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

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.