Final Contract Report

Pay for Performance: A Decision Guide for Purchasers

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No./Assignment No. 290-01-0001/298

Prepared by:
R. Adams Dudley, M.D., M.B.A.
University of California, San Francisco

Meredith B. Rosenthal, Ph.D.
Harvard School of Public Health

AHRQ Publication No. 06-0047
April 2006

Acknowledgment

This Guide is intended to be used by public and private purchasers of health care services, including health plans, who are considering sponsorship of a pay-for-performance (P4P) initiative. Twenty questions, identified in collaboration with purchasers, are presented for consideration along with options and any available evidence for each. The authors gratefully acknowledge the valuable comments made by 10 public and private purchasers, who generously contributed their time and expertise in reviewing a formative draft.

Dudley RA, Rosenthal MB. Pay for Performance: A Decision Guide for Purchasers. Rockville, MD: Agency for Healthcare Research and Quality; 2006. AHRQ Pub. No. 06-0047.

Contents

The P4P Checklist: 20 Questions For Purchasers To Consider
Phase 1. Contemplation
√ Question 1. Is our community ready?
√ Question 2. Should we partner with other purchasers or go it alone?
√ Question 3. When and how should we engage providers in P4P discussions?
Phase 2. Design
√ Question 4. Which providers should we target first? Hospitals or physicians? Specialists or primary care providers?
√ Question 5. For physicians, what are the advantages and disadvantages of targeting individual clinicians versus medical groups? In the case of hospitals, what are the advantages and disadvantages of targeting individual hospitals versus hospital systems?
√ Question 6. Should provider participation be voluntary or mandatory?
√ Question 7. Should we use carrots or sticks—bonuses or penalties—or a combination?
√ Question 8. How should the bonus be structured?
√ Question 9. Should we use relative or absolute performance thresholds?
√ Question 10. What are our options for phasing in pay for performance?
√ Question 11. Where do we find the money?
√ Question 12. How much money should we put into performance pay?
√ Question 13. What measure characteristics make them attractive candidates for inclusion in an initial measure set?
Phase 3. Implementation
√ Question 14. How do we address providers' concerns about whether risk adjustment adequately captures the severity of illness of their patients?
√ Question 15. If we currently sponsor a private or public report card, will P4P offer more of an incentive? If we are considering both a public report and P4P, which should we pursue first?
√ Question 16. Should we tailor pay for performance for subsets of a particular group of providers, e.g., safety-net hospitals?
√ Question 17. How should we think about pay for performance and its relationship to benefit design, including tiered networks?
√ Question 18. Is there any special advice for Medicaid agencies and Medicaid managed care plans interested in pay for performance?
Phase 4. Evaluation
√ Question 19. How can we tell if the P4P program is working?
√ Question 20. What unintended consequences should we look for?

Introduction

Recent surveys suggest that the number of pay-for-performance (P4P) initiatives nationwide sponsored by a variety of health plans, employer coalitions, and public insurance programs now exceeds 100.[1, 2, 3] Through these various programs, most physicians and hospitals in the United States currently face or are in discussions with local purchasers about some form of pay for performance.[4] The sponsors of these incentive programs state that either rewarding or improving quality of care is a primary goal; the other goal is usually controlling costs either directly or indirectly by reducing errors and inappropriate utilization.

We define "pay for performance" broadly and include any type of performance-based provider payment arrangements including those that target performance on cost measures. Despite the growing use of P4P initiatives, there is little evidence on how best to design incentive programs in the health sector.[5] Perhaps as a result of the paucity of evidence, there is tremendous variety in the approaches used in existing incentive programs.[, 6]

Existing P4P initiatives are sponsored by government purchasers—Medicare and Medicaid—as well as private employers, coalitions of employers, and health plans. We use the term "purchasers" to refer to all these potential sponsors. Although this Guide is developed for a purchaser audience, we note that some P4P programs have been initiated by providers.

There are many decisions that go into the design of a P4P program, and each decision affects the likelihood that a program will achieve its goals. In this Guide, we isolate and sequence 20 questions purchasers face in considering pay for performance, review options and any available evidence—from empirical evaluations and economic theory—that may inform future decisionmaking, and discuss potential effects and unintended consequences. We group questions into one of four phases through which a purchaser considering P4P might evolve: contemplation, design, implementation, and evaluation.

We recognize that pay for performance is only one among many possible and valuable strategies that purchasers may undertake to improve the quality and affordability of health care. Purchasers contemplating P4P need to consider the appropriate role and limitations of payment incentives in comparison to other potential strategies including physician and patient education, private and public report cards, disease management, and technical assistance.

Finally, we must note that, while P4P programs create explicit incentives to reward or improve performance, the pre-existing, underlying payment system exerts its own set of (mostly implicit) incentives. For example, fee-for-service payment creates an incentive to increase utilization while capitation payment involves incentives to reduce services. Purchasers must account for the pre-existing payment system incentives when contemplating additional ones. Also, the value of a P4P program will be a function of both gains in the quality of care and the total costs of the program, including additional payments to providers (if any) and the costs of implementation and monitoring.

Phase 1. Contemplation

Purchasers contemplating the adoption of pay for performance should initially consider the three key questions discussed in this section:

1. Is our community ready?
2. Should we partner with other purchasers or go it alone?
3. When and how should we engage providers in P4P discussions?

Question 1. Is our community ready?

P4P initiatives are currently underway in a number of different types of communities, involving large and small purchasers, for-profit and not-for-profit providers and all types of market structures. The Centers for Medicare & Medicaid Services (CMS) demonstration project with Premier, Inc. hospitals, for example, includes urban, rural, and even critical access hospitals[i] (see: www.cms.hhs.gov/HospitalQualityInits/35_HospitalPremier.asp).

Although there have been no studies of the type of community in which P4P is most likely to succeed, two factors are likely to be important: sponsor influence and the pre-existing capacity of local providers to engage in quality measurement and improvement.

Pay for performance is most likely to be effective when it is introduced by a powerful stakeholder in the market. Purchasers representing a large share of targeted providers' patients are better positioned to introduce significant changes in behavior than those representing a small share of patients.[7, 8] For example, in Hawaii, other than Kaiser, the Hawaii Medical Service Association (HMSA) covers almost all patients with commercial insurance. Large market share facilitated HMSA's introduction of a P4P program in 1998, well before most other purchasers were even considering this option. State Medicaid agencies, which cover roughly half of long-term care spending, would be similarly positioned to implement meaningful incentives for facilities such as nursing homes.[9]

In an equal and opposite way, highly organized providers with high market share are more able to resist changes in incentives that they do not like.[10] The presence of powerful providers does not rule out P4P, however, but implies that the purchaser must consider the provider groups' input.

Another major factor that can facilitate or inhibit a P4P initiative is the local market capacity for quality measurement and improvement. There may be little value in establishing ambitious performance targets based on process or outcome measures if providers have weak information systems and poor office systems for managing patient care. Purchasers in such communities might initially focus on rewards based on measures that do not require well-developed information system capacity, such as patient ratings of their experience with care or measures of infrastructure (so-called structural quality measures).

[i] Critical access hospitals are rural, acute-care hospitals that are eligible for cost-based reimbursement by Medicare based either on State designation as a "necessary provider" or distance from the nearest acute-care facility.

For example, the Bridges to Excellence program and the California Integrated Healthcare Association (IHA) both reward physicians for information technology adoption, in addition to other dimensions of quality. The Medicare Payment Advisory Commission (MedPAC) also has taken this position in its recommendation that CMS focus on encouraging information technology adoption as a first step in introducing P4P for physicians in traditional Medicare.[11] In Massachusetts, under the Massachusetts Healthcare Quality Partnership, capacity building to support P4P has focused on quality measurement and data aggregation across a set of participating health plans. Similarly, purchasers in communities where quality information is scarce might initially focus on rewards for reporting of measures as was done by CMS in the Hospital Quality Alliance (HQA) program.

Question 2. Should we partner with other purchasers or go it alone?

Purchasers face tradeoffs in the decision of whether to coordinate with other purchasers or undertake their own P4P program.[ii] Reasons that a purchaser might find coordination for P4P desirable include:

- The purchaser's market share is small, such that any unilaterally imposed incentive would likely have little or no effect.
- Providers will be frustrated with multiple, uncoordinated data requests from different sources.[12]
- If the purchaser acts on its own, other purchasers will benefit from the P4P program without directly participating and there may be a "free-rider" problem because of overlapping networks.[7, 8]

For these reasons, purchasers may want to coordinate with each other and a broader group of stakeholders to establish agreements about what quality indicators to collect. While there may be substantial benefits from coordination, purchasers will need to consider carefully the antitrust implications of doing so [see box].

Aside from minimizing antitrust issues, some purchasers may be prepared to "go it alone" for the following reasons:

- For a commercial health plan, P4P could be used as a differentiating factor in attracting business from employers and enrollees.
- Purchasers can align their program with their own or broader ongoing data collection and quality improvement efforts. For example, purchasers could design their P4P programs around the data collected by hospitals on cardiac and lung patients for the Joint Commission on Accreditation of Healthcare Organizations (JCAHO), for the Centers for Disease Control and Prevention (CDC) on hospital-acquired infections, and for various specialty societies about procedures such as coronary bypass.[13] Similarly, existing public reports of provider performance sponsored by the CMS, JCAHO, the Leapfrog Group, and others could be used as low cost data sources for pay for performance.[14]

[ii] This means without coordinating with other *purchasers* in the market. See Question 3 for issues in collaborating with providers.

Antitrust Issues: A Consultation With Federal Trade Commission Staff

Joint decisions between or among otherwise competing payers that relate to what/how they will pay providers under a P4P (or any other) program **always** raise an antitrust issue, though not necessarily an antitrust problem. How these types of agreements among competitors are analyzed under the antitrust laws is discussed at considerable length in the joint Federal Trade Commission/Department of Justice Antitrust Guidelines for Collaborations Among Competitors (see: www.ftc.gov/os/2000/04/ftcdojguidelines.pdf) and in the Commission's decision in *Polygram Holding, Inc.*, 5 Trade Reg. Rep. (CCH) ¶15,453 (FTC 2003) (see: www.ftc.gov/os/2003/07/polygramopinion.pdf) (affirmed by the Court of Appeals, *Polygram Holding, Inc. v. FTC*, 416 F.3d 29 (D.C. Cir. 2005). In addition to consulting these resources for further information, seeking advice from competent antitrust counsel would be advisable if payers are contemplating acting jointly, particularly regarding payment/provider contracting issues.

Development or adoption of uniform P4P quality/performance standards and data reporting requirements—rather than actual payment terms by payers under such programs—would appear to be less problematic and easier to justify as pro-competitive. There are potential justifications for such joint activity that relate to the inability of small, individual payers to do this effectively on their own. Competitors working together to make P4P programs possible, more efficient, or less onerous to providers may be more likely to be considered a legitimate justification for joint behavior under antitrust law because of potentially pro-competitive effects in the marketplace.

Overall, from an antitrust perspective it is far less risky to develop or adopt a set of "best practice" standards for P4P programs and allow individual payers to decide whether or not to use the "gold standard" in their programs, rather than agreeing to do so (particularly regarding terms of paying providers). Similarly, adopting uniform standards for data compilation and reporting, so that all programs would have access to a broader data source to evaluate their respective programs, would appear to have far less potential for raising anti-competitive concerns. It is important to note, however, that even otherwise potentially efficient and arguably pro-competitive agreements among competitors regarding this type of activity still could raise antitrust concerns where the competitors together have market power and thus are able to compel acceptance of their standards by providers in the market.

Note: Views expressed by members of the FTC staff do not necessarily represent those of the Federal Trade Commission or of any individual commissioner.

Question 3. When and how should we engage providers in P4P discussions?

Basic options for provider-engagement include information only, advice, and shared decisionmaking. Providers are both a potential ally and a potential source of resistance to P4P. Understandably, providers may have particular concerns about the quality of the data and the validity of measures created using the data.[15, 16] Historically, providers have been very skeptical about data produced by outside stakeholders such as government agencies or employer

coalitions.[16, 17, 18] Physicians also have expressed concern over their ability to influence many outcomes measures of quality—such as smoking cessation or control of high blood pressure—because of the substantial role played by patient actions and preferences.

Attempting to meet provider concerns in the development of a P4P program could improve the effectiveness of the effort and its long-term chances but may also slow down the initiative if there is substantial resistance. The existence of provider-sponsored quality measurement programs—e.g., the Society of Thoracic Surgeons' coronary bypass database—suggests that at least some providers value performance measurement for internal quality improvement purposes. They may not as a whole, however, advocate payments in which some are winners and some are losers. Further, they may not trust groups that are not clinical to develop valid metrics and truly focus on quality (rather than cost). [12, 19] Involving providers early could help purchasers identify performance indicators or measurement systems that meet providers' standards for validity and could facilitate cooperative relations needed to maintain provider participation.

In several communities, public reporting and P4P programs have been successfully developed using a multi-stakeholder approach that involved key providers. The California IHA program is a leading example of this approach. Individual health plans also have successfully involved providers in the development of their P4P programs, despite the sometimes contentious environment that surrounds contracting. In Washington State, for example, the Premera Blue Cross health plan worked closely with the major clinics that provide care to its enrollees to develop performance reporting on both cost and quality measures, and then to use the same performance data to support financial rewards. In the development of the program, the participating physicians often supported tougher standards than the plan initially proposed.

> "*Allowing physicians to be involved at every step of the program's development has been critical to our ability to set meaningful performance goals and truly engage the clinics in quality improvement*."
>
> Mark Sollek, MD, Medical Director, Premera Blue Cross plan of Washington State

Phase 2. Design

In this section, we discuss Questions 4-13, which purchasers need to address once they have decided that they will undertake a P4P initiative. They are:

4. Which providers should we target first? Hospitals or physicians? Specialists or primary care providers?
5. For physicians, what are the advantages and disadvantages of targeting individual clinicians versus medical groups? In the case of hospitals, what are the advantages and disadvantages of targeting individual hospitals versus hospital systems?
6. Should provider participation be voluntary or mandatory?
7. Should we use carrots or sticks—bonuses or penalties—or a combination?

8. How should the bonus be structured?
9. Should we use relative or absolute performance thresholds?
10. What are our options for phasing in P4P?
11. Where do we find the money?
12. How much money should we put into performance pay?
13. What measure characteristics make them attractive candidates for inclusion in an initial measure set?

Question 4. Which providers should we target first? Hospitals or physicians? Specialists or primary care providers?

A recent study suggested that the majority of P4P programs now target both primary care physicians (PCPs) and specialists and about 25 percent target hospitals.[1] Three key factors help determine which types of providers should be the initial focus of P4P programs:

- **Most significant performance (quality or cost) problems.** All else equal, payment incentives should be introduced where the greatest gains may be achieved. Uncovering local quality problems might require claims data analysis but also could be informed by reviewing existing data such as HEDIS® data (health plan report cards); the Dartmouth Atlases, which report a variety of utilization, cost, and quality measures by geographic area; and the National Healthcare Quality Report (NHQR),[20] which annually tracks nearly 200 measures on a nationwide basis and includes measures of care in a variety of settings. In addition, the online State Snapshots based on the NHQR identify potential areas for quality improvement in every State in the Nation.
- **Share of covered services delivered by different categories of providers.** If few covered beneficiaries ever use a type of provider (e.g. rehabilitation facilities), then the value of changing practice patterns may be small.
- **Available performance measures and existing data for each type of provider.** A prerequisite for P4P is that there must be valid and reliable performance measures to capture the relevant dimensions of provider behavior and/or patient outcomes. The existence of a set of validated measures is important, not only for the effective design of the payment system but also for securing the support of providers. There has been a great deal of collective investment in quality measurement focused on certain areas, e.g., preventive care. For some specialist physicians and hospital departments, however, there are few accepted measures of clinical quality of care. Structural measures—such as those found in the NCQA's Physician Practice Connections tool—and patient experience measures—such as CAHPS®—may be applicable to a wide range of physician specialties.

Resources for identifying performance measures that are in use and have been validated include:

- Joint Commission on Accreditation of Healthcare Organizations.
- National Quality Measures Clearinghouse.
- National Quality Forum.

- National Committee for Quality Assurance.
- Hospital Quality Alliance.
- Ambulatory Care Quality Alliance.

In addition, the CAHPS® family of measures offers several validated instruments for measuring patient experience with physicians, medical groups, hospitals, hemodialysis centers, and nursing homes in addition to health plans.

> *"Purchasers should be actively looking to augment these [nationally accepted] measure sets particularly in many specialty areas. P4P should apply to all high-volume specialists and not purely PCPs. Most of the real cost and quality drivers involve chronic disease processes that are more often managed by specialists."*
>
> Nicholas Bonvicino, MD, MBA, Senior Medical Director
> Clinical Network Management, Horizon Blue Cross Blue Shield of New Jersey

Question 5. For physicians, what are the advantages and disadvantages of targeting individual clinicians versus medical groups? In the case of hospitals, what are the advantages and disadvantages of targeting individual hospitals versus hospital systems?

Some payers may have a choice of whether to institute P4P at the level of an individual hospital or physician rather than a hospital system or medical group.[iii] There are pros and cons of targeting incentives at the individual provider versus the medical group or hospital system.

The advantages of targeting incentives at the individual provider are:

- Incentive schemes that directly link payment to those responsible for improving care provide stronger motivation than incentives linked to group behavior. If an individual physician is paid a bonus for the quality of care provided to her own patients, she has more opportunity to influence the chances that she receives a bonus than if she is 1 of 10 physicians whose practice patterns are aggregated for bonus determination.
- Measuring and rewarding performance at the individual physician or hospital level may provide more actionable feedback than relying on more aggregate data and may enhance accountability.

The advantages of targeting incentives at the medical group or hospital system are:

- Many believe that system failures are the key to quality problems and that system reforms are needed to overcome the problems.[21] Moreover, many medical groups and independent

[iii] Some purchasers may want to consider rewarding clinical teams rather than contracting entities, such as medical groups or independent practice associations, because many chronic care models rely on the concept of a clinical team as the locus of care management. Although such an approach would be more consistent with how care is delivered, it would likely pose challenges for data collection and payment since these entities are not generally recognized for contracting or billing purposes.

practice associations would argue that they exist in large part to improve the coordination and quality of care. Providing incentives to "systems" so they can invest in improvement would be more consistent with this idea than paying individuals. For example, Blue Cross Blue Shield of Alabama and the Dean Health Plan in Wisconsin offer incentives for the adoption of electronic medical records (EMRs). Given the investment required to introduce an EMR, targeting individual physicians with an incentive is less likely to drive behavior than targeting groups. This "bigger is better" notion may not extend to hospital systems, where the evidence suggests that the advantages of larger scale operations are limited.

- Rewarding groups of providers may be advantageous is related to errors in the measurement of performance. In some instances, particularly where the clinical process or outcome of interest occurs relatively rarely, there will be random variation in performance (and therefore payment), unrelated to the actions of the provider. With larger numbers of patients, there is less of this type of uncontrollable variation. This problem will be of greater concern for outcome measures generally but can be relevant for process measures that apply to small populations.

Question 6. Should provider participation be voluntary or mandatory?

Many pilot P4P and public reporting programs, such as the CMS/Premier demonstration, are voluntary, which is generally preferred by providers.[22] The major advantage of a voluntary program is the relative ease with which it can be implemented because not all providers need be ready and willing to participate. Voluntary programs will be likely to attract those providers who expect to perform well—usually those that are already performing well[23]—while the poor performers remain on the sideline, which may limit the potential of a voluntary program to improve care among poor or mediocre performers.

Other programs mandate participation in the sense that it becomes a requirement for contracting, such as with most P4P programs implemented broadly (as opposed to pilots) by health plans. The main advantage of a mandatory program is fairness and the ability to promote quality across the market or network. (We note that a mandatory program where P4P takes the form of a bonus may be, in practice, exactly the same as a voluntary program because not all providers will find it worthwhile to respond.)

In practice, the decision of whether to make a program voluntary or mandatory is intertwined with considerations of data availability, the respective clout of providers relative to purchasers in the community, and the basic structure of the P4P program. A mandatory withhold, for example, appears much different from a mandatory bonus, as described above.

Question 7. Should we use carrots or sticks—bonuses or penalties—or a combination?

There is disagreement among researchers and industry leaders on whether threats or rewards are more effective motivators. Some analysts argue that penalties may be more effective motivational tools than bonuses because people view potential losses differently from potential gains.[24, 25] Although some documented evidence supports this theory, the conclusions are

somewhat mixed.[25, 26, 27] Others argue that providers dislike penalty-based approaches and, when faced with such negative incentives, they "game" the system.[28, 29]

In practice, only a few P4P programs—such as the new general practitioner contract in the United Kingdom (UK) and the CMS/Premier hospital P4P demonstration— incorporate penalties for consistent poor performance. And even these programs plan for only very rare use of the penalties. In the first year of the UK program, almost 90 percent of physicians attained the program's maximum rewards, to a large extent because performance goals were set very low.[30]

Similarly, in the CMS/Premier demonstration, CMS agreed that there would be no penalties in the first 2 years and that the penalty threshold for the third year would be set as quality at or below the 10th percentile of performance in the baseline year. Avoiding the penalty requires a relatively low level of quality improvement and all providers have at least 2 years to accomplish this goal.[31, 32]

The impact of these strategies on quality of care is not yet known. Use of penalties to set a floor for performance expectations may prove to be an effective strategy. As overall performance improves, the floor could be moved upward over time.

Question 8. How should the bonus be structured?

The answer to this question in part depends on the overarching aim—to reward high-performing providers versus to encourage improvement.

At least four options in designing a bonus exist (see Table 1):

- Rewarding only those providers that meet or exceed a single threshold of performance.
- Differentially rewarding providers for achievements along a continuum of performance thresholds.
- Rewarding providers that meet or exceed a single threshold of performance combined with incentive rewarding of those that improve, regardless of whether they meet the threshold.
- Rewarding providers in a continuous manner in proportion to their achievement.

The most common approach to P4P is to set a single benchmark level of performance that represents "good" quality and pay a bonus to providers that meet or exceed this threshold. As noted in Table 1, in its first year, the PacifiCare of California Quality Incentive Program rewarded all medical groups that exceeded a single threshold, which was pegged at the previous year's 75th percentile for each measure. This approach is consistent with a strategy to reward high-quality providers (rather than to improve performance) and has the advantage of simplicity. This approach does not uniformly provide incentives for improvement, however. High-quality providers may receive bonuses without making any improvements, and low-quality providers may find the single threshold too difficult to meet and opt not to engage. Some early empirical evidence on the impact of recently implemented P4P programs supports this understanding.[23]

Table 1. Four strategies for designing a bonus structure, with purchaser examples

Strategy	Example
Bonus to providers that meet or exceed single benchmark level of performance, one benchmark for all providers	• PacifiCare of California Quality Incentive Program, year 1: All medical groups that score above the prior-year 75th percentile of performance in the network receive per member per month bonus.
Graduated or tiered bonus based on more than one level of performance	• PacifiCare of California Quality Incentive Program, year 2: All medical groups that score between the prior-year 75th and 85th percentile of performance in the network receive 50 percent of the bonus potential; providers scoring above the 85th percentile receive full bonus. • Bridges to Excellence Physician Office Link: Physicians receive per patient bonus for meeting a set of standards related to office systems that promote quality care; incremental rewards are associated with higher levels of achievement (basic, intermediate, advanced).
Combination of bonus for meeting threshold and bonus for improvement	• Premera Blue Cross of Washington State: Rewards clinics based on process and outcome measures of quality (as well as other efficiency- and access-related metrics). Points, which determine each clinic's allocation, are awarded based both on rank among peers *and* improvement.
Continuous rewards	• Hudson Health Plan (a Medicaid managed care plan in New York): Pays $200 for every 2-year-old who receives all recommended immunizations on time.

As an alternative, purchasers may wish to consider tiered awards, in which differential incentives are offered to providers at different performance levels, such as 70 percent compliance, 80 percent compliance, 90 percent compliance. The more thresholds, the greater the likelihood that providers at different levels of quality performance will have an incentive to engage and improve. Again using the PacifiCare example, in the second year, it offers the full bonus to groups whose performance is above the prior year's 85th percentile level and 50 percent of that amount to groups that perform above the prior year's 75th percentile but below the 85th percentile.

Alternatively, purchasers might explicitly tie payment to improvement either in addition to or instead of a benchmark level of attainment. Premera Blue Cross of Washington State rewards clinics based both on their rank among peers and the degree of improvement over the prior year.

For measures that reflect concerns about underuse of effective services (e.g., retinal exams for patients with diabetes), another alternative would to pay an additional fee for each appropriately managed patient or for each "recommended" service that the purchaser is targeting. Unlike setting a bonus threshold at a single level, under the additional fees-for-service model, physicians always do better financially by bringing more patients into compliance with the standard.

Although the incentive properties of rewarding improvement or using additional fees each time a service is performed are preferable to a single fixed threshold, some may object in principle to rewarding physicians at levels of performance that are below acceptable norms (whatever these are.) To accommodate such concerns purchasers could set a minimum threshold—such as 60 percent adherence to the evidence-based guideline in question—below which physicians are ineligible for any payment.

Question 9. Should we use relative or absolute performance thresholds?

Asked another way, should the incentive be structured such that all providers could theoretically receive some reward, or should we structure the program such that there are only a limited number of winners?

In contrast to Question 8, which examines he relationship between performance and payments, Question 9 addresses whether providers compete against one another or are held to some external standard. Many current P4P programs pay bonuses based on the ranking of performance relative to other providers in the network. For example, Anthem Blue Cross Blue Shield of New Hampshire rewards physicians whose performance on clinical quality measures places them in the top two quartiles of the distribution (with larger bonuses for the top quartile). This type of reward structure is sometimes referred to as a *tournament*.

Tournaments may be desirable for the following reasons:

- Relative performance measures can filter out common sources of uncontrolled variation in performance. For example, if a purchaser who wanted to target flu shots for improvement compared an individual physician's performance in 2004 against 2003, the physician's quality might appear to have declined in 2004 due to a decrease in vaccination rates, even though these lower rates primarily reflected vaccine shortages over which the physician had no control. However, examining the change in vaccination rates over time nationally or among physicians in the same market would produce a different picture of physician efforts to improve quality.
- Tournaments provide strong incentives to improve continuously because there is no level at which it is guaranteed that a provider will be ranked sufficiently high to receive a reward.
- Because not everyone receives a bonus, a tournament program with the same maximum bonus potential for those who will receive one will cost less than a program where all providers could get the bonus.

Important disadvantages of tournament-style rewards also exist, such as these:

- Because providers cannot be certain beforehand what level of performance must be achieved to result in a bonus payment, they may judge investments in quality improvement to be unacceptably risky.
- Providers that have already determined how to deliver good-quality health care along the targeted dimensions will be at an advantage (the same is true with a non-tournament program with a single high threshold). Providers that are ranked low among their peers are less likely

to find it worthwhile to strive for these bonuses because of the low likelihood of surpassing the competition. This rewarding of historical investments in quality, although possibly justified, may not yield as much quality improvement across the population as other approaches.

- When providers know their payment may be determined by relative performance, they may be less willing to engage in one of the most commonly used quality improvement models—the local collaborative in which successful local providers advise and assist less successful ones.

For questions 8 and 9, purchasers need to decide whether the *primary* goal of their P4P program is to improve the quality of care delivered by all eligible providers or to begin paying more to high-quality providers than to low-quality providers. These objectives are not incompatible, but some approaches to P4P (in particular, using tournaments or high fixed benchmarks) will favor the latter. Alternatively, the approach of paying additional fees-for-service achieves both goals, since higher performing providers receive more fees but all providers have a reason to improve.

Question 10. What are our options for phasing in P4P?

Most purchasers that have introduced P4P have started in a limited way and expanded over time. Advantages for phasing in P4P are that it permits testing of measures before full scale implementation, gives providers time to gear up for a P4P initiative; and enables purchasers to evaluate the small scale impact before applying it to the larger group of providers.

Options for phasing in P4P include the following:

- Pilot test a payment scheme in a limited geographic area.
- Focus on specific provider types or clinical areas.
- Begin with pre-existing, national measure sets and add measures over time.
- Rely on existing data (most likely billing data) and incorporate additional data as needed over time.
- Begin with a voluntary system.
- Begin with private quality reports and introduce incentives over time.
- Begin with a modest benchmark for performance and raise the standard over time.
- Begin with requiring or rewarding data collection and reporting and introduce performance incentives over time.

The CMS experience with hospital incentives illustrates one approach to phasing in a P4P effort. CMS introduced a pay-for-reporting program to encourage hospital participation in the Hospital Quality Alliance, in which participating hospitals receive 0.4 percent of their payment update if they publicly report a set of quality measures; non-participating hospitals lose this revenue stream. Because of the large market share represented by Medicare, more than 98 percent of hospitals nationwide report on the set of measures.

Question 11. Where do we find the money?

Potential sources of funds for a P4P initiative include:

- New money.

- Redirection of annual payment updates.
- Reallocation of payment among providers, e.g., through a combination bonus-penalty payment scheme.
- Cost savings resulting from improved quality and special cases of shared savings.
- For Medicaid, disproportionate share funds, and for the special case of Medicaid managed care, preferential auto-assignment formulas, which provide financial incentives in the form of greater volume of patients.

Many private payers that have introduced P4P programs frame the bonus potential as "increased" payments to providers, but it is difficult to imagine sustaining such increases given the recent double-digit growth rates in spending. If performance pay is to account for more than a small share of provider compensation in the near term, there will have to be significant redistribution (winners and losers); savings will be needed to at least partially offset the additional costs associated with improving quality; or, most likely, the funds will come from cumulatively directing all or a portion of annual updates to incentive pay. From an employer perspective, the possibility of offsetting savings should account for increased employee productivity.

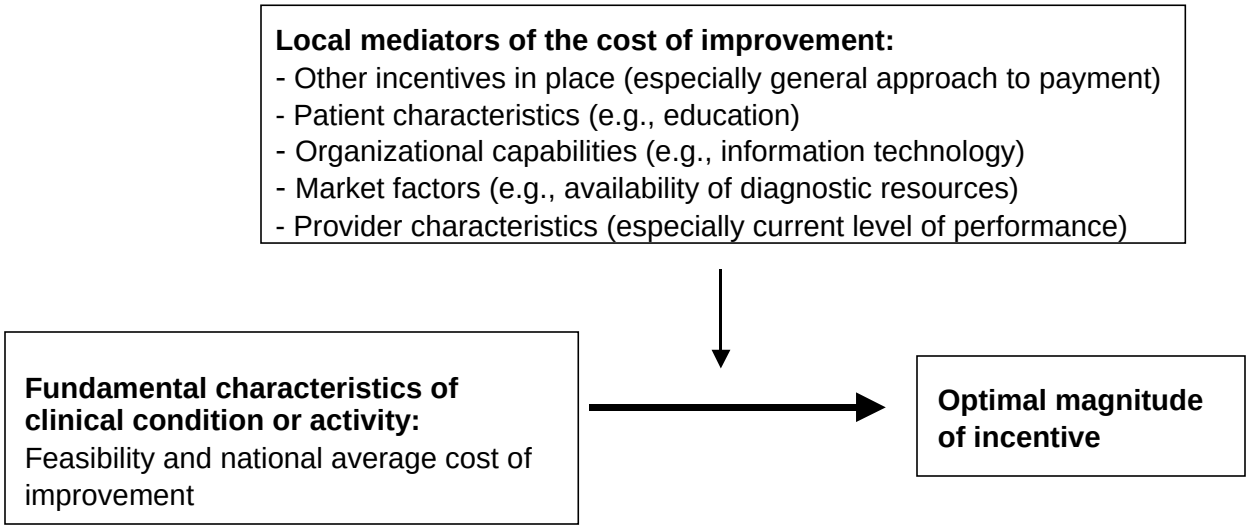
Whether improved quality will pay for itself in financial terms is an issue of some debate. In general, however, it is clear that the question of whether the costs of quality improvement will be offset by savings (e.g., from reduced hospitalizations) will first depend on whether the incentive seeks to remedy misuse, underuse, or overuse. Even within each of these categories, however, there will be differences by clinical area. For incentives to correct underuse of cancer screening, for example, most of the health and financial gains are long term, so a positive financial return in the short term is unlikely. Reducing underuse of prescription drugs and educational services for patients with chronic illnesses and substantial risks of high-cost hospitalizations or procedures may be more likely to yield savings in the near term.

At present, for many of the most commonly used measures of quality (e.g., HEDIS® measures), the promise that improvement will result in cost offsets sufficient to ensure financial sustainability of pay for performance is uncertain at best. In light of these clinical realities and current budgetary constraints, purchasers might choose to explicitly incorporate cost-saving measures into their P4P programs. For example, the Wellpoint Physician Quality Incentive Program rewards physicians for generic prescribing and administrative efficiency measures such as electronic claims filing. The Integrated Healthcare Association in California is also currently evaluating the inclusion of an aggregate cost-efficiency measure in its P4P program. On the other hand, there is some risk that orienting a P4P program toward cost control may undermine the credibility of the quality improvement aspects of pay for performance.

Question 12. How much money should we put into performance pay?

There is no single answer to the question of how much money is needed. Some P4P schemes have provided as little as $2 per visit and had an impact, while others offering bonuses of up to $10,000 had no effect.[33, 34] One can, however, identify a number of factors relevant to the decision about how large performance pay needs to be (see diagram).[5]

Factors Affecting the Necessary Size of Incentive



Key issues include the following:

- **Characteristics of the clinical condition or treatment**. Some changes are easier to achieve than others. It is easier to get patients to take flu shots than to quit smoking. Some interventions are less costly than others, even among screening tests. For instance, Pap smears are much less expensive than colonoscopies. Improving performance in areas with good feasibility and low cost should require smaller incentives than improving results in other areas.

- **Other incentives already in place**. For example, if medical groups are capitated for their services, then incentives to increase screening tests would need to be larger than in a fee-for-service system in which providers already receive basic fees for the associated visits and procedures.

- **Organizational capabilities**. Larger groups may have the resources to hire a dedicated asthma patient educator and ensure excellent communication between pulmonologists and primary care providers, while smaller groups and solo practices may find patient education and inter-provider communication more difficult.

- **Patient and market variables**. Providers with highly educated patients traditionally experience better patient adherence and cooperation, which may affect their performance ratings.[35] Rural diabetics may have a harder time getting eye exams than their urban counterparts because of a dearth of local ophthalmologists. Market share of the payer may also be a factor in determining the necessary size of the bonus, particularly if investments in infrastructure or training are needed to achieve the quality goal. A purchaser with large market share, like Medicare, may be able to promote change with a relatively smaller proportional bonus compared to a purchaser with small market share.

In light of all the uncertainty on this topic, it is not surprising that the P4P programs in place today—most of which are fairly new—typically place 5 percent or less of contracted revenues at risk for performance, although there is some indication that the amounts at risk are increasing.[1,

[36] In the case of hospital programs, the percentages are often lower. For example, the CMS/Premier demonstration involves a reward of 1 to 2 percent for top-performing hospitals.

Another perspective on how much to pay to improve performance can be found by considering shared savings where savings are anticipated from quality improvement. This approach has been used by purchasers such as the Alliance of Wisconsin as well as the Bridges to Excellence program. For example, if hospitals reduce complication rates among patients receiving a particular procedure and those avoided complications save the purchaser $10,000 in additional treatment costs, the hospital might receive 50 percent or $5,000 of those savings.

Question 13. What measure characteristics make them attractive candidates for inclusion in an initial measure set?

Measure types span structure, process, and outcome and include technical (clinical) as well as interpersonal attributes of care. Developing a robust measure set is crucial to P4P success. Not surprisingly, surveys of providers indicate that performance measurement that lacks clinical face validity or sufficient scope and sophistication will be poorly received and actively resisted.[12, 37]

Table 2 lists characteristics to consider in evaluating candidate quality indicators. One major issue is whether the indicator generates information about a single condition (e.g., use of appropriate antibiotics in pneumonia) or is relevant to a broad population (e.g., rates of medication errors). Although measures that apply to larger numbers of patients are attractive, the disadvantage of this approach is that precise measurement definitions and standards are less often available for process or outcome measures of this type; valid measurement may require adjustment for differences in the types of patients across providers (and methods for such adjustment may not be developed for some measures). On the other hand, precisely defined, condition-specific measures suffer from a general lack of availability for many diseases and treatments.

Table 2. Indicator characteristics to consider in developing a measure set

1. Does the indicator measure care that is a priority for quality improvement?
2. Does the indicator apply to a single disease or across multiple patient groups?
3. Does the indicator generate information about cost efficiency, health care processes, outcomes, or structure?
4. Does the indicator reflect technical competency or patient experiences with care?
5. Is the indicator actionable?
6. Is there a valid source for the data needed to calculate the indicator? What is the cost of acquisition and validation of those data?
7. Is the indicator nationally accepted or locally developed?

Other considerations include the following:

- Providers generally prefer process measures, which assess whether the right clinical decision was made and the appropriate diagnostic test or treatment was used, rather than outcomes, which are more strongly influenced by patient factors beyond a provider's control.[12, 18, 38]

- Structural measures—such as the volume of procedures a provider performs or their capacity for computerized order entry—have been favored by some purchasers because they do not require collection of detailed clinical data and can be measured by survey. This approach largely avoids the issue of patient differences, but structural measures are often only weakly related to outcomes. In addition, this strategy runs counter to the idea that incentives should be established to encourage suppliers to find the most effective and efficient production systems on their own.
- Some purchasers may wish to reward the reduction of disparities in the quality of care or access. Reductions in differences in quality would not be appropriate quality measures to use as the basis of rewards because differences could be reduced by decreasing the quality of the better served group. Instead, purchasers could provide incentives for improving care to the underserved group.

Decisions about measures require evaluation of sources of data. The main sources, in order of increasing expense of data collection, are:

- Pre-existing administrative databases (generally created through the submission of claims and/or discharge abstracts) or data that have been collected for another purpose such as accreditation.
- Provider surveys.
- Patient surveys.
- Medical record abstractions.

Each data source comes with its own set of strengths and limitations:

- Administrative data are readily available and algorithms for using them to examine the quality of care are established, although providers may not believe those algorithms yield valid performance measures.[39, 40]
- Administrative data are a reasonably good source of process information, although this is less true in the hospital setting, in part because hospitals are typically paid a set fee per day or per discharge so that details about individual therapies that a patient received while admitted are not captured.
- Administrative data yield fewer outcome measures than medical records and contain few of the variables perceived as necessary for risk adjustment of those outcomes.[19, 41] It is noteworthy that an increasing number of health plans capture pharmacy claims and lab results in their electronic data systems, which strengthen a purchaser's ability to judge quality of care through claims data.
- In general, provider acceptance of the validity of the data is least for administrative data and greatest for medical record data.[12]
- Chart abstraction, done correctly, can address many of the limitations of administrative data, but it is expensive. In the future, information technologies may be adopted that greatly reduce the cost of collecting the data generally sought through chart abstraction, but implementation of electronic medical records with such capabilities has been slow.
- Provider surveys are the most feasible way of collecting information on structural measures (e.g., whether a hospital has computerized order entry) but are limited by the reliability of self-report and the fact that standardized methods for auditing them are not yet available.

- Patient (or family) surveys are the source for information about patients' experiences, and there are validated survey measures that could be readily used for almost any provider type. Patients are less reliable sources for technical information about their own diagnoses and care.[42, 43]

Another major tension in measure selection is the choice between using nationally adopted indicators versus developing local measures. When feasible, it is clearly preferable to use measures endorsed by CMS, JCAHO, National Committee for Quality Assurance, the Hospital Quality Alliance, the Ambulatory Quality Alliance (AQA) or the National Quality Forum (NQF). However, the number and scope of measures available from these sources is limited. There are more indicators endorsed by NQF, but the work of developing measurement specifications is ongoing, so one cannot implement all NQF measures at the current time.[iv] (Table 3 presents specific examples of various types of quality measures currently used by purchasers.)

Table 3. Types of quality measures, with purchaser examples and specific measure used

Type of measure	Purchaser example and measure
Structure	• Empire Blue Cross Blue Shield: Leapfrog Group measures including computerized physician order entry and staffing of intensive care units with intensivists.
Process	• Integrated Healthcare Association, year 1: Hemoglobin (Hb)A1c testing, LDL cholesterol testing, childhood immunizations, cervical cancer screening, and mammography.
Health outcome	• Premera Blue Cross of Washington State: HbA1c, LDL cholesterol, and blood pressure control, among other measures.
Patient experience	• Integrated Healthcare Association, year 1: 40 percent of P4P is based on the following patient satisfaction measures: 1) satisfaction with specialty care, 2) timely access to care, 3) doctor-patient communication, and 4) overall ratings of care.
Locally developed measures	• Hawaii Medical Services Association: Locally developed measure of surgical complications.
Nationally developed measures	• Empire Blue Cross Blue Shield: Leapfrog Group measures.

Note: Some purchasers may use a mix of various types of measures.

On the other hand, local development of measures may be advantageous for two reasons: 1) developing measures that are relevant to a local population and delivery system may be an effective means for engaging providers, and 2) there may be important local public health priorities for which nationally vetted measures do not exist.

[iv] Summaries of NQF reports on issues in developing quality measurement specifications and recent NQF-endorsed consensus standards may be found at www.qualityforum.org.

"[W]e do not yet have many quality areas with generally accepted measures established. Pioneers in this work will have to validate their own measures much of the time."

James Mortimer, former President, Midwest Business Group on Health

Phase 3. Implementation

Questions 14-18 need to be addressed in the implementation of a P4P program:

14. How do we address providers' concerns about whether risk adjustment adequately captures the severity of illness of their patients?
15. If we currently sponsor a private or public report card, will P4P offer more of an incentive? If we are considering both a public report and P4P, which should we pursue first?
16. Should we tailor P4P for subsets of a particular group of providers, e.g., safety-net hospitals?
17. How should we think about P4P and its relationship to benefit design, including tiered networks?
18. Is there any special advice for Medicaid agencies and Medicaid managed care plans interested in P4P?

Question 14. How do we address providers' concerns about whether risk adjustment adequately captures the severity of illness of their patients?

Providers who treat a larger proportion of higher risk or less adherent patients may receive lower ratings on process and outcome measures, despite making equal efforts to practice high-quality care. Thus, providers legitimately want to make sure that a P4P program accounts fairly for patient differences. Risk adjustment models to correct patient outcome estimates (usually mortality rates) for underlying differences in patient populations have been under development for many years.[19, 41] Nonetheless, providers worry about the adequacy of risk adjustment.[18, 38, 44] Furthermore, refusal to address such concerns may threaten the legitimacy and sustainability of any incentive program.[10, 45, 46, 47]

Risk adjustment is generally less effective when administrative data are used because detailed clinical information (e.g., blood pressure) is typically unavailable. Analysts have shown, however, that in some cases the addition of a few simple clinical variables to administrative data would be sufficient to make risk adjustment comparable to that which can be achieved with the sophisticated databases many specialty societies have developed.[19, 41] This is especially the case in States such as California that are adding "condition present on admission" indicators to their administrative data to distinguish pre-existing comorbidities from treatment-related complications.

A P4P sponsor could engage providers in the design of a clinical data collection system that is either consistent with one of the growing number of national databases (specialty societies, CDC,

JCAHO, others) or a less burdensome augmentation of administrative databases and test whether the additional data actually make a difference in the distribution of rewards.

Finally, some approaches to P4P will be less sensitive to differences in patient characteristics than others. In particular, if a purchaser decides to reward providers for improvement relative to their own baseline rather than for meeting a common standard, risk adjustment will be less of an issue than if a tournament approach is used where only the top ranked providers receive a bonus.

Question 15. If we currently sponsor a private or public report card, will P4P offer more of an incentive? If we are considering both a public report and P4P, which should we pursue first?

No studies have compared the effects of report cards relative to P4P. There is evidence that providers respond to public reports about their performance,[18, 48, 49] although hospital executives have indicated that their response to public reporting may wane over time, especially if there are no supporting financial incentives.[12, 45] Thus, the approaches may best be viewed as complementary, rather than mutually exclusive.

Public reporting may be part of a phase-in strategy for P4P; this appears to be the strategy chosen by CMS in the case of the Hospital Quality Alliance data, although the specifics of a P4P program for hospitals have not yet been determined. An advantage of this approach is that it gives providers time to improve their data collection and become more proficient in using methods of performance measurement before the measures become economically significant. This may facilitate the use of a measure set of greater scope than would be acceptable to providers if P4P were to start with the initial measurement period.

In some cases, public reporting and P4P may differ somewhat in focus. For example, research has shown that it is preferable not to include a large number of technical quality measures in a public report card if the goal is to affect consumer choice. So a report card might display a few composite measures of evidence-based care and patient experience, while the P4P program could separately target specific processes and outcomes where the purchaser has identified a shortfall in quality.

Question 16. Should we tailor P4P for subsets of a particular group of providers, e.g., safety-net hospitals?

Providers treating patient populations that are low income and/or have low educational attainment or literacy may be disadvantaged by a "one size fits all" approach to P4P because these communities have poorer health behavior than others (patient differences could also affect patient experience of care, for example, because of cultural issues). To the extent that a payer is concerned about improving performance of all providers or is particularly interested in reducing disparities in the quality of care, a more targeted approach might be warranted.

Purchasers could tailor a P4P initiative in a variety of ways:

- Purchasers could make the reward larger for some providers—either those providers with the lowest performance ratings or, for example, safety-net providers. One argument for

increasing payments is that the costs of improving care will be greater for some providers because of geographic, linguistic, financial, and other barriers that they or their patients face or a lack of infrastructure and poor human resource capacity for quality improvement.

- Purchasers could provide capital grants and/or technical assistance to poor-performing providers again as a way of offsetting their presumed higher costs of complying with performance standards. Independent Health in New York, for example, assists providers serving large numbers of Medicaid patients in planning quality improvement programs.
- Purchasers could allow performance measures to vary across providers. Again, Independent Health involved providers with large numbers of Medicaid patients in the selection of site-specific quality metrics.[50]

A final strategy for tailoring P4P would be for purchasers to set lower performance standards for certain kinds of providers that have lower performance or fewer resources—for example, small practices or rural hospitals. To illustrate, a plan might provide a bonus to all urban hospitals that give at least 90 percent of their patients beta-blockers after a heart attack but advise rural hospitals (who in this example are assumed to have lower rates of beta-blocker usage) they need only achieve 80 percent adherence to receive a bonus. The important argument against this approach is that it will institutionalize disparities in quality. For this reason, approaches that differentially empower low-resource providers and those serving disadvantaged populations are preferred.

Question 17. How should we think about P4P and its relationship to benefit design, including tiered networks?

P4P programs have been implemented in the context of health maintenance organizations (HMOs), point-of-service plans, preferred provider organizations, indemnity plans, and consumer-directed health plans.[1] In principle, provider incentives can be established independently of benefit design, but in practice there will be important interactions to consider, including assignment of accountability and alignment of physician and patient incentives.

The first consideration is *assigning accountability*. In many HMO arrangements, patients must select a physician or medical group to act as a primary care "home" and possibly as a gatekeeper for referrals. These providers will then be a natural unit of accountability for the quality of primary prevention and chronic illness management. In contrast, in a setting where patients do not have identified or assigned primary care providers, attributing responsibility becomes somewhat more complex, but not insurmountably so.

Two basic strategies for attribution of responsibility for the quality of care of individual patients based on contact have been used in practice, each with advantages and disadvantages: 1) all physicians with a minimum level of contact are accountable for a patient's care; or 2) a primary responsible physician is determined retrospectively based on contact.

With regard to the first approach, if multiple physicians share responsibility for delivering a specific test or service, all have a reason to ensure quality, but shirking of responsibility also might occur. In addition, physicians might order redundant tests or services if they do not receive information about services provided by the other physicians.

With regard to the second strategy, a key disadvantage is that during the course of the year, physicians will be uncertain as to whether any given patient will affect their performance estimate because attribution is determined retrospectively.

The second important connection between P4P and benefit design is the *congruence of physician and patient incentives*. Although there is no empirical evidence of a connection, it is logical to conclude that patient and provider incentives each will be more powerful if they are aligned. For example, some purchasers in the Bridges to Excellence Diabetes Care Link program offer their employees rewards for participating in improving the management of their diabetes.[51] Similarly, purchasers who have constructed or are considering tiered provider networks may want to consider focusing on the same sets of performance measures for P4P to intensify the impact.[v]

Question 18. Is there any special advice for Medicaid agencies and Medicaid managed care plans interested in P4P?

In many States— including Michigan, Pennsylvania, and New York—Medicaid agencies offer auto-assignment and/or financial bonuses to managed care organizations that perform well on clinical quality and patient satisfaction measures. Medicaid managed care organizations also have implemented P4P. For example:

- The Local Initiative Rewarding Results program in California offers financial rewards based on the quality of ambulatory care for MediCal beneficiaries.
- Hudson Health Plan, a Medicaid managed care plan in New York, also has a number of P4P initiatives including rewards for childhood immunization and effective management of patients with diabetes.
- The Neighborhood Health Plan of Rhode Island uses P4P to target asthma care.
- In North Carolina, the Primary Care Case Management program has introduced both financial bonuses and recognition for physicians that either reach a best practice performance goal (85th percentile of baseline performance) or improve by 20 percent and exceed the median level of baseline performance. Performance measures in the first incentive year (through June 2006) are related to care for asthma, diabetes, and prescribing patterns.

Purchasers such as Medicaid and Medicaid managed care plans face many of the same obstacles discussed above, particularly with regard to the need to protect safety-net providers and their patients (see Question 16). In addition, constrained Medicaid budgets have resulted in below-market provider reimbursements so that program participation is an ongoing concern. These issues highlight the need to involve providers early and continuously in the development and evolution of an incentive program. The experiences of two New York Medicaid plans corroborate this observation. The Hudson Health Plan focused intently on provider communication. Health Now management developed its initial P4P program internally, albeit with the intention of creating a program that providers would find easy to understand and implement. A survey by the Center for Health Care Strategies found better provider acceptance

[v] Some measures, however, may be appropriate for tiering but not for P4P—for example, the volume of certain kinds of procedures.

of the Hudson Health program than the Health Now program, and Health Now has moved to increase provider participation in the redesign of its program.[50]

In addition, because of particular concerns with patient adherence in populations with low literacy and other challenges, Medicaid programs and plans may find it particularly beneficial to emphasize patient incentives alongside provider incentives, which is likely to improve provider perception of the P4P program as well. Patient incentives are currently in use by some Medicaid managed care plans to encourage appropriate use of services such as adolescent wellness visits and prenatal care.[50] Executives at CalOptima, a Medicaid managed care program in California, believe that participants in a beneficiary incentive program in which department store gift cards are offered for adherence to preventive care recommendations are more likely to receive appropriate immunization and prenatal care.[50]

Medicaid programs may wish to consider P4P in one market in which they are the dominant payer and thus could have substantial impact: nursing home care. Legislation passed in 2005 in Ohio outlines such a program and sets aside 2 percent of average payments to be allocated to the best-performing facilities with regard to a set of structure, process, and outcome measures of quality (and casemix). Performance data on nursing homes are currently being collected and publicly reported by CMS; these data would be a natural platform for P4P. In addition, CMS has recently begun designing a nursing home P4P demonstration project, which may provide both momentum and information for State Medicaid agencies interested in implementing programs of their own.[52]

Finally, Medicaid programs will need to consider regulatory requirements, particularly if they intend to receive a Federal match for the payment incentive (see box).

Programmatic Issues for State Medicaid Programs Considering Pay for Performance

The method by which a State may choose to accomplish its quality-based purchasing program can vary greatly because of the variety of approaches available to a State to administer its Medicaid and State Children's Health Insurance Programs. In general, States have broad flexibility, within established Federal regulations, to decide on medically necessary services that will be covered and rates that will be paid to providers or plans. CMS may review these plans through a State plan or a Medicaid demonstration project application or amendment and through various other mechanisms.

In general, if the pay-for-performance program is a part of a fee-for-service delivery system, a State may include its initiative in its State plan. While the requirements for payment for managed care are somewhat more complicated, CMS will work with States to determine the proper method to implement such an initiative. A waiver under Sections 1115, 1915(b), or 1915(c) of the Social Security Act may be necessary when the initiative will not be statewide; will impact the amount, duration, and scope of benefits; will affect the comparability of benefits across the eligible population; or will restrict beneficiary choice of provider.

Source: Jean Moody-Williams, Centers for Medicare & Medicaid Services.

Phase 4. Evaluation

P4P programs are a work in progress and, because there is little evidence as to the effects of specific approaches, will need to be monitored and improved on an ongoing basis. Although evaluation will naturally follow implementation, the two questions in this section need to be asked during the design phase to assure that the implementation of the program will support meaningful evaluation. They are:

19. How can we tell if the P4P program is working?
20. What unintended consequences should we look for?

Question 19. How can we tell if the program is working?

Learning about the impacts of a P4P program can be particularly challenging because a multitude of additional forces simultaneously affect the quality of patient care and costs. Ideally, purchasers would implement P4P in one market or sub-market and track the same performance measures on a set of comparison providers. Some large purchasers and CMS may be in a position to implement P4P is this way, but most purchasers will not design their programs as controlled trials. Therefore, some care is needed to disentangle the effects of the program from other trends. At a minimum, purchasers should collect baseline data on the targeted quality measures (this will be a critical part of implementation too, of course, because providers without a clear understanding of their performance can hardly be expected to respond optimally to P4P). Then, as performance data are collected for payment purposes, the main effect of the program can be evaluated in terms of the change in performance, preferably compared either to some comparable but unaffected population or the trend in performance prior to implementation.

Purchasers will have to decide how rigorous an evaluation needs to be to ascertain whether a program is working and how to improve it. To adhere strictly to scientific standards of evidence may be too costly and produce evidence too late to be useful for decisionmaking. On the other hand, erroneous conclusions that may be drawn from anecdotal or incomplete information may have substantial costs as well.

Question 20. What unintended consequences should we look for?

In addition to the hoped-for effects of the program, purchasers will need to monitor, and try to minimize, unintended negative consequences. Three important negative effects to look for are patient selection, diversion of attention away from other important aspects of care, and widening gaps in performance among providers.

- **Patient selection.** Providers may avoid sicker patients in the belief that risk adjustment is not adequate and that caring for such patients will reduce their measured performance. Surveys done after New York instituted public reporting for coronary bypass found that two-thirds of cardiac surgeons admitted to avoiding the most severely ill patients.[53] To minimize the potential for the P4P program to result in selection of the "easiest" patients or exclusion of high-risk or non-adherent patients, purchasers can focus on structural or process measures of quality. Risk adjustment of performance measures, particularly those that relate to patient

outcomes such as complication or readmission rates, should help to minimize selection incentives as long as providers believe the risk adjustment is adequate. In addition, including explicit reporting of casemix data—which would show providers who are avoiding or accepting the more difficult cases—or providing differential rewards for meeting performance goals with more difficult patients could increase providers' willingness to take on these cases. Another possibility would be to collect and report information about patients who change from one provider to another. A provider who was avoiding sicker patients would be identified by the high casemix scores of patients leaving his practice.

- **Diverting attention from other aspects of care**. Targeting specific performance measures may focus provider attention on the conditions or care processes for which there is measurement and payment, to the detriment of performance in other areas.[15] At a minimum, this problem suggests the need for careful measure selection and attention to interrelationships among targeted and untargeted domains of performance. Rewarding providers for performance on some broader measures of outcome, such as patient experience or decubitus ulcer (bed sore) rates and pain scores in hospitals, would mitigate this problem as well.
- **Widening performance gaps.** This may be particularly likely to occur if the purchaser chooses to reward only providers that meet a high standard of performance or those that are the highest ranked among peers. If P4P results in a substantial redistribution of resources then some providers may actually worsen with respect to quality of care. This will be a particular concern if those providers serve large numbers of beneficiaries/enrollees or are part of the safety net, and/or if there are not enough suitable choices for the population that receives care from these poor-performing providers. If these adverse consequences are anticipated or noted, purchasers can consider the solutions described in Question 16 above.

These examples give important clues about what evidence to seek in evaluating programs for unintended consequences. Clinician feedback should be sought about unexpected problems with the measures used, including difficulties with both access to care and pressure to offer inappropriate care. Since such data would come from clinician surveys (and unhappy clinicians would be expected to be motivated to respond), getting this feedback should not be too burdensome. Similarly, purchasers should consider tracking a set of performance indicators that are outside of the P4P program to better understand both negative and positive spillover effects from the program onto untargeted clinical domains. Finally, evaluation of the program should not just look at average performance but at the effects of P4P on different parts of the delivery system including providers with high and low baseline performance.

A Final Note—Sustaining Quality Improvement

Even the best-designed P4P program will require maintenance. For example, if the program uses fixed targets, the targets will need to be advanced as providers improve. We note, however, that if providers see that targets are fully adjusted to reflect gains in prior year performance, incentives to improve quality in the current period may be dampened. For most measures, there are also natural "ceiling" effects that will lead to diminished opportunities to improve quality over time. As adherence rates to evidence-based guidelines approach 100 percent, the incremental cost of improving quality is likely to increase as only the cases that failed to respond to initial quality improvement efforts remain.

As clinical evidence about best practices changes, structural (e.g., information technology requirements) and process measures will also need to be updated. Purchasers will have to balance the need to keep P4P programs effective by retiring measures that are no longer useful against the concern that P4P programs provide some stability so that providers can undertake larger investments with the expectation that the reward structure will not be dramatically altered in the short run (and hence a reasonable return on investment can be expected). To this end, explicitly including providers in the decisions about measure selection and retention may be desirable. One approach that has been adopted by some programs, including the California IHA, is to commit to medium-term plans (2 or 3 years) with regard to measure sets and introduce measures in a "testing set" prior to their full inclusion.

To the extent possible, purchasers should use their P4P programs to promote continuous innovation rather than institutionalize a single approach to delivering high-quality care. This concern might be addressed by rewarding, at least in part, outcome measures. Vigorous attempts to keep structure and process measure targets up-to-date with the latest technology will also reduce system rigidity, but political and bureaucratic barriers to change will be inherently limiting.

References

[1] Baker G, Carter B. The evolution of pay for performance models for rewarding providers. Introduction to: Case Studies in Health Plan Pay-for-Performance. Washington, DC: Atlantic Information Services; 2004.

[2] Casalino L, Gillies RR, Shortell SM, et al. External incentives, information technology, and organized processes to improve health care quality for patients with chronic diseases. Journal of the American Medical Association. 2003;289(4):434-41.

[3] The Leapfrog Group. Leapfrog Compendium. <www.leapfroggroup.org/ircompendium.htm> Accessed August 25, 2005.

[4] Steiger, B. Poll finds physicians very wary of pay-for-performance programs. The Physician Executive, November/December 2005: 6-11.

[5] Dudley RA, Frolich A, Robinowitz DL, et al. Strategies To Support Quality-based Purchasing: A Review of the Evidence. Technical Review 10. (Prepared by the Stanford-University of California San Francisco Evidence-based Practice Center under Contract No. 290-02-0017). AHRQ Publication No. 04-0057. Rockville, MD: Agency for Healthcare Research and Quality. July 2004.

[6] Baker G. Pay for Performance Incentive Programs in Healthcare: Market Dynamics and Business Process. San Francisco, CA: Med-Vantage; 2004

[7] Glied S, Zivin J. How do doctors behave when some (but not all) of their patients are in managed care? Journal of Health Economics 2002; 21(2): 337-53.

[8] McGuire TG, Pauly MV. Physician responses to fee changes with multiple payers. Journal of Health Economics. 1991; 10(4):385-410.

[9] O'Brien E. Long-term care: understanding Medicaid's role for the elderly and disabled. Washington, DC: Henry J. Kaiser Family Foundation. November 2005.

[10] Neuhauser D, Harper DL. Too good to last: did Cleveland Health Quality Choice leave a legacy and lessons to be learned? Quality & Safety in Health Care. 2002;11:202–3.

[11] Medicare Payment Advisory Commission. Strategies to Improve Care: Pay for Performance and Information Technology. Chapter 4,Report to Congress: Medicare Payment Policy. March 2005.

[12] Mehrotra A, Bodenheimer T, Dudley RA. Employers' efforts to measure and improve hospital quality: determinants of success. Health Affairs (Millwood). 2003; 22(2):60-71.

[13] National Health Foundation. Findings from the Survey of Hospital Performance Measurement Activities in California. Oakland, CA: California Healthcare Foundation. 2004.

[14] Shearer A, Cronin C, Feeney D. The State-of-the-Art of Online Hospital Public Reporting: A Review of Forty-Seven Websites. Easton, MD: The Delmarva Foundation; 2004.

[15] Werner RM, Asch DA. The unintended consequences of publicly reporting quality information. Journal of the American Medical Association. 2005;293(10):1239-44.

[16] Luce JM, Thiel GD, Holland MR, et al. Use of risk-adjusted outcome data for quality improvement by public hospitals. Western Journal of Medicine. 1996;164(5):410-4.

[17] Berwick DM, Wald DL. Hospital leaders' opinions of the HCFA mortality data. Journal of the American Medical Association. 1990;263(2):247-9.

[18] Romano PS, Rainwater JA, Antonius D. Grading the graders: how hospitals in California and New York perceive and interpret their report cards. Medical Care. 1999;37(3):295-305.

[19] Jones RH, Hannan EL, Hammermeister KE, et al. Identification of preoperative variables needed for risk adjustment of short-term mortality after coronary artery bypass graft surgery. The Working Group Panel on the Cooperative CABG Database Project. Journal of the American College of Cardiology. 1996;28(6):1478-87.

[20] Agency for Healthcare Research and Quality. 2005 National Healthcare Quality Report. Rockville, MD: Department of Health and Human Services, Agency for Healthcare Research and Quality; 2005. <www.qualitytools.ahrq.gov/qualityreport/2005/>

[21] Institute of Medicine, Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: National Academy Press; 2001.

[22] American Medical Association. Principles for pay-for-performance programs; 2005.<www.ama-assn.org/ama1/pub/upload/mm/-1/finalpfpprinciples.pdf>

[23] Rosenthal MB, Frank RG, Li Z, Epstein AM. From concept to practice: early experience with pay-for-performance. Journal of the American Medical Association. October 12, 2005;294(14):1788-93.

[24] Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. Econometrica. 1979; 47(2):263-91.

[25] Camerer CF. Prospect theory in the wild: evidence from the field. California Institute of Technology, Division of the Humanities and Social Sciences, Working Papers: No. 1037; 1998.

[26] Oliver P. Rewards and punishments as selective incentives for collective action: theoretical investigations. American Journal of Sociology. 1980;85(6):1356-1375.

[27] Town R, Wholey DR, Kralewski J, Dowd B. Assessing the influence of incentives on physicians and medical groups. Medical Care Research and Review. 2004;61(3 Suppl):80S-118S; discussion 119S-123S.

[28] Werner R M, Alexander GC, Fagerlin A, Ubel PA. The "hassle factor": what motivates physicians to manipulate reimbursement rules? Archives of Internal Medicine. 2002;162(10):1134-9.

[29] Morreim E H. Gaming the system. Dodging the rules, ruling the dodgers. Archives of Internal Medicine. 1991; 151(3):443-7.

[30] Roland M. Linking 30% of UK GP's pay to quality of care: a unique collaboration between funders, physicians and academics. Presentation to the Annual Meeting of AcademyHealth, Boston, MA, June 27, 2005.

[31] Tieman J. Experimenting with quality. CMS-Premier initiative to reward best, punish worst. Modern Healthcare. 2003; 33(28):6.

[32] Rewarding superior quality care: the Premier Hospital Quality Incentive Demonstration. Centers for Medicare & Medicaid Services Fact Sheet. <www.cms.hhs.gov/HospitalQualityInits/downloads/HospitalPremierFS200602.pdf> Accessed August 25, 2005; updated January 2006.

[33] 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 .September 1987;80(3):344-350.

[34] 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. Preventive Medicine. March 2003;36(3):291-99.

[35] Zaslavsky AM, Hochheimer JN, Schneider EC, et al. Impact of sociodemographic case mix on the HEDIS measures of health plan quality. Medical Care. October 2000;38(10):981-92.

[36] Rosenthal MB, Fernandopulle R, Song HR, Landon B. Paying for quality: providers' incentives for quality improvement. Health Affairs. 2004; 23 (2):127-41.

[37] Berwick DM, Wald DL. Hospital leaders' opinions of the HCFA mortality data. Journal of the American Medical Association. 1990;263(2):247-9.

[38] Gupta N, Kotler PK, Dudley RA. A report card approach to ICUs. Journal of Intensive Care Medicine. 2002; 17(5):211-218

[39] Agency for Healthcare Research and Quality. AHRQ Quality Indicators--Guide to Patient Safety Indicators. Version 2.1, Revision 3. Rockville, MD: Agency for Healthcare Research and Quality; 2003, revised January 17, 2005. AHRQ Pub. No. 03-R203.

[40] Agency for Healthcare Research and Quality. AHRQ Quality Indicators--Guide to Inpatient Quality Indicators: Quality of Care in Hospitals--Volume, Mortality, and Utilization. Revision 4. Rockville, MD: Agency for Healthcare Research and Quality; 2002, revised December 22, 2004. AHRQ Pub. No. .02-RO204

[41] Krumholz HM, Chen J, Wang Y, et al. Comparing AMI mortality among hospitals in patients 65 years of age and older: evaluating methods of risk adjustment. Circulation. 1999;99(23):2986-92.

[42] Tilley BC, Barnes AB, Bergstralh E, et al. A comparison of pregnancy history recall and medical records: Implications for retrospective studies. American Journal of Epidemiology. 1985;121, 269-281.

[43] Quam L, Ellis L, Venus P, et al. Using claims data for epidemiologic research: the concordance of claims-based criteria with the medical record and patient survey for identifying a hypertensive population. Medical Care. 1993;31: 498-507.

[44] Sirio CA, Shepardson LB, Rotondi AJ, et al. Community-wide assessment of intensive care outcomes using a physiologically based prognostic measure: implications for critical care delivery from Cleveland Health Quality Choice. Chest. 1999; 115:793-801.

[45] Burton TM. Operation that rated hospitals was success but the patience died. Wall Street Journal. August 23, 1999:A1.

[46] Clough JD, Engler D, Canuto PE. Cleveland Health Quality Choice was a failure, not a martyr. Quality & Safety in Health Care. 2002 Dec;11(4):391.

[47] Marshall MN, Shekelle PG, Davies HTO, Smith PC. Public reporting on quality in the United States and the United Kingdom. Health Affairs (Millwood). May/June 2003; 22(3):134-148.

[48] Marshall MN, Shekelle PG, Brook RH, Leatherman S. Use of performance data to change physician behavior. Journal of the American Medical Association. September 6, 2000;284(9):1079.

[49] Hibbard JH, Stockard J, Tusler M. Hospital performance reports: impact on quality, market share, and reputation. Health Affairs (Millwood). July-August 2005;24(4):1150-60.

[50] Verdier J, Felt-Lisk S, Smieliauskas F, et al. Quality-Related Provider and Member Incentives in Medicaid Managed Care Organizations. Princeton, NJ: Center for Health Care Strategies; 2004.

[51] Bridges to Excellence. GE, Ford, UPS, P&G, Verizon, others back new pay-for-quality initiative for physicians. [Press release]. April 10, 2003 .<www.bridgestoexcellence.org/bte/news/pr_1.htm>

[52] Nursing home pay-for-performance demonstration draft design. [Slides from special open door forum on the nursing home pay-for-performance demonstration, September 20, 2005]. <www.cms.hhs.gov/DemoProjectsEvalRpts/downloads/NHP4P_Handout.pdf>

[53] Burack JH, Impellizzeri P, Homel P, Cunningham JN Jr. Public reporting of surgical mortality: a survey of New York State cardiothoracic surgeons. Annals of Thoracic Surgery. October 1999;68(4):1195-200.