Strategies To Support Quality-based Purchasing: A Review of the Evidence

Agency for Healthcare Research and Quality

www.ahrq.gov

The mission of AHRQ is to improve the quality, safety, efficiency, and effectiveness of health care by:

- Using evidence to improve health care.
- Improving health care outcomes through research.
- Transforming research into practice.

Introduction

Deficiencies in quality have been widely documented in the U.S. health care system. A recent component of purchaser response to these data has been the pursuit of quality-based purchasing (QBP). However, purchasers have been uncertain both how to measure quality and what incentives to offer to stimulate performance improvement. Furthermore, there has been dispute in the literature about the validity of quality measures, especially outcomes indicators, and the potential for chance variation in outcomes to unduly influence reported performance. Therefore, despite the release of public reports of providers' outcomes by several States, purchasers have been slow to use outcomes reports to drive QBP policies. Without more information about how to proceed with QBP, purchasers risk investing time, resources, and good will without a reasonable expectation of achieving a good return.

In this report,[1] we sought to describe and evaluate the evidence regarding the effectiveness and potential of QBP strategies to improve the quality of care provided in the U.S. health care system. For this report, QBP is defined as payment or reputational strategies aimed at providers that individual employers, employer coalitions, or government programs could plausibly adopt to stimulate the improvement of quality in health care. With respect to providers, the primary issue within the purchaser's purview is the establishment of incentives—for individual providers or for provider organizations such as medical groups and hospitals—that either stimulate or inhibit provider behaviors to improve quality (strategies aimed at consumers, such as variable copayments, were not considered). Specifically, this report focuses on the two types of incentives in widespread use—*performance-based payment* and *reputational incentives* arising from the public release of performance data.

U.S. Department of Health and Human Services
Public Health Service

Advancing Excellence in Health Care

Objectives

Because quality-based purchasing is in its infancy, the first objective was to develop a conceptual model of how QBP strategies could be used to create incentives for providers to improve care. The second objective was to identify all the published, peer-reviewed randomized controlled trials of QBP and to summarize what is known about the relative effectiveness of different strategies.

Because the literature on QBP is sparse, a third objective was to identify ongoing research that might increase our knowledge. Finally, since one of the main issues purchasers face is whether to use reports of outcomes of care, the fourth objective was to determine whether outcomes reports convey meaningful information or are too influenced by chance events to be useful.

Conceptual Model

There is extensive theoretical literature about the determinants of the effectiveness of incentive arrangements in several disciplines, including economics, psychology, and organizational behavior. An expansive review of that literature is beyond the scope of this report. However, this research has pointed out, among other things, the influence of the characteristics of the incentive itself and of the context in which it is applied on the likelihood that the incentive will be effective.

- **Characteristics of the incentive.** Important financial characteristics include whether it is directed to the optimal recipient. Recipients could include, for instance, the individual provider, provider groups, or even community organizations, with "optimal recipient" varying depending on the goal and degree of coordination among providers required. Other important financial

factors are the potential impact on revenue (based on the magnitude of the incentive and the proportion of encounters or patients to which it applies) and the cost of complying with the performance measure. Nonfinancial characteristics are more numerous and subtle. These include perceived attainability of the performance goals set, the acceptability of those goals (their congruence with professionalism, altruism, and intrinsic motivation and with provider preferences for domain of performance measured), and the approach to reinforcement (e.g., positive vs. negative reinforcement).

- **Contextual factors.** Although these factors are likely very important, they have received little attention, especially in the empirical literature. In particular, we posit that there are predisposing factors—such as the mix of other incentives in the market and individual provider characteristics or a provider organization's understanding of its mission—that that will determine the likelihood of a provider having any interest in responding to a newly introduced QBP program. Furthermore, we also hypothesize that there are enabling factors— especially at the organization level, where many aspects of the structure of care are determined, and at the patient level—that will facilitate or inhibit any efforts a provider makes to improve care.

In emphasizing both the characteristics of the incentive itself (the QBP stimulus to improve) and the predisposing and enabling factors that may vary among providers and markets, we believe this model complements and can integrate most of the existing theories of incentives. It is offered simply to ensure that adequate consideration is given to all key factors in designing both studies of quality-

based purchasing and future QBP programs.

Methods for the Literature Search and Identification of Ongoing Research

Literature Searches

To be considered an article that provided evidence regarding QBP, the intervention in the trial had to be a performance-based payment or reputational incentive strategy that could plausibly be introduced by a purchaser. The focus was on articles that provided definitive primary data from randomized controlled trials, because most non-randomized designs in this domain are severely confounded, especially by selection bias in which providers were willing to accept new incentives, regression to the mean (since organizations may have chosen to introduce incentives targeted at problem areas that would have improved anyway), the Hawthorne effect, and other sources of variation in performance over time not related to the incentive. Articles that did not have clear inclusion and exclusion criteria and greater than 75% followup were excluded.

Standard search strategies were used. These strategies involved the querying of two online databases (MEDLINE® and Cochrane) using key words, followed by evaluation of the bibliographies of relevant articles, Web sites of relevant organizations (especially of funding agencies providing project summaries and of employer organizations pursuing QBP), and reference lists provided by the Technical Expert Panel. At least two investigators screened titles, abstracts, and articles, as necessary, to determine if they met inclusion criteria. From each included article, the following data were extracted, when available: information describing financial and nonfinancial characteristics of the incentive, financial characteristics of the environment including dominant proportion of income from fee-for-service or capitation and other incentives faced, provider characteristics, organizational capabilities, and patient factors, as well as references in the bibliography that might meet inclusion criteria.

Identifying Ongoing Research

The online databases HSRProj and GOLD—the Grants-On-Line Database of the Agency for Healthcare Research and Quality (AHRQ)—were searched, as well as the Web sites of other funders or coordinators of projects (e.g., the Leapfrog Group). Finally, staff at AHRQ, the Robert Wood Johnson Foundation (RWJF), the California HealthCare Foundation (CHCF), and the Commonwealth Fund were asked whether ongoing research that met the inclusion criteria was being funded by those organizations. Two investigators reviewed the abstracts of projects identified from the database searches to assess relevance to the Technical Review. Discrepancies in inclusion were resolved by discussion and re-review and by discussion with project officers at funding agencies or with the principal investigator of the project under consideration.

Results From the Literature Search and Identification of Ongoing Research

Articles Included in the Literature Search

The literature searches identified 5,045 unique candidate articles for inclusion, of which 4,882 were eliminated after review of their abstracts. The remaining 163 articles underwent full text review. Among these there were only nine randomized controlled trials, eight using performance-based payment as the intervention and one using reputational incentives.[1-10]

Completeness of the Literature

In every article reporting the results of a randomized controlled trial of performance-based payment incentives, there were significant variables from our conceptual model that were either not reported at all or that were incompletely described. The only variables that were reported in all trials were characteristics of the incentive itself: the recipient of the incentive, its magnitude, and the domain of performance measured. Several potentially critical variables were never reported in any trial, including payment incentive as a proportion of total income, the costs of complying with the incentive, and most enabling factors at the organizational level.

Findings From Trials of Performance-based Payment

The eight trials of performance-based payment were neither consistent in their design of the independent variable (the financial incentive offered) nor comparable in terms of their dependent variable (the performance indicator measured). Thus, their results are presented as a function of several of the variables within the conceptual model (those that are actually reported for all papers). In total, ten hypotheses and ten dependent variables were tested because one study had two intervention arms (a fee-for-service arm and a bonus arm) compared to controls, and one had two dependent variables (screening for smoking and smoking cessation).

Recipient of incentive. In four studies, the recipient of the incentive was an individual provider, while in the other four the recipient was the provider group or could be either an individual provider or a group. Among the studies targeting individual providers, there were five positive and two negative results; among the studies in which the target was or could be the provider group, there were one positive and two negative results. (In general,

the term "positive" is used to mean an effect in the desired direction—the incentive worked—and "negative" to mean there was no significant effect of the incentive on the outcome measure.)

In seven studies, with a total of nine dependent variables, the target of the incentive was a physician. Of the nine dependent variables assessed, five showed a significant relationship to the incentive in the expected direction and four showed no significant change after the incentive was introduced. A single study involved pharmacists and was positive.

Magnitude of the incentive. Incentives ranged in magnitude from $0.80/flu shot to a bonus of up to $10,000 per clinic per year. There was no consistent relationship between the magnitude of the incentive and response (though the lack of similar interventions and dependent variables make it unlikely that any pattern could be detected, even qualitatively).

Fee-for-service vs. bonus. There were five dependent variables in fee-for-service studies (that is, the intervention involved paying providers a higher than usual fee for each encounter if and only if a performance standard was met) and five in bonus studies. Among the fee-for-service studies, four were positive and one was negative. Among the bonus studies, two were positive and three were negative.

Performance domain measured. Among the articles included, there were seven studies of preventive care with nine dependent variables assessed. Among these nine outcomes, five were positive and four were negative. The single study addressing chronic care was positive.

Patient factors. Authors did not report the burden adherence would place on patients in any of the articles. However, in a general sense, incentives to achieve performance were found to be more effective when the indicator to be followed required less patient cooperation (e.g., receiving vaccinations or answering questions about smoking) than when significant patient cooperation was needed (e.g., to quit smoking).

Findings From Trials of Reputational Incentives

There was only one randomized controlled trial of reputational incentives. This study showed that hospitals with low performance scores were more likely to engage in quality improvement activities. This was especially true for hospitals whose performance was released to the public (as opposed to being kept confidential).

Ongoing Research Identified

We identified no currently ongoing randomized controlled trials of QBP strategies from any funding source. There were 18 ongoing research projects about QBP. For many of these, the exact nature of the performance measures and the incentive were still being determined. For some, the study design is observational; that is, health plans are making decisions about incentives without input from the investigators, but the investigators are assessing the response.

Expected Knowledge To Be Gained From Ongoing Research

Ongoing research being conducted by AHRQ, the Robert Wood Johnson Foundation, the California HealthCare Foundation, and the Commonwealth Fund will provide some important additional information about quality-based purchasing. For example, several studies will describe the type and frequency of use of QBP strategies; others will investigate provider reactions to incentives in terms of willingness to participate in programs and awareness of the incentives offered. In addition, some investigators will obtain quantitative and qualitative information about attitudes towards incentives used and performance targets set (such as salience, clinical validity, and whether the performance measures were within the providers' scope of control). These studies may be useful for understanding providers' motivation to respond and organizational decisionmaking when incentives are offered. Still other projects will report on the tools used to communicate incentives, rather than the provider or consumer response to the incentive.

The Rewarding Results projects (with components sponsored by RWJF, CHCF, and AHRQ) as well as several others will provide assessments of the impact of incentives on traditional performance measures of structure, process, and outcomes. Although none of these is randomized and all involve organizations that self-select to adopt or participate in incentive programs, taken together they will provide preliminary evaluations of QBP in Medicaid, Medicare, and commercial insurance settings and will cover many different approaches to incentives.

Among the interventional studies, there are also some major differences in the characteristics of the incentives themselves between the prior literature and the ongoing research. For instance, the ongoing studies involve actual health plans or government programs making an ongoing commitment to an incentive strategy, rather than a researcher making a short-term payment intervention (which was the situation in the prior studies). Similarly, all the studies included in the literature review above involved incentives directed at only a small number (usually just one) performance indicator for a single condition or type of patient. However, all the ongoing interventional studies identified involve multiple measures (often ten or more) across a variety of conditions and distinct patient populations. Both these factors—that the incentive comes from a payer (e.g., health plan, government)

and that there are multiple quality indicators—will provide more broadly applicable evidence about the probability that provider investments in quality improvement (e.g., installing a new information system) can be recouped relative to previously studied incentive strategies.

Methods for Simulations To Assess the Usefulness of Outcomes Reports

To examine the role of random variation versus true hospital quality differences in assessing reported hospital outcomes, simulations were developed to determine how often hospitals would be mislabeled in public reports. To do this, first assumptions were made about what the population of hospitals looks like in terms of both the proportion of hospitals with good and poor quality and the difference in outcomes between these groups of hospitals. The second step was to calculate, given the first assumptions, the probability that an individual hospital with known characteristics will receive a particular label (e.g., "poor" vs. "good" vs. "superior") and how often those labels will be misapplied (e.g., that a poor quality hospital will be labeled "good"). This mislabeling is possible because random variation in patient outcomes can occur such that, by chance, a good hospital could potentially have a significantly worse than expected mortality rate. (This is discussed in terms of mortality rates, but the same logic applies to any other outcome.) How often this happens is a function of the difference in performance rates between good and bad hospitals and the sample size at each hospital (which determines the standard deviation of measured performance for like hospitals).

Assumptions for the Simulations

Prior studies have suggested that the influence of chance is very great,

perhaps enough to cause outcomes reporting to do more harm than good. However, these were based on assumptions—usually based on implicit reviews of overall performance rather than explicitly assessing compliance rates for specific aspects of care—that included a relatively simple performance distribution (e.g., only "good" and "bad" hospitals) with small differences in performance between the groups.[11] For completeness sake, some simulations were performed using assumptions taken from prior research. However, some simulations were done in which assumptions about hospital performance were based on published California data about acute myocardial infarction mortality rates from 1991-1998. These data showed approximately 10% of hospitals had been labeled "better than expected," 80% had been labeled "no different than expected," and 10% had been labeled "worse than expected" in most years. Furthermore, hospitals labeled "better than expected" had been shown in validation studies to have superior processes of care compared to hospitals labeled "worse than expected." Thus, although a simplification (hospital performance is likely aligned along a spectrum, rather than divided into only three groups), these results support the assumption of a distribution of hospital performance that included 10% poor quality, 10% superior quality, and 80% good (or expected) quality hospitals.

Estimates were obtained of probability of death at poor, good, and superior quality hospitals using 3-year grouped data from the published California study of acute myocardial infarction outcomes. Hospitals that were found consistently—i.e., over two or three of the 3-year periods included in the data (1991-1993, 1994-1996, and 1996-1998)—to have statistically significantly higher than expected mortality were included in the group of poor hospitals; those with consistently lower than expected mortality were included in the

Advancing Excellence in Health Care

group of superior hospitals, and all others were in the good or expected group.

Assessments of Outcomes Reports and Labels

Using these assumptions, simulations were run to determine the proportion of hospitals from each group (i.e., hospitals that were truly poor, good, or superior) that would be designated into each group (i.e., the proportion that would receive the labels "poor," "good," or "superior"). Since hospitals that have generally been performing well which have a single event in which they are labeled "poor" might face few consequences, simulations were performed not just for a single point in time, but also for two or three measurement periods. The impact of varying sample sizes at a hospital was also considered.

Results From Simulations To Assess the Usefulness of Outcomes Reports

Simulations Using Assumptions From the Literature

As expected, when the assumptions used previously are made again, the results suggest that random variation causes frequent mislabeling of hospitals in a single period, with potentially more than half the hospitals labeled "poor" actually coming from the population of good hospitals. However, when the analysis is extended over as few as 3 years, mislabeling more than once becomes extremely unusual for good hospitals; fewer than 0.2% of good hospitals would have this outcome even if one assumes small mortality differences between poor and good hospitals.

Simulations Using Assumptions From California Data

The mortality rates for acute myocardial infarction for poor, good, and superior hospitals in California in 1996-1998 were 17.1%, 12.2%, and 8.6%, respectively. Using these mortality rates, superior hospitals were almost never labeled "poor" and vice versa. Over a 3-year period (with reports each year), 92.5% of poor hospitals would be labeled as such at least once (vs. only 8.7% of good hospitals) and almost all the hospitals that were labeled poor more than once would in fact be poor. Similarly, most superior hospitals would receive at least one such label, and almost all hospitals labeled superior more than once would actually be superior.

Discussion and Future Research

Quality-based purchasing is a relatively new topic, and very few studies were found that address the key questions about QBP. Comparison of our conceptual model to the available research also points out that the studies available are incomplete in their reporting of potentially key mediators of the effects of incentives.

Nonetheless, there is evidence that, in some circumstances, both performance-based payment and reputational incentives can work. Preliminary evidence suggests that, consistent with theory, the revenue potential from incentives and the costs of achieving performance goals may influence response, as will enabling or inhibiting factors at the patient level. In addition, ongoing research will inform us about the extent of use of QBP, provider attitudes toward both incentives and the use of various types of performance measures, and preliminary estimates (though the data will come from non-randomized studies) of the impact of QBP on quality.

Much additional research is needed, including both qualitative and quantitative designs. Since randomized trials are expensive and providers often will not agree to randomization, funders might consider looking for natural experiments or situations in which non-randomly selected control groups could reasonably be used (as when a health plan decides to roll out a QBP approach first in one city, then in another; of course, even in these situations there will probably be a reason as to why one city was chosen to be first that could bias results). One such example may be the recently initiated Centers for Medicare & Medicaid Services Premier Hospital Quality Incentive Demonstration that recognizes and provides financial rewards to hospitals that demonstrate high quality performance in a number of areas of acute care (see: www.cms.hhs.gov/researchers/demos/phqidemo.asp).

Furthermore, subsequent research should explicitly address the elements from conceptual models that have largely been ignored. Investigators should address the reality that while much of performance is ultimately determined by the actions of individual providers, enabling factors at the organizational and community levels that determine the structure and processes of care are also important and could be targets for incentive strategies. In addition, studies that address the combination of performance-based payment with reputational incentives are needed.

Finally, one must recognize that a prominent barrier to QBP is that the science of performance measurement is still underdeveloped. Purchasers interested in QBP have limited choices for performance measures and these disproportionately target preventive care and structure or processes rather than outcomes. That is, the available set of metrics is not broadly representative of all care, while purchasers must pay for care across the entire clinical spectrum. This suggests that research into QBP should be accompanied by further development of the basic tools of performance measurement.

Conclusion

The environment in which purchasers and providers interact is rapidly changing. There is clearly growing interest in QBP and some evidence that both payment and reputational incentives can work; but, to date, there is little unequivocal data on which to base QBP strategy selection. Our modeling suggests that, with appropriate caution, outcomes measures can be included among the performance indicators used for QBP. Furthermore, the notion of using incentives to encourage high quality (as well as actually measuring quality) is much more acceptable than it was a few years ago, and this has increased the number of opportunities to study QBP. Researchers have responded with a broad portfolio of ongoing research that promises to both outline current trends in the use of QBP and offer some preliminary evaluations of several different incentive approaches. Additional policy-relevant research, including studies incorporating in their designs conceptual considerations such as those outlined here, may rapidly advance our understanding of how to use performance measurement and incentives to improve the quality of health care Americans receive.

For More Information

Printed copies of the Technical Review from which this summary was taken may be obtained free of charge from the AHRQ Publications Clearinghouse by calling 800-358-9295. Requesters should ask for Technical Review 10, *Strategies To Support Quality-based Purchasing: A Review of the Evidence* (AHRQ Pub. No. 04-0057).

Additionally, the Technical Review and this summary will be available online through AHRQ's Web site at www.ahrq.gov.

References

1. Christensen DB, Holmes G, Fassett WE, et al. Influence of a financial incentive on cognitive services: CARE project design/implementation. *J Am Pharm Assoc.* Sep-Oct 1999;39(5):629-639.

2. Christensen DB, Hansen RW. Characteristics of pharmacies and pharmacists associated with the provision of cognitive services in the community setting. *J Am Pharm Assoc.* Sep-Oct 1999;39(5):640-649.

3. Davidson SM, Manheim LM, Werner SM, Hohlen MM, Yudkowsky BK, Fleming GV. Prepayment with office-based physicians in publicly funded programs: results from the Children's Medicaid Program. *Pediatrics.* Apr 1992;89(4 Pt 2):761-767.

4. Fairbrother G, Siegel MJ, Friedman S, Kory PD, Butts GC. Impact of financial incentives on documented immunization rates in the inner city: results of a randomized controlled trial. *Ambul Pediatr.* 2001;1(4):206-212.

5. Hickson GB, Altemeier WA, Perrin JM. Physician reimbursement by salary or fee-for-service: effect on physician practice behavior in a randomized prospective study. *Pediatrics.* Sep 1987;80(3):344-350.

6. Hillman AL, Ripley K, Goldfarb N, Nuamah I, Weiner J, Lusk E. Physician financial incentives and feedback: failure to increase cancer screening in Medicaid managed care. *Am J Public Health.* Nov 1998;88(11):1699-1701.

7. Hillman AL, Ripley K, Goldfarb N, Weiner J, Nuamah I, Lusk E. The use of physician financial incentives and feedback to improve pediatric preventive care in Medicaid managed care. *Pediatrics.* Oct 1999;104(4 Pt 1):931-935.

8. Kouides RW, Bennett NM, Lewis B, Cappuccio JD, Barker WH, LaForce FM. Performance-based physician reimbursement and influenza immunization rates in the elderly. The Primary-Care Physicians of Monroe County. *Am J Prev Med.* Feb 1998;14(2):89-95.

9. Roski J, Jeddeloh R, An L, et al. The impact of financial incentives and a patient registry on preventive care quality: increasing provider adherence to evidence-based smoking cessation practice guidelines. *Prev Med.* Mar 2003;36(3):291-299.

10. Hibbard JH, Stockard J, Tusler M. Does publicizing hospital performance stimulate quality improvement efforts? *Health Aff* (Millwood). Mar-Apr 2003;22(2):84-94.

11. Thomas JW, Hofer TP. Accuracy of risk-adjusted mortality rate as a measure of hospital quality of care. *Medical Care.* January 1999;37(1):83-92.

Advancing Excellence in Health Care