Improving Evaluations of Value-Based Purchasing Programs

Megan McHugh and Maulik Joshi

Although value-based purchasing (VBP) holds promise for encouraging quality improvement and addressing rising costs, currently there is limited evidence about how best to structure and implement VBP programs. In this commentary, we highlight several issues for improving evaluations of VBP programs. Implementation research can be enhanced through early and continuous assessment and greater variation in program designs. Impact research can be improved by creating better outcome measures, increasing the availability of linked patient-level data, and advancing synthesis research. We offer several recommendations for improving the foundation to conduct evaluations of VBP programs to better inform policy and practice.

Key Words. Payment reform, value-based purchasing, bundled payment, implementation research, impact research

---

There is widespread agreement that fee-for-service payment, which remains the most common form of provider reimbursement (Steinwald 2008), encourages overuse of health services and offers little incentive for providers to improve quality, shift patients to lower-cost settings, or coordinate care. Value-based purchasing (VBP), which creates a stronger link between payment and performance, has been proposed as a way to address these shortcomings. For our purpose, VBP is defined as programs for which financial incentives or disincentives are overlaid on top of the existing reimbursement mechanism (commonly called pay-for-performance [P4P]); or fundamental payment reform, such as bundling payments across settings (ambulatory, hospital, or postacute) or providers (physicians and institutions), or global capitation payment for a wide array of services. If structured appropriately, VBP holds the potential to motivate providers to improve

clinical quality, coordinate care, reduce adverse events, encourage patient-centered care, avoid unnecessary costs, and invest in information technologies and other tools proven effective in improving quality (CMS Hospital Pay for Performance Workgroup 2007). Results from the CMS' Premier Hospital Quality Incentive Demonstration, the largest hospital demonstration project that ties payment to performance, lend support to the notion that financial rewards and penalties are associated with improved hospital performance and patient outcomes (Premier Inc., 2008).

To date, VBP has been limited to P4P projects and a small number of projects that fundamentally alter reimbursement methods, so there is little information available to guide the adoption of VBP broadly. As articulated during a 2008 Senate Finance Committee roundtable on VBP (Baucus 2008), policy makers struggle with basic questions, including the following:

- How should VBP policies be structured to best promote quality improvement efforts?
- How should VBP policies be implemented to ensure that goals are achieved?

Nevertheless, policy makers have heeded the calls from advisory groups (e.g., MedPAC, Institute of Medicine) and others to expand VBP under Medicare. The Patient Protection and Affordable Care Act includes provisions to establish a VBP program for hospital payments based on the hospital quality reporting program; a national, voluntary, 5-year bundled payment pilot program; and a new payment structure for providers organized as accountable care organizations. Additionally, the law creates a Center for Medicare and Medicaid Innovation within CMS to test, evaluate, and expand innovative payment and service delivery models that improve quality and reduce program expenditures.

These efforts and others will provide a laboratory for evaluating new ways to pay for health services, and the results will be of interest to policy makers and payers from the private sector. But there are several challenges to conducting evaluations of VBP programs. The purpose of this paper is to discuss the data and methods needed to facilitate evaluations of VBP programs, and to recommend actions to address those needs.

Address correspondence to Megan McHugh, Ph.D., Director, Research, Health Research & Educational Trust, 155 N. Wacker Drive, Suite 400, Chicago, IL 60654; e-mail: mmchugh@aha.org. Maulik S. Joshi, Dr.P.H., President, Health Research & Educational Trust, Senior Vice President of Research, is with the American Hospital Association, Chicago, IL.

A FRAMEWORK FOR EVALUATING VBP PROGRAMS

In an oversimplified diagram, Figure 1 depicts two key areas for research—implementation and impact—needed to advance our understanding of VBP. Implementation research assesses the set of activities that put into practice an activity or program of known dimensions (National Implementation Research Network 2008). As applied to VBP programs, these studies critically assess the building blocks of the programs, including participants (providers and patients), program design elements (e.g., goals, the reimbursement structure, size of incentive payments) and the process of rolling out the program or bringing it to scale (e.g., provider and patient recruitment, technical assistance). Impact research explores changes in outcomes associated with VBP programs. In the context of VBP, impact studies identify changes in provider behavior, patient outcomes, and program spending, and ultimately explain whether the program achieved the intended goals.

IMPLEMENTATION RESEARCH ON VBP

The science of implementation research is evolving (Institute of Medicine 2007), and researchers studying the implementation of VBP programs now have several frameworks to guide their work (Fixsens et al. 2005; Graham and Tetroe 2009). However, implementation research still often receives short shrift from researchers and funders, even though questions about why or how

Figure 1: Framework for Value-Based Purchasing Program Evaluations

a program works are at least as important as whether it works (Pawson and Tilley 1997; Berwick 2008). When a VBP program fails, it is critical to understand the reason for failure (e.g., flawed design, poor implementation) so that future efforts do not repeat the same mistakes. When a program succeeds, it is critical to understand the context in which the program was implemented to inform future replication efforts.

Two improvements to the current approach for conducting implementation research would accelerate our knowledge about when and how VBP programs work. First, implementation studies should be launched early. Studies that begin data collection after the program is launched may miss the opportunity to gather information on the program decisions that were made during the planning stage. As a first step to conducting an evaluation of a VBP program, researchers should interview program designers to understand the program's logic model and rationale for the program structure. Second, implementation studies should be longitudinal. Implementation studies, as the name suggests, typically examine only the time associated with the implementation of a program, often the first 6–12 months, and findings are often based on site visits or interviews that occur at a single point in time. However, VBP programs will evolve over time. Modifications to the structure of the program may occur in response to challenges or successes encountered, program goals or targets may change, and participants (providers and patients) may have different enrollment dates into the program. There may be important structural changes (e.g., technology investments, expanded use of teams) that occur over time as organizations shift away from an FFS culture. Capturing these changes is essential because the presence or absence of the structural changes may contribute to outcomes. Focusing on the first 6 months of a program will not permit evaluators to accurately capture all of these factors.

There are at least a couple of challenges that must be overcome to promote early and continuous data collection on VBP programs. First, there must be an early recognition on the part of funders and program designers about the importance of implementation research. Those well positioned to sponsor implementation research on VBP (e.g., CMS, AHRQ, and private funders) should partner with researchers as early in the process as possible so that implementation studies can incorporate the planning stages. Further, funders and researchers should make sure that the scope of work and resources devoted to data collection reflect the importance of continuous program monitoring.

The second challenge to conducting longitudinal implementation research is the lengthy delay associated with the Office of Management and

Budget (OMB) clearance process. Under the Paperwork Reduction Act passed by Congress in 1980 and amended in 1995, OMB clearance is required for federally funded research studies for which the data collection activities will involve 10 or more respondents and where the questions are standardized in nature. The clearance process affords an opportunity for the public, the OMB, and the sponsoring federal agency to evaluate the utility and appropriateness of the information to be collected and to assess the burden (i.e., time, effort, financial resources) on respondents (AHRQ 2008). According to AHRQ, and based on our own experience, the OMB clearance process takes 7–9 months to complete, during which data collection is on hold. Because program designers are typically unwilling to wait to begin planning and implementing their programs, evaluations supported by federal dollars get a very late start collecting data. The lengthy process compromises the value of the research, and one could envision a much more efficient clearance process, potentially housed outside of OMB, which would permit more timely data collection.

Another hindrance to implementation research is the lack of variation in program designs and participants. Many VBP programs are implemented in a single site or in a uniform way across providers, which limits researchers' ability to draw conclusions about whether and how the program can succeed in other sites. For example, Geisinger Health System's ProvenCare coronary artery bypass surgery program offers a single-episode price that includes preoperative evaluation and workup, hospital and professional fees, routine postdischarge care, and management of related complications within 90 days (Paulus, Davis, and Steele 2008). A case study on ProvenCare suggests that implementation was fostered by a number of factors, including Geisinger's integrated delivery system, electronic health record system, and history of innovation. But the study provides little information on whether bundled payment is appropriate for other systems or hospitals.

The opportunity to learn about the implementation—and impact—of VBP programs relies largely on payers' adoption of practices that vary in design and context and include a diverse group of participants. The VBP programs established under the health reform law will help, but policy makers should continue to facilitate greater experimentation with VBP and promote variation in the ways in which programs are implemented. It is essential that these VBP projects include a variety of participants, and incentives may be needed to encourage participation by providers that have traditionally been overlooked in VBP pilots (e.g., nonphysician-hospital organizations or nonintegrated academic medical centers). For example, program designers

could develop “reinsurance” for VBP programs, thus providing a layer of coverage if the VBP program has unintended financial consequences. One challenge to the recruitment of diverse providers is the requirement for budget neutrality for Medicare demonstration projects. The requirement may limit participation to a small percentage of provider organizations with a more advanced infrastructure (e.g., IT, EHRs) or integrated provider network. The cost neutrality requirement should be revisited so that participants in Medicare demonstration program are more representative of the general provider population.

To enhance generalizability, program designers should also be encouraged to experiment with different implementation strategies, for example, provider recruitment, program materials and education, technical assistance, or reporting requirements. The variability of program structures, participants, and implementation elements will produce data necessary for researchers to identify when, how, and why VBP programs may be successfully implemented and maintained. However, this call for increased VBP experimentation requires careful planning; poorly conceived projects implemented within organizations clearly incapable of succeeding under the program would do little to improve knowledge that would enhance generalizability.

IMPACT RESEARCH ON VBP

Investigating the impact of VBP policies requires information on the interventions (i.e., the VBP program), the outcomes (i.e., performance measures and health outcomes), and the units (i.e., patients, providers). Data on the interventions are relatively easy to obtain and readily available, assuming an implementation analysis was conducted. However, there are a number of shortcomings concerning the availability and appropriateness of data on outcomes and units that are needed for inclusion in multivariate models to determine the impact of VBP policies.

Meaningful Outcome Measures

The implied goal of VBP is improved quality in return for the same payment, or lower payment for the same level of quality (Tompkins, Higgins, and Ritter 2009). Payment is comparatively easy to measure; payers can calculate spending per encounter or per beneficiary over a given time period.

Although progress in the development, adoption, and reporting of quality indicators within the past 5 years should be celebrated, a number of shortcomings remain. Most outcome measures in use today focus on the delivery of preventive services (e.g., smoking cessation, vaccines), process of care for certain clinical conditions (e.g., heart failure, pneumonia, diabetes), and complications (e.g., surgical inflections), rather than patient outcomes or cost. The measures do not address all six of the Institute of Medicine's quality aims (safe, timely, efficient, effective, equitable, and patient-centered), nor are they appropriate for measuring episodes of care or care across a continuum of different providers. Further, the measures are not appropriate for all patients. For example, Hospital Quality Alliance measures, which are currently tied to hospital reimbursement, focus on care processes associated with subsets of clinical practice that are more appropriate for older patients, rather than children. As such, the measures are not well suited for Medicaid.

If payment reform moves towards reimbursement for bundles of services, rather than a single visit to a provider, new measures are needed to encourage care coordination and integration across settings, and convey a shared accountability for patients. Bundled measures, which could be individual measures or a composite of several measures, will be challenging to develop, as they must reflect the quality of care delivered by all providers for a given episode of care. There are a number of challenging data issues that need to be addressed for the construction of meaningful measures. For example:

- **Sample sizes:** The sample sizes determined to be adequate by statisticians may be impractical or impossible to collect in practice, particularly for smaller providers.
- **Risk adjustment:** Although risk adjustment exists in specific settings, such as the intensive care unit or the hospital, there are no risk adjustment methods to account for severity of illness across providers or at the community level.
- **Data validation:** Current approaches for validating and auditing data for accuracy can be expensive and time consuming, and little is known about the efficacy of data validation approaches. However, the increasing use of electronic health records with electronic forcing functions and data "checks" may help to ensure that the collected data are valid.
- **Measurement composition:** Composite measures are becoming the norm. However, one of the challenges is assigning relative weights to the individual measures.

- Indication of relative performance: Ultimately the measures should provide information on the relative performance of providers. However, much of the variability in current, individual outcome measures cannot be easily explained in statistical models, and it raises questions about the appropriateness of measures to assess and differentiate providers (Tompkins, Higgins, and Ritter 2009).

Once measures are developed, they must be periodically reevaluated and updated. Over time, if VBP programs become more comprehensive (i.e., a majority of health services are reimbursed through mechanisms that are fundamentally different from FFS), measures could include population-based indicators of health. We believe that the CMS and AHRQ are the best positioned to fund and coordinate the development of new measures, and the National Quality Forum should be tasked with assessing and endorsing the measures. Policies to expand the capacity of these organizations may be necessary.

Availability of Patient-Level Data across Care Settings

Evaluations of VBP strategies' impact will require analysis of patient-level diagnosis, treatment, and discharge status (if appropriate), which are typically available in claims data and medical records. Ideally, the analyses will also include patient-specific characteristics such as age, race, and ethnicity to better understand how different groups may fare under the change in payment policy. But as VBP strategies encourage greater collaboration among providers through bundled or global payments, and as outcome measures are developed to better capture the whole patient experience for a single episode, data must be available to analyze patient encounters longitudinally, across different providers in different settings. Although the focus is typically on linking data from physicians, hospitals, and postacute care facilities, other providers should also be included as they represent important services on the continuum of care, for example, prehospital emergency medical services, home health, pharmacy, and school health. The availability of these data will be aided by the penetration and use of health information technology systems and electronic medical records, assuming the compatibility of systems across providers. However, the linking of data across providers requires unique patient identifiers, which introduces difficult questions about protection of patient confidentiality. Needless to say, stringent data use agreements must be developed for evaluators who are permitted to access these data.

Table 1: Summary of Recommendations

<i>Problem</i>	<i>Recommendation</i>	<i>Focus</i>	<i>Target Audience</i>
Limited information on implementation and management of VBP programs	Early and continuous collection of data on implementation	Methods/ infrastructure support	Researchers/policy makers/providers
Limited generalizability of findings	More experimentation and greater variation in VBP	Data/methods	Researchers/policy makers/providers
Lack of meaningful outcome measures	Improved methods for risk adjustment, data validation, and measurement composition	Data	Policy makers/ providers/ researchers
Lack of integrated and aggregated data	Support for EHR/IT systems	Infrastructure support	Policy makers/ providers
Limited ability to synthesize learning from diverse VBP efforts	Better practices/methods for synthesizing VBP program findings	Methods	Researchers/policy makers

Synthesis Research

It is unlikely that any single evaluation will provide enough information to inform ongoing policy. The proliferation of P4P programs in the public and private sectors has permitted researchers to conduct synthesis research, leading to at least preliminary lessons for future P4P policy (Rosenthal and Dudley 2007; McNamera 2009). But as VBP programs become more complex and varied through bundled or global payment, it will become more challenging to conduct synthesis research. For example, outcome measures may not be common across studies and program structures may have little in common. There is an opportunity for researchers—methodologists—to consider ways in which we may distill findings from smaller VBP evaluations into guidance better suited for policy making. For example, a synthesis of individual VBP evaluations may reveal certain contextual factors or program elements that are common among successful VBP efforts. We propose that AHRQ convene a roundtable of experts—both in quantitative and qualitative methods—to consider ways in which researchers could conduct synthesis research or even meta-analyses of VBP evaluations. As a field, health services research could benefit from additional effective methods that seek to leverage knowledge from prior research that may be in the same topic area but

disparate in design, measures, analysis, and learnings. Because many VBP programs are operated by private payers (79 percent of P4P programs were administered in the private sector in 2007; Baker and Delbanco 2007), synthesis research should involve public-private collaboration.

CONCLUSION

To effectively learn from VBP program and facilitate the design of future policies, a number of barriers to implementation and impact research must be overcome. Our eagerness to understand the impacts of VBP must not overshadow the need to learn more about best ways to implement, maintain, and manage VBP programs. But implementation research must advance to overcome its methodological shortcomings of narrow focus and limited generalizability. With regard to impact research, the major challenge is the collection and integration of more and better data. And findings from single program evaluations need to be organized and synthesized in a way that identifies best practices for future policy. Table 1 summarizes our recommendations for advancing data and methods to evaluate VBP.

As important as VBP is to health reform, so is the evaluation of VBP. Generating a better understanding of VBP programs is critical to the effective implementation of successful programs. However, the foundation for research needs to be improved for health services researchers to answer the most pressing policy questions.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This commentary was supported by a grant from AcademyHealth.

Disclosures: None.

Disclaimers: None.

REFERENCES

- AHRQ. 2008. *Navigating the OMB Clearance Process: Guidance for ACTION Task Orders*. Rockville, MD: AHRQ.
- Baker, G., and S. Delbanco. 2007. *2006 Longitudinal Survey Results with 2007 Market Updates*. San Francisco: Med Vantage.

- Baucus, M. 2008. *Opening Statement*. Paper presented at the Senate Finance Committee Roundtable on Value Based Purchasing, Washington, DC.
- Berwick, D. M. 2008. "The Science of Improvement." *Journal of the American Medical Association* 299 (10): 1182–4.
- CMS Hospital Pay for Performance Workgroup. 2007. *Medicare Hospital Value-Based Purchasing Plan Development*. Baltimore, MD: CMS.
- Fixsens, D. L., S. F. Naom, K. A. Blase, R. M. Friedman, and F. Wallace. 2005. *Implementation Research: A Synthesis of the Literature (No. 231)*. Tampa, FL: University of South Florida.
- Graham, I., and J. Tetroe. 2009. "Learning from the U.S. Department of Veterans Affairs Quality Enhancement Research Initiative: QUERI Series." *Implementation Science* 4 (1): 13.
- Institute of Medicine. 2007. *The State of Quality Improvement and Implementation Research: Expert Views*. Washington, DC: National Academies Press.
- McNamara, P. 2009. *Pay for Performance: The US Experience*. Rockville, MD: AHRQ.
- National Implementation Research Network. 2008. "What Do We Mean by Implementation?" [accessed on January 14, 2010]. Available at http://www.fpg.unc.edu/~nirn/implementation/01_implementationdefined.cfm
- Office of Management and Budget. 2009. *Budget of the U.S. Government, Fiscal Year 2010: Summary Tables*. Washington, DC: OMB.
- Paulus, R. A., K. Davis, and G. D. Steele. 2008. "Continuous Innovation in Health Care: Implications of the Geisinger Experience." *Health Affairs* 27 (5): 1235–45.
- Pawson, R., and N. Tilley. 1997. *Realistic Evaluation*. London: Sage Publications.
- Premier Inc. 2008. *Patient Lives Saved as Performance Continues to Improve in CMS, Premier Healthcare Alliance Pay-for-Performance Project*. Charlotte, NC: Premier Inc.
- Rosenthal, M. B., and R. A. Dudley. 2007. "Pay-for-Performance: Will the Latest Payment Trend Improve Care?" *Journal of the American Medical Association* 297 (7): 740–4.
- Steinwald, A. B. 2008. *Primary Care Professionals: Recent Supply Trends, Projections, and Valuation of Services*. Washington, DC: General Accountability Office.
- Tompkins, C., A. R. Higgins, and G. A. Ritter. 2009. "Measuring Outcomes and Efficiency in Medicare Value-Based Purchasing." *Health Affairs* 28 (22): w251–61.

SUPPORTING INFORMATION

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

Appendix SA1: Author Matrix.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.