BS, Pitts MM, Maclean R, Cangialose C, Kishel M, Imai H, Etchason J. The effect of explicit financial incentives on physician behavior. Archives of Internal Medicine. 2001 May 28;161(10):1261–1266.

127. 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 May;110(2–3):115–130.

128. Emmert M, Eijkenaar F, Kemter H, Esslinger AS, Schoffski O. Economic evaluation of pay-for-performance in health care: A systematic review. The European Journal of Health Economics. 2012 Dec;13(6):755–767.

129. Greene SE, Nash DB. Pay for performance: An overview of the literature. American Journal of Medical Quality. 2009 Mar–Apr;24(2):140–163.

130. 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 Services Research. 2010;10(1):247.

131. 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 Affairs. 2009 Mar–Apr;28(2):526–32.

132. Stefan MS, Pekow PS, Nsa W, Priya A, Miller LE, Bratzler DW, Rothberg MB, Goldberg RJ, Baus K, Lindenauer PK. Hospital performance measures and 30-day readmission rates. Journal of General Internal Medicine. 2013 Mar;28(3):377–385.

133. Nicholas LH, Osborne NH, Birkmeyer JD, Dimick JB. Hospital process compliance and surgical outcomes in medicare beneficiaries. Archives of Surgery. 2010 Oct;145(10):999–1004.

134. Bradley E, Herrin J, Elbel B, McNamara R, Magid D, Nallamothu B, Wang Y, Normand S, Spertus J, Krumholz H. Hospital quality for acute myocardial infarction: Correlation among process measures and relationship with short-term mortality. JAMA. 2006;296(1):72–78.

135. Werner RM, Bradlow ET. Relationship between Medicare's hospital compare performance measures and mortality rates. JAMA. 2006 Dec 13;296(22):2694–2702.

136. Sidorenkov G, Haaijer-Ruskamp FM, de Zeeuw D, Bilo H, Denig P. Review: Relation between quality-of-care indicators for diabetes and patient outcomes: A systematic literature review. Medical Care Research and Review. 2011 Jun;68(3):263–289.

137. Ryan AM, Doran T. The effect of improving processes of care on patient outcomes: Evidence from the United Kingdom's quality and outcomes framework. Medical care. 2012 Mar;50(3):191–199.

138. Kralewski JE, Dowd BE, Xu YW. Medical groups can reduce costs by investing in improved quality of care for patients with diabetes. Health Affairs. [Research Support, Non-U.S. Gov't]. 2012 Aug;31(8):1830–1835.

139. Glickman SW, Boulding W, Roos JM, Staelin R, Peterson ED, Schulman KA. Alternative pay-for-performance scoring methods: Implications for quality improvement and patient outcomes. Medical care. 2009 Oct;47(10):1062–1068.

140. Jha AK, Orav EJ, Li Z, Epstein AM. The Inverse Relationship Between Mortality Rates And Performance In The Hospital Quality Alliance Measures. Health Affairs. 2007 July 1, 2007;26(4):1104–1110.

141. Krumholz H, Lin Z, Keenan P, Chen J, Ross J, Drye E, Bernheim S, Wang Y, Bradley E, Han L, Normand S. Relationship between hospital readmission and mortality rates for patients hospitalized with acute myocardial infarction, heart failure, or pneumonia. JAMA. 2013;309(6):587–593.

142. Popescu I, Werner RM, Vaughan-Sarrazin MS, Cram P. Characteristics and outcomes of America's lowest-performing hospitals: An analysis of acute myocardial infarction hospital care in the United States. Circulation: Cardiovascular Quality and Outcomes. 2009 May;2(3):221–227.

143. Quattromani E, Powell ES, Khare RK, Cheema N, Sauser K, Periyanayagam U, Pirotte MJ, Feinglass J, Mark Courtney D. Hospital-reported data on the pneumonia quality measure “Time to First Antibiotic Dose” are not associated with inpatient mortality: Results of a nationwide cross-sectional analysis. Academic Emergency Medicine. 2011 May;18(5):496–503.

144. Ashton CM, Kuykendall DH, Johnson ML, Wray NP, Wu L. The association between the quality of inpatient care and early readmission. Annals of Internal Medicine. 1995 Mar 15;122(6):415–421.

145. Li S, Liu J, Gilbertson D, McBean M, Dowd B, Collins A. An instrumental variable analysis of the impact of practice guidelines on improving quality of care and diabetes-related outcomes in the elderly Medicare population. American Journal of Medical Quality. 2008 May–Jun;23(3):222–230.

146. Harman JS, Scholle SH, Ng JH, Pawlson LG, Mardon RE, Haffer SC, Shih S, Bierman AS. Association of Health Plans' Healthcare Effectiveness Data and Information Set (HEDIS) performance with outcomes of enrollees with diabetes. Medical care. 2010 Mar;48(3):217–223.

147. 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 Sep 11;310(10):1051–1059.

148. Petersen LA, Simpson K, Pietz K, Urech TH, Hysong SJ, Profit J, Conrad DA, Dudley RA, Woodard LD. Effects of individual physician-level and practice-level financial incentives on hypertension care: A randomized trial. JAMA. 2013 Sep 11;310(10):1042–1050.

149. Chen JY, Tian H, Juarez DT, Yermilov I, Braithwaite RS, Hodges KA, Legorreta A, Chung RS. Does pay for performance improve cardiovascular care in a “real-world” setting? American Journal of Medical Quality. 2011 Sep–Oct;26(5):340–348.

150. Young G, Meterko M, White B, Sautter K, Bokhour B, Baker E, Silver J. Pay-for-performance in safety net settings: Issues, opportunities, and challenges for the future. Journal of Healthcare Management. 2010 Mar–Apr;55(2):132–141; discussion 41–42.

151. Larson BK, Van Citters AD, Kreindler SA, Carluzzo KL, Gbemudu JN, Wu FM, Nelson EC, Shortell SM, Fisher ES. Insights from transformations under way at four brookings-dartmouth accountable care organization pilot sites. Health Affairs. 2012 Nov;31(11):2395–2406.

152. Premier. CMS/Premier Hospital Quality Incentive Demonstration (HQID). Charlotte, NC Premier. 2013 [cited 2013 July 10]. As of November 21, 2013: https://www.premierinc.com/p4p/hqi/

153. Grossbart SR. What's the return? Assessing the effect of "pay-for-performance" initiatives on the quality of care delivery. Medical Care Research and Review. 2006 Feb;63(1 Suppl):29S–48S.

154. Atkinson JG, Masiulis KE, Felgner L, Schumacher DN. Provider-initiated pay-for-performance in a clinically integrated hospital network. Journal for Healthcare Quality. 2010 Jan–Feb;32(1):42–50; quiz

155. Berthiaume JT, Chung RS, Ryskina KL, Walsh J, Legorreta AP. Aligning financial incentives with quality of care in the hospital setting. Journal for Healthcare Quality. 2006 Mar–Apr;28(2):36–44, 51.

156. Berthiaume JT, Tyler PA, Ng-Osorio J, LaBresh KA. Aligning financial incentives with "Get With The Guidelines" to improve cardiovascular care. American Journal of Managed Care. 2004 Jul;10(7 Pt 2):501–504.

157. Nahra TA, Reiter KL, Hirth RA, Shermer JE, Wheeler JR. Cost-effectiveness of hospital pay-for-performance incentives. Medical Care Research and Review. 2006 Feb;63(1 Suppl):49S–72S.

158. Mehrotra A, Pearson SD, Coltin KL, Kleinman KP, Singer JA, Rabson B, Schneider EC. The response of physician groups to P4P incentives. American Journal of Managed Care. 2007 May;13(5):249–255.

159. Campbell SM, Reeves D, Kontopantelis E, Sibbald B, Roland M. Effects of pay for performance on the quality of primary care in England. New England Journal of Medicine. 2009 Jul 23;361(4):368–378.

160. Drake DE, Cohen A, Cohn J. National hospital antibiotic timing measures for pneumonia and antibiotic overuse. Quality Management in Health Care. 2007 Apr–Jun;16(2):113–122.

161. McDonald R, Roland M. Pay for performance in primary care in England and California: Comparison of unintended consequences. The Annals of Family Medicine. 2009 Mar–Apr;7(2):121–127.

162. Werner RM, Goldman LE, Dudley RA. Comparison of change in quality of care between safety-net and non-safety-net hospitals. JAMA. 2008 May 14;299(18):2180–2187.

163. Weinick RM, Chien AT, Rosenthal MB, Bristol SJ, Salamon J. Hospital executives' perspectives on pay-for-performance and racial/ethnic disparities in care. Medical Care Research and Review. 2010 Oct;67(5):574–589.

164. 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. New England Journal of Medicine. 2006 Jul 27;355(4):375–384.

165. Pham HH, Coughlan J, O'Malley AS. The impact of quality-reporting programs on hospital operations. Health Affairs. 2006 Sep–Oct;25(5):1412–1422.

166. Muhlestein D. Why has ACO growth slowed? 2013 [cited 2013 October 31]. As of November 21, 2013: http://healthaffairs.org/blog/2013/10/31/why-has-aco-growth-slowed/

167. Sebelius K. Physician Group Practice Evaluation: Report to Congress. Washington DC: Department of Health and Human Services. 2009.

168. Audet AM, Kenward K, Patel S, Joshi MS. Hospitals on the path to accountable care: Highlights from a 2011 national survey of hospital readiness to participate in an accountable care organization. Issue Brief (Commonw Fund). 2012 Aug;22:1–12.

169. MacKinney AC, Mueller KJ, McBride TD. The march to accountable care organizations: How will rural fare? The Journal of Rural Health. 2011 Winter;27(1):131–137.

170. McClellan M, McKethan AN, Lewis JL, Roski J, Fisher ES. A national strategy to put accountable care into practice. Health Affairs. 2010 May;29(5):982–990.

171. Berkowitz SA, Miller ED. Accountable care at academic medical centers--lessons from Johns Hopkins. New England Journal of Medicine. 2011 Feb 17;364(7):e12.

172. Kastor JA. Accountable care organizations at academic medical centers. New England Journal of Medicine. 2011 Feb 17;364(7):e11.

173. Tallia AF, Howard J. An academic health center sees both challenges and enabling forces as it creates an accountable care organization. Health Affairs. 2012 Nov;31(11):2388–2394.

174. Higgins A, Stewart K, Dawson K, Bocchino C. Early lessons from accountable care models in the private sector: Partnerships between health plans and providers. Health Affairs. 2011 Sep;30(9):1718–1727.

175. Bailit M, Hughes C. Key design elements of shared-savings payment arrangements. Issue Brief (Commonw Fund). 2011 Aug;20:1–16.

176. Meyer H. Many accountable care organizations are now up and running, if not off to the races. Health Affairs. 2012 Nov;31(11):2363–2367.

177. Kreindler SA, Larson BK, Wu FM, Carluzzo KL, Gbemudu JN, Struthers A, Van Citters AD, Shortell SM, Nelson EC, Fisher ES. Interpretations of integration in early accountable care organizations. Milbank Quarterly. 2012 Sep;90(3):457–483.

178. Bachrach D, Bernstein W, Karl A. High-performance health care for vulnerable populations. New York, NY: The Commonwealth Fund. 2012.

179. Weissman JS, Bailit M, D'Andrea G, Rosenthal MB. The design and application of shared savings programs: lessons from early adopters. Health Affairs. 2012 Sep;31(9):1959–1568.

180. Kroch ER, Champion W, DeVore SD, Kugel MR, Lloyd DA, Rothney-Kozlak L. Measuring Progress Toward Accountable Care: Premier Research Institute. 2012.

181. Hester J, Lewis J, McKethan A. The Vermont accountable care organization pilot a community health system to control total medical costs and improve population health. [Book; Computer File; Internet Resource Date of Entry: 20100804]: New York : Commonwealth Fund. 2010; 22 p. : digital, PDF file, ill., maps.]. As of November 21, 2013: http://www.commonwealthfund.org/~/media/Files/Publications/Fund%20Report/2010/May/1403_Hester_Vermont_accountable_care_org_pilot.pdf

182. Fisher ES, McClellan MB, Safran DG. Building the path to accountable care. New England Journal of Medicine. 2011 Dec 29;365(26):2445–2447.

183. James MH. Navigating the road ahead: lessons from a pioneer ACO. Healthcare Financial Management. 2012 Aug;66(8):64–69.

184. Davis K, Schoenbaum SC. Toward high-performance accountable care: Promise and pitfalls. New York, NY: The Commonwealth Fund. 2010.

185. Guterman S, Schoenbaum SC, Davis K, Schoen C, Audet A-MJ, Stremikis K, Zezza MA. High performance accountable care building on success and learning from experience. [Book; Computer File; Internet Resource Date of Entry: 20110621]: [New York, N.Y.] : Commonwealth Fund. 2011 [cited 2011 June 21]; xvii, 40 p. : digital, PDF file, ill.]. As of November 21, 2013: http://www.commonwealthfund.org/~/media/Files/Publications/Fund%20Report/2011/Apr/1494_Guterman_high_performance_accountable_care_v3.pdf

186. Pollack CE, Armstrong K. Accountable care organizations and health care disparities. JAMA. 2011 Apr 27;305(16):1706–1707.

187. Fisher ES, Shortell SM. ACOs: making sure we learn from experience. New York, NY: The Commonwealth Fund; 2012 [cited 2013 July 12]; Available from:

http://www.commonwealthfund.org/Blog/2012/Apr/ACOs-Making-Sure-We-Learn-from-Experience.aspx.

188. Gosden T, Forland F, Kristiansen IS, Sutton M, Leese B, Giuffrida A, Sergison M, Pedersen L. Impact of payment method on behaviour of primary care physicians: a systematic review. Journal of Health Services Research and Policy. 2001 Jan;6(1):44–55.

189. Grabowski DC, Huckfeldt PJ, Sood N, Escarce JJ, Newhouse JP. Medicare postacute care payment reforms have potential to improve efficiency of care, but may need changes to cut costs. Health Affairs. 2012 Sep;31(9):1941–1950.

190. Mechanic RE. Opportunities and challenges for episode-based payment. New England Journal of Medicine. 2011 Sep 1;365(9):777–779.

191. Sood N, Huckfeldt PJ, Escarce JJ, Grabowski DC, Newhouse JP. Medicare's bundled payment pilot for acute and postacute care: analysis and recommendations on where to begin. Health Affairs. 2011 Sep;30(9):1708–1717.

192. Maddux FW. Impact of the bundled end-stage renal disease payment system on patient care. Blood Purification. 2012;33(1–3):107–111.

193. Wiler JL, Beck D, Asplin BR, Granovsky M, Moorhead J, Pilgrim R, Schuur JD. Episodes of care: is emergency medicine ready? Annals of Emergency Medicine. 2012 May;59(5):351–357.

194. Winkelmayer WC. Potential effects of the new Medicare Prospective Payment System on drug prescription in end-stage renal disease care. Blood Purification. 2011;31(1–3):66–69.

195. Iglehart JK. Bundled payment for ESRD—including ESAs in Medicare's dialysis package. Minnesota Medicine. 2011 May;94(5):38–39.