Concerns about quality of care have accelerated since the 1990s, as studies by Wennberg, Fisher, and others have documented large and unexplained variations in rates of health care utilization and clinical outcomes across geographic areas, calling into question the traditional approach of relying on the medical profession to deliver high-quality care uniformly (Davis & Guterman, 2007; Wennberg et al., 2002). Since about 2000, several landmark publications have highlighted widespread problems with patient safety and quality of care, most notably from the Institute of Medicine (IOM) and the RAND Corporation (IOM Board on Health Care Services, 2001; Kohn et al., 1999; McGlynn et al., 2003). These studies helped to galvanize a policy consensus, leading the federal government and private health insurance plans to increasingly focus policy, regulatory, and management interventions more directly on quality of care measurement, quality improvement programs, and financial incentives for quality improvement through pay for performance (P4P).

P4P programs have focused primarily on quality of care measures to assess provider performance. Although other performance evaluation approaches, such as efficiency measures, are possible for P4P, those in policy circles currently perceive the lack of incentives for improved quality in the prevailing fee-for-service (FFS) payment systems as a major problem in the US health care system. As a result, P4P programs have focused mainly on addressing this problem.

This chapter reviews issues regarding the application of quality measures in P4P programs. The first section of the chapter provides background, including conceptual frameworks for quality of care, and reviews organizations that develop and certify quality measures. The second section discusses different types of quality measures, including structure, process, and outcome measures (Donabedian, 1966). The third section reviews issues in selecting quality measures for P4P programs. The fourth section describes methods for
analyzing quality measures for P4P. The fifth section discusses public reporting of quality measures and how that separate approach to quality improvement can be integrated with P4P programs.

**Background**

Two major conceptual frameworks have been developed for health care quality, one by Donabedian (1966) and the other by the IOM (IOM Board on Health Care Services, 2001). Researchers and policy makers can use both to guide development and implementation of quality measures for P4P. Other models are available (IOM Board on Health Care Services, 2006), but the Donabedian and IOM frameworks are the most widely used. This section describes both models, although we emphasize Donabedian's framework because developers of P4P programs use it more frequently than the IOM model.

Donabedian's model focuses on the concepts of structure, process, and outcome for defining quality of care. Despite being first published more than 40 years ago, this model remains a leading paradigm. The key elements can be described as follows:

**Structure**—the inputs into the health care production process. These include physicians, nurses, and other staff; medical equipment; facilities; information technology; administrative support systems; medical supplies; pharmaceuticals; and other resources. Problems may arise if inputs are not available when needed to treat a patient or when health professionals do not view the capabilities of inputs as optimal. For example, from a structure perspective, high-quality care may entail using clinical teams, including board-certified cardiologists, to treat patients with advanced heart failure rather than relying solely on primary care physicians.

**Process**—the procedures used to diagnose a patient, prescribe a course of testing or treatment, and ensure that the testing and treatment are carried out in accordance with clinical guidelines or norms of medical practice. Process problems are often classified as underuse, misuse, or overuse of tests or treatments (IOM Board on Health Care Services, 2001; Chassin & Galvin, 1998). For example, from a process perspective, high-quality care may be associated with laboratory testing of diabetic patients at least once a year for their levels of glycosylated hemoglobin (HbA1c). To be useful, process measures must have been demonstrated to be statistically and clinically associated with corresponding outcome measures. For example, appropriate
Quality Measures for Pay for Performance

colorectal cancer screening is a process measure known to reduce mortality attributable to colon cancer.

**Outcome**—the ultimate goals of reducing morbidity and mortality and improving quality of life (QOL) and patient satisfaction. Quality analysts can identify problems by comparing outcomes achieved for patients with the outcomes expected for similar patients with the same disease. For example, from an outcome perspective, high-quality care may be associated with a reduced frequency of relapses for patients with multiple sclerosis (MS) because such care reduces the morbidity that patients suffer.

Quality measures focused on structure are easier to measure, but they may have only limited impact on the final outcomes of interest. Process measures assess the actual medical treatment that physicians and other health professionals provide. However, they may require detailed data collection through costly medical record reviews to obtain the clinical data necessary to identify patients for the measures’ denominators and clinical events for their numerators. Administrative data such as insurance claims may enable less costly measurement for some types of process measures, but most cannot be measured in this way. Combinations of administrative and chart review data collection (“hybrid” measures) have been encouraged by the National Committee for Quality Assurance (NCQA) as an efficient approach. Electronic medical records (EMRs) may someday reduce the data collection burden for process measures, but they are still not widely implemented. The technical specifications for process measures may also be costly to develop and keep updated because of changes in medical practices and technologies and development of new pharmaceuticals; such changes may cause shifts in the lists of inclusion and exclusion criteria for denominators and numerators used to calculate performance rates. Outcome measures may be ideal in theory, because they represent the ultimate goals of interest, but they are often difficult to measure, especially in a timely fashion. For example, variation in mortality outcomes may appear only many years after patients have received medical treatment. In addition, many factors can affect variation in mortality and other outcomes besides the quality of medical care, for example, age, comorbidities, diet, exercise, and risky behavior. Consequently, physicians and provider organizations, such as hospitals and physician groups, may consider it inappropriate to hold them accountable for quality of care measured by outcomes unless complex risk adjustments are applied. QOL and other patient-reported outcomes are costly to measure because they require primary data
collection through direct responses from patients in formal surveys. Patient-reported outcomes data cannot be collected from secondary data, such as insurance claims, that do not include patient surveys. Finally, poor outcomes, which are often rare and therefore more difficult to measure, lead to sample size issues.

As a result, not one of these three categories—structure, process, or outcome—is always better than the others for quality measurement, and P4P programs have applied all of them in practice to assess quality. P4P programs use process measures more frequently than the other types of measures because they represent a middle ground, physicians and other clinicians are more familiar with them, and process measures make clear what must be improved in care processes in comparison with outcome measures. However, process measures also have shortcomings and are often complemented by structure measures, outcome measures, or both.

The IOM presented its conceptual model of factors affecting health care quality in its *Crossing the Quality Chasm* report (IOM Board on Health Care Services, 2001). This report focused on six goals for improving health care. As the report noted, health care should be

- **Safe**—avoiding injuring patients with the care that is intended to help them.
- **Effective**—providing services based on scientific knowledge to all who could benefit and refraining from providing services to those not likely to benefit (avoiding underuse, misuse, and overuse).
- **Patient-centered**—providing care that respects and is responsive to individual patient preferences, needs, and values and ensuring that patient values guide all clinical decisions.
- **Timely**—reducing waits and sometimes harmful delays for both those who receive and those who give care.
- **Efficient**—avoiding waste, in particular waste of equipment, supplies, ideas, and energy.
- **Equitable**—providing care that does not vary in quality because of personal characteristics such as gender, ethnicity, geographic location, and socioeconomic status (pp. 39-40).

The Donabedian and IOM models overlap in many areas. The IOM’s goal of safe care relates to all three of Donabedian’s concepts. For example, a structure intervention to implement computerized physician order entry
(CPOE) for drug prescriptions may prevent overdoses of chemotherapy drugs, or dangerous interactions between drugs a patient may be taking. A process intervention could also prevent overdoses by requiring multiple nurses to check dosages before they administer drugs. If these interventions are successful, then the beneficial outcomes are reduced rates of morbidity and mortality for patients taking these drugs.

The IOM's aims for effective, patient-centered, and timely care all relate to Donabedian's concept of process. The aim of patient-centered care also relates to outcomes that are measured using patient surveys of QOL, patient satisfaction, or experience of care.

However, the IOM model includes additional concepts of cost and access in its domains of efficient and equitable care. We prefer to maintain the conceptual distinctions between the overall health policy goals of increasing quality, reducing cost, and improving access, which are often used as a larger conceptual framework for analyzing health services. For example, quality is often associated with measuring the performance of clinicians or provider organizations. Access is often associated with measuring the performance of health care systems that may cover regions or an entire country. Cost is usually considered separately from quality, and discussions of cost tend to focus on the analysis of financial resources and budgets used for providing care.

Researchers and policy makers can use the IOM's aims for efficient and equitable care to develop performance measures for P4P programs that are separate from measures of quality. We view those concepts as useful for P4P programs, although we present efficiency measures in a separate chapter of this book (Chapter 5) to maintain the conceptual distinctions between quality, cost, and access goals. Researchers and policy makers can also develop access measures of performance that are separate from quality measures. The access measures may, for example, be included alongside quality measures in P4P programs that focus on vulnerable populations, such as Medicaid enrollees.

**Types of Quality Measures**

Following our focus on Donabedian's model, we categorize quality measures for P4P programs into structure, process, and outcome. This section describes examples of all three types and illustrates how P4P programs have applied them.
Structure Measures

Health professionals and policy makers sometimes view structure measures as less valuable than process or outcome measures because they are further removed from the ultimate goal of improving outcomes. Structure measures indicate only the potential for providing or improving quality of care; they do not directly measure the clinical processes of care or health outcomes that more closely represent true quality. Also, fewer structure measures are available for ambulatory care than for inpatient care (Birkmeyer et al., 2006), although that situation may change in coming years, with expanded emphasis on implementing systems that support health care delivery, such as EMRs and chronic disease registries.

Some individual health professionals may view structure measures as unfair if the individuals score low on them but have high quality in terms of outcomes (Birkmeyer et al., 2006). Moreover, linkages between structural measures and outcomes may be evident at health system or community levels, but they may not differentiate individual clinicians well.

In recent years, health care accreditation organizations have moved away from their traditional reliance on structure measures to focus more on process measures. For example, the Joint Commission has moved toward using measures of process and outcome (Hurtado et al., 2001). Several quality monitoring organizations have also begun to focus more on process and outcome indicators than they previously did. For example, the NCQA has developed and periodically updated a set of Healthcare Effectiveness Data and Information Set (HEDIS) indicators used to measure quality in private managed health care plans, Medicare, and Medicaid (National Committee for Quality Assurance [NCQA], 2006). Federal quality improvement efforts, including Medicare's Physician Quality Reporting Initiative and the Hospital Quality Initiative, have also focused on process and outcome measures (Centers for Medicare & Medicaid Services [CMS], 2008a, 2008b).

Nonetheless, some structure measures have been found effective in promoting quality. In addition, they are usually easier to measure than process or outcomes, so data collection is both less challenging and less expensive. They may also be efficient in the sense that one structure measure may relate to several different diseases or outcomes (Birkmeyer et al., 2006).

The Leapfrog Group is a proponent of several specific structure measures of quality (Birkmeyer & Dimick, 2004; Leapfrog Group, 2008). Its focus includes three structure measures for hospitals:
• **Computerized physician order entry**—Studies have shown that physicians can significantly reduce prescribing errors when they use CPOE to highlight incorrect dosages, drug interactions, or patients’ allergies to prescribed drugs.

• **Intensive care unit (ICU) physician staffing**—ICUs staffed with critical care specialists (sometimes referred to as intensivists) can reduce the risk of patients’ dying in the ICU.

• **Evidence-based hospital referral**—For patients needing certain types of complex medical procedures, referral based on scientifically evaluated factors, such as the number of times a hospital has performed a procedure each year, has been shown to reduce the risk of death.

All of these measures can potentially be applied in P4P programs, and the Leapfrog Group has provided assistance to health plans and payers using them in P4P. A recent study confirmed the Leapfrog Group’s claims about the value of these structure measures of quality (Jha et al., 2008). It found that hospitals that implemented these three types of patient safety–oriented interventions also had improved process and outcome measures of quality—including lower 30-day mortality rates—for patients with acute myocardial infarction, congestive heart failure, and pneumonia.

Although the Leapfrog Group has received the most attention in policy circles for its focus on structure measures of quality, several payer groups have also used structure measures in their P4P programs. For example, some programs have provided incentives for health care professionals to invest in information technology, implement electronic health records (EHRs), or use EHRs. These programs include the Integrated Healthcare Association, Bridges to Excellence, and the Hawaii Medical Services Association (IOM Board on Health Care Services, 2007; Bridges to Excellence, 2008; Gilmore et al., 2007; McDermott et al., 2006).

Payers and health care plans are particularly interested in the development and use of EHRs or EMRs because they have the potential to improve coordination of care and reduce medical errors. For more than a decade, many commentators have noted this potential, but the high costs of these systems in relation to the benefits received at the physician practice level have hampered implementation. As a result, most small and medium-sized physician practices have been slow to adopt EHRs and EMRs. This means that P4P programs with incentives for EHR or EMR implementation could provide a useful catalyst for improving the business case for these systems at the physician practice level.
Organizational interventions are another type of structure measure. The United Kingdom's (UK's) P4P program applies a substantial number of these interventions to assess performance of family practitioners in the National Health Service (Department of Health, 2004; Doran et al., 2006a, 2006b). The UK P4P system is noteworthy because it has 146 quality measures, the largest number of any P4P program. Of these, 76 are classified as clinical quality indicators, and another 70 are classified as organizational and patient experience quality measures. Our review of these measures reveals that many of them are structure measures of quality. For example, the 76 clinical quality indicators are classified into 11 chronic disease domains. For each domain, the first indicator is a structure measure of whether the practice has a register of the patients with that disease (e.g., “DM 1. The practice can produce a register of all patients with diabetes mellitus”; Department of Health, 2004). Similar indicators are repeated for the other 10 disease domains (e.g., hypertension, chronic obstructive pulmonary disease, and cancer), thus the 76 clinical quality indicators are actually 11 structure measures and 65 clinical process measures.

The UK's 70 organizational and patient experience quality measures include 31 structure measures, according to our review (e.g., “Records 3: The practice has a system for transferring and acting on information about patients seen by other doctors out of hours”; Department of Health, 2004). The wording of this measure uses structure language about the presence of a system, rather than process language about the percentage of patients seen by other doctors who were transferred to new doctors and for whom clinical information was acted on. The other structure measures from among these organizational and patient experience indicators are worded in similar ways, often referring to the presence of a system rather than to how the system should be applied in medical practice.

Therefore, the UK’s P4P program uses 42 structure measures, which is 29 percent of the overall total of 146 quality measures. This is a higher percentage of structure measures than most US P4P programs use, although the difference may reflect the much larger overall number of indicators in the UK program compared with US programs. The UK program uses 104 process measures—in absolute numbers, more process measures than are used in any US P4P programs, even if the percentage of process measures is lower in the UK program than in some US programs.

The current interest in “medical homes” in the United States can be viewed as another type of structure intervention for quality of care. Medical homes
are sometimes proposed for additional per capita or per visit bonus payments to physicians because they are expected to improve coordination of care, case management, information technology, and continuity of primary care. As a result, medical homes could be a focus of implementation incentives in P4P programs as structure measures of quality, similar to P4P programs that provide incentives for implementing EMRs.

**Process Measures**

Process measures are procedures or treatments that are designed to improve health status or prevent future complications or comorbidities. In most cases, a process measure is a dichotomous indicator of whether the process was performed during the recent past (e.g., whether patients taking interferon drugs had liver function tests in the past 6 months). When characterizing health professional and provider organization performance, a process measure is expressed as the proportion of eligible patients who received the procedure. Process measures are often limited to certain subgroups of patients for whom a particular treatment process applies.

A benefit of process measures is that health professionals recognize them as reflecting routine clinical care. In many cases research studies have found them to be associated with outcomes, although this is not always well established and it is becoming increasingly less acceptable to use process measures that lack an evidence base. Process measures may also provide positive spillover effects, such as raising clinicians’ awareness about quality measures and clinical guidelines (Birkmeyer et al., 2006).

Process measures are important to consider because they are usually more practical for data collection and monitoring than outcomes are for quality improvement programs (IOM Board on Health Care Services, 2006; Eddy, 1998; Jencks et al., 2000). Four characteristics of process measures make them more feasible than outcome indicators for routine quality monitoring. First, outcomes often occur with lower frequency than do associated process indicators. For example, breast cancer deaths occur at a rate of only about 1 per 1,000 women older than 50 years of age (an outcome indicator). In contrast, NCQA and Medicare apply process indicators specifying that all women ages 50 to 69 should be receiving biennial mammograms for breast cancer screening (Kautter et al., 2007; NCQA, 2006).

Second, outcomes often require long periods for evaluation of effects (Palmer, 1997). For example, to get outcomes measured as 5- to 10-year cancer survival rates, it will take at least 5 to 10 years and probably longer because of
data reporting lags. Routine evaluation of process indicators can usually be done annually or even more frequently, depending on how many patients with a particular disease physicians treat in any given month or year.

Third, factors outside the control of health plans, health care organizations, or clinicians who treat patients with chronic diseases often affect outcomes. In contrast, process of care measures are, by definition, primarily under health professionals' control and usually do not require risk adjustment.

Fourth, significant improvements in processes are generally larger in relative terms than improvements in outcomes, which makes it easier to measure the former and easier to identify significant changes. This aspect enables P4P programs to base incentive payments on more statistically reliable data.

Process measures have another appealing aspect. One of the key steps in quality improvement is identifying the cause of problems and improving the associated care processes. Unlike outcome measures, process measures target which area of care needs to be improved, although the health care organization still needs to ascertain how to achieve the improvement needed.

Nonetheless, just because process measures are usually easier to specify, measure, and track from year to year does not mean that P4P programs should use them exclusively. An important consideration with process measures is whether they are clearly linked to improved outcomes or at least to a higher likelihood of improved outcomes. Researchers have developed a range of methods to assess the strength of scientific evidence that underlies clinical practice guidelines, quality measures, and quality improvement programs (Lohr, 2004). However, the extent of currently available evidence to support links between process indicators and outcomes varies widely (Birkmeyer et al., 2006). Process measures recommending routine laboratory testing may be good clinical practice, but the results of testing, and the degree of follow-up that health professionals provide, are more closely linked to outcomes than to whether testing was done.

**Outcome Measures**

Ultimately, people care most about outcomes, including morbidity, mortality, QOL, functioning, and patient satisfaction. Improved outcomes are the desired consequences of quality improvement efforts. For example, for treatment of MS, outcomes may be measured through physical and mental functional status indicators (which can be either physician reported or patient reported), disability, complication rates (e.g., urinary tract infections, pressure
ulcers), frequency of relapses, standardized measures of health-related QOL, standardized measures of patient satisfaction, and other indicators.

In addition, implementing a system of outcome measurement may itself improve outcomes—a “Hawthorne effect”—beyond the interventions that may be related to particular outcomes (Birkmeyer et al., 2006). For example, surgical morbidity and mortality in VA hospitals fell dramatically after measurement began in 1991, to an extent too large to explain solely by organizational or process improvements.

To date, P4P programs have used outcome measures less frequently than process measures, even though outcome measures are preferable in theory because they represent the ultimate health care goals. As noted, researchers have raised concerns regarding the strength of the relationship between structure or process quality measures and the outcomes they target. In addition, focusing on outcome measures is expected to encourage innovation in health care services more than would focusing on process measures (Sorbero et al., 2006). Unless process measures are updated frequently, which could be costly, they may reinforce existing care patterns rather than encourage development of new treatment methods that improve outcomes even more.

Thus, moving P4P programs toward more direct use of outcome measures where possible may be beneficial. Physicians and other clinicians may want to maintain a mix of process and outcome measures in P4P programs, however, given that process measures provide more specific information about particular care processes that need to be improved.

One concern is that multiple factors outside of the health care system can affect outcome measures, a problem that is commonly cited. As a result, physicians and other clinicians may not consider it fair to be held accountable for outcome performance. For example, many different physicians and other health professionals may treat patients with cardiovascular disease, and patient factors regarding diet, exercise, and adherence to medications may play a large role in mortality rates. Risk adjustment for outcome measures can be expensive if it is done in detail using data from medical records, and it may be inadequate if done using administrative data that, though usually less expensive, contain less clinical detail (Birkmeyer et al., 2006). However, recent efforts to add present on admission (POA) codes in hospital medical records will enable better analysis of outcomes for hospitals (Jordan et al., 2007; Pine et al., 2007). POA codes help to determine whether complications and comorbidities were acquired by patients during a hospital stay, and thus can be attributed to the care provided at the hospital.
Another problem with measuring outcomes is that sample sizes may be small for surgical outcomes or rare diseases. This means that statistical analysis of performance improvement may be unreliable, so P4P programs cannot pay bonuses with confidence in these situations.

An issue with patient-reported outcomes, such as those reflecting QOL or health status, is that they require patient surveys, which may impose costs that provider groups find difficult to sustain. Lower cost options are not readily available because these types of outcomes require primary data collection from patients.

In general, outcomes can be categorized into two types: clinician-reported outcomes and patient-reported outcomes. Both can be applied in P4P programs.

Clinician-Reported Outcomes

Clinician-reported outcomes are those that physicians or other health care professionals measure and record. They can be further classified as “intermediate” outcomes (e.g., blood pressure levels or HbA1c levels that put patients at risk for severe complications or comorbidities, or stage of cancer at diagnosis) and final outcomes (e.g., decubitus ulcers causing morbidity, or mortality). Medical records are primary data sources for collecting clinician-reported outcomes, but P4P programs can also use laboratory databases and claims data for some types of outcome measures.

**Intermediate outcomes.** Intermediate outcomes measure clinical results, so they can be viewed as outcomes rather than process measures, but they are not final outcomes in the sense of being direct measures of morbidity or mortality. Blood pressure levels are important outcomes that provide information on patients’ risks for heart disease and stroke. As a result, control of blood pressure is a goal that makes sense to reward through P4P programs. Similarly, HbA1c levels can indicate risks for diabetics to develop several severe complications, including retinopathy, nephropathy, and neuropathy.

A positive feature of intermediate outcomes is that they are closer than process measures to the final clinical outcomes of interest, so they provide a closer link to final outcomes. For example, HEDIS process measures include measuring blood pressure periodically for patients with heart disease or hypertension, and testing for HbA1c levels periodically for diabetics. However, just because the tests were conducted does not mean that the clinical indicators of interest were brought under control. Thus, focusing on the levels
themselves—the intermediate outcomes—is preferable to targeting only the frequency of testing.

Another positive feature of intermediate outcomes is that they can be measured more frequently than final outcomes. As a result, P4P programs that focus on providing routine performance assessments and periodic (often annual) bonus payments to physicians and other health professionals can use them more easily. For example, among patients with diabetes, neuropathy can result in foot or leg infections that require amputations, but these events occur much less frequently than elevated levels of HbA1c. Amputation rates can be tracked as performance measures, but they may require much larger samples of diabetic patients than are available for most physician practices or even larger group practices. Because amputations occur less frequently than HbA1c tests, they may not provide the routine data needed for annual performance assessments for P4P bonus payment determinations.

A third positive feature of intermediate outcomes is that they may not require risk adjustment for appropriate performance assessment, in contrast to mortality and other final outcomes. Appropriate levels of blood pressure and HbA1c are standardized for most patients, and although patient factors enter into the levels achieved, physicians and other clinicians can usually be held accountable for average levels achieved over groups of patients. Physicians may not be able to control patients’ diet and exercise patterns completely, but most accept responsibility for working with their chronic disease patients to control blood pressure and HbA1c, especially when patients are at risk for complications associated with elevated blood pressure or HbA1c. Risk adjustment may still be indicated for some types of intermediate outcomes, but it may be implemented more easily than for final outcomes, with fewer variables and data collection requirements.

As a result, a promising approach for P4P programs would be to work more aggressively to expand the range of intermediate outcomes that they use to assess provider performance. They can also increase the weighting provided to these measures relative to others. Intermediate outcomes represent a middle ground between the more controllable process indicators, which may not be closely linked to final outcomes, and the final outcomes of interest, which would be ideal performance measures—if they were easier to measure frequently and if it were easier for providers to link the final outcomes to their efforts.
In addition, Current Procedural Terminology (CPT-II) codes have now been developed for some intermediate outcomes, such as HbA1c levels for diabetics, so these outcomes can now be measured using administrative claims data instead of relying solely on more expensive chart review (American Medical Association, 2008). More work is needed to expand the list of CPT codes for intermediate outcomes, and to expand the extent to which health professionals use them for billing for clinical services, but the technical groundwork has been laid in the CPT coding system.

**Final outcomes.** Clinician-reported final outcomes can include a range of morbidity, functional status measures, and mortality measures. Morbidity measures include medical and surgical complications that can be used in P4P programs, although they apply primarily to hospitals or other institutional providers. Decubitus ulcers are an example of a preventable complication that can develop during hospital stays or among nursing home residents. Because they are preventable for most patients, they can serve as a useful outcome measure for P4P programs. They usually occur infrequently, however, so they may need to be measured as average rates over large groups of patients.

Other types of hospital-related complications, such as postsurgical infections, readmission rates within 1 to 3 months of discharge, and “never” events, such as surgery on the wrong body part, can also serve as final outcome measures. These outcome measures include patient safety quality indicators that the US Agency for Healthcare Research and Quality (AHRQ) and others have developed (AHRQ, 2003). An advantage of these complication-related outcome measures is that they do not require risk adjustment in most cases, because patient safety indicators such as avoiding postsurgical infections apply to most patients. In addition, these indicators are clearly under the control of hospitals and their medical staff because they occur during the patient’s stay in the hospital, nursing home, or other medical facility. As a result, clinicians are more willing to accept responsibility for these types of final outcome measures. For example, when Medicare recently announced that it would not reimburse hospitals for admissions that resulted in “never” events, there was little resistance from the hospital or physician community.

Functional outcomes comprise measures of activities of daily living (ADLs), instrumental activities of daily living (IADLs), time to walk 25 feet, established scales such as the Functional Independence Measure (FIM), and others. Health professionals often use such outcomes in rehabilitation services assessments, to judge patients’ progress in recovery from illness, or to assess levels of disability. These measures have promise for P4P because they can be measured
frequently and can show significant changes resulting from effective treatment in many situations. Assessments of MS patients routinely use clinician-reported outcome measures of physical and cognitive function, including the Expanded Disability Status Scale, Multiple Sclerosis Functional Composite, neuropsychological tests, and others (Cohen & Rudick, 2007; Coulthard-Morris, 2000; Joy & Johnston, 2001; Rothwell et al., 1997).

However, functional outcomes suffer from at least two concerns. First, they may require risk adjustment like other types of outcomes, because factors unrelated to the quality of medical care can affect them. Second, many functional outcomes rely to some extent on the clinician’s judgment for scoring each patient on the measures or scales. This can make the functional outcomes more vulnerable to gaming by health professionals and providers, especially when P4P programs use scores to calculate bonus payments.

Mortality is the ultimate final outcome, although mortality measures can be a sensitive topic for both patients and clinicians. In principle, P4P programs could use mortality rates or risk-adjusted mortality rates for performance assessment. Aligning health professionals’ financial interests in keeping the patient alive as long as possible may improve mortality outcomes. However, patients and their families may understandably be concerned if the presence of a P4P program implied that physicians would not be doing all they could to keep patients alive in the absence of P4P financial incentives.

At the same time, researchers have conducted much statistical analysis in recent years to create risk-adjusted mortality rates for several diseases and populations. Quality improvement efforts and public reporting of mortality outcomes have used these rates. For example, risk-adjusted mortality rates have been reported publicly for several years on Medicare’s Dialysis Facility Compare Web site (Trisolini & Isenberg, 2007), the State of New York has reported publicly on risk-adjusted mortality rates for cardiac surgeons for many years (Jha & Epstein, 2006), and Medicare recently began reporting risk-adjusted mortality rates for some types of patients on its Hospital Compare Web site (CMS, 2009). These measures have been well tested, so they presumably could be extended for use in P4P programs.

Patient-Reported Outcomes

Patient-reported outcomes have the advantage of providing data on outcomes that can be collected only from patients; broadly speaking, such outcomes can include QOL, patient satisfaction, and patient experience of care. Patient satisfaction data are already used in P4P programs, including those sponsored
by the Integrated Health Care Association, the Hawaii Medical Services Association, and the British National Health Service (Doran et al., 2006b; Gilmore et al., 2007; McDermott et al., 2006). Health professionals might be expected to object to P4P programs tying financial rewards to subjective indicators such as patient satisfaction, but the success of these three large P4P programs in implementing these patient-reported outcome measures indicates that clinician acceptance is possible.

Standardized patient satisfaction scales for quality measurement and public reporting have become widely accepted in recent years, which has helped to promote their use in P4P programs. AHRQ developed the Consumer Assessment of Healthcare Providers and Systems in the 1990s, and it now includes a family of standardized patient surveys that have broad acceptance for assessment of health plans, hospitals, physician groups, and other provider organizations (AHRQ, 2007, 2009).

Clinical trials of new drugs and evaluations of health service interventions have used QOL scales, such as the SF-36 or SF-12, to monitor outcomes of care, and these types of scales have potential for use in P4P programs. Medicare has also publicly reported QOL scales in recent years through the Health Outcomes Survey (NCQA, 2006). These scales can include broad, generic measures of functioning, such as the Physical Component Summary (PCS) and Mental Component Summary (MCS) scales for the SF-36 or SF-12. They can also include more specific measures of particular symptoms, such as the Modified Fatigue Impact Scale (MFIS; National Multiple Sclerosis Society, 1997a). With a wide range of both general health and symptom-specific QOL scales developed in recent years, P4P programs have many options if they wish to measure QOL performance.

Some QOL scales have been developed for particular diseases, such as the Kidney Disease Quality of Life Scale (Hays et al., 1994). The National Multiple Sclerosis Society developed a multipurpose patient survey instrument for measuring a range of outcomes, the Multiple Sclerosis Quality of Life Inventory (MSQLI; National Multiple Sclerosis Society, 1997a, 1997b). The MSQLI includes the 21-item MFIS and nine other scales that measure outcomes related to generic physical and mental health, pain, sexual satisfaction, bladder control, bowel control, visual impairment, perceived deficits (cognition), and social support. Several other disease-specific QOL measures for MS have also been developed in recent years (Burks & Johnson, 2000; Nortvedt & Riise, 2003).
Patient-reported outcomes can also include other types of functioning scales; some overlap with clinician-reported outcomes such as those assessing ADLs, IADLs, and mobility for rehabilitation programs. For example, for MS two disability scales focus mainly on walking ability: the Extended Disability Status Scale (EDSS), which neurologists assess, and the Patient-Determined Disease Steps (PDDS), which patients can assess. In theory, P4P programs could use either or both measures to assess performance, although in MS, the goal is usually slowing the decline in function rather than improving function.

It is interesting that P4P programs have used patient satisfaction scales to date but not QOL scales. This may stem from health professional and provider organization concerns that factors outside their control can affect QOL and thus would require risk adjustment. For example, one study that used QOL scales to assess the performance of Medicare providers used several demographic and comorbidity variables for risk adjustment (Trisolini et al., 2005). In contrast, patient satisfaction is more under the control of physicians and other health professionals and providers because it largely reflects the patient’s experience of receiving care from the clinician. Moreover, private health insurance plans may include patient satisfaction in P4P programs because it helps them attract enrollees into their plans and thus affects their ability to compete against other health insurance companies.

**Issues in Selecting Quality Measures for P4P Programs**

**Data Sources and Administrative Burdens**

The three basic data sources for measuring quality of care indicators are medical records (paper-based or electronic), patient surveys, and administrative data (including enrollment records, insurance claims, and facility records). Each has advantages and disadvantages (Berlowitz et al., 1997).

**Medical records.** Medical records have the advantage of including much more detailed clinical information than do administrative data: for example, the specific clinical values provided by laboratory test results for HbA1c for diabetics, assessments of the patient’s severity of illness, physical examinations, pharmaceutical prescriptions, neurological tests that physicians conducted, results of magnetic resonance imaging (MRI) tests and other radiology examinations, and clinicians’ or providers’ notes about treatments and the patient’s status. They also provide more complete information than do claims data on diagnoses, complications, and comorbidities because claims rely on
coding that information, and coding efforts may be incomplete for some types of diagnoses and complications.

As a result, medical record abstracts are important data sources for quality measures when process interventions or outcomes depend on identifying patients with a particular clinical or functional status that cannot be identified through claims data or patient surveys. For example, appropriate interventions and expected outcomes will vary between MS patients depending on whether they have relapsing-remitting or progressive forms of the disease (Noseworthy et al., 2000).

The main disadvantage of medical records is the high cost of collecting those data in many circumstances, particularly when the records are paper-based or when EMRs do not include the specific data necessary for quality measures. The manual medical record abstraction process necessary in such circumstances can be very labor intensive; usually a trained nurse must ensure accuracy, and medical record coders and administrators may also be involved. However, large sample sizes may become increasingly available in EMRs as implementation of EMRs spreads, at least for larger physician groups and integrated delivery systems. In theory, EMRs could reduce the cost of data collection substantially, by enabling access to data already stored in digital format, like claims data. However, at present EMRs are available only in a limited number of hospitals and physician groups, and smaller physician practices have had even lower implementation rates. Comprehensive availability of EMRs for all health professionals and provider organizations across the country remains a long-term goal that may take many years to achieve despite the initiative in the American Recovery and Reinvestment Act of 2009 to fund implementation of EMRs.

Another weakness of medical records, and even EMRs, is that a given patient's medical data can be fragmented across the multiple medical records maintained by the different physician practices, hospitals, and other providers treating the patient. Efforts to develop community-wide health information exchanges (HIEs), to enable more comprehensive access to a patient's data, are still in the pilot phase. EMR vendors are working to make their systems compatible with one another to better promote development of HIEs, but this effort, too, remains in the development phase.

**Patient surveys.** Surveys can provide unique types of data for measuring quality indicators. For example, some types of outcomes, such as patient satisfaction, can be measured only through patient surveys. Surveys can be used to collect data on physical functioning, mental functioning, and social
support for a range of diseases. Disease-specific symptoms, such as fatigue, urinary dysfunction, bowel dysfunction, and sleep satisfaction, can also be captured in survey scales. Standardized QOL survey instruments often capture both generic and disease-specific outcomes data (e.g., National Multiple Sclerosis Society, 1997a, 1997b) by including a mix of scales. Researchers and policy makers can analyze those patient-reported quality measures independently or in conjunction with physician-reported measures of complementary outcomes that may be included in a patient’s medical record.

Patient surveys have two main disadvantages, however. First, they can be costly, depending on how they are administered, whether by trained interviewers (in person or by phone) or not. Mail surveys may be a relatively low-cost option in many cases, but they often suffer from lower response rates and higher rates of missing data. Conversely, when trained interviewers conduct in-person interviews, the costs of administering the survey are higher, but the data may be more complete. Many studies have struck a middle ground, using telephone surveys, which can be conducted by interviewers or with computer assistance.

In recent years, online surveys have become more common, and they may enable less expensive survey data collection to become more widespread in the future. At present, the more limited availability of Internet access for low-income respondents and the more limited willingness of elderly or chronically ill patients to participate in online surveys pose problems. However, these concerns will likely diminish considerably in the future as online access and Web use become more routine for most Americans. Online surveys also have the advantage of enabling automated skip patterns and immediate prompts to respondents for out-of-range values and missing data.

The second disadvantage of patient surveys is reliance on patient recall. For infrequent events (such as use of some types of health services) or long recall periods, this drawback may result in inaccurate data. Where possible, combining patient surveys with administrative data can avoid this problem, such as by using surveys for QOL outcomes that require patient responses and administrative claims data for hospital days and other utilization or cost outcomes.

Researchers must also guard against unexpected variations in patient responses due to cultural, racial, ethnic, language, educational, or socioeconomic differences among respondents. Survey instruments often require translations into multiple languages, and researchers may conduct
cognitive testing, reading level testing, and other types of pretesting with
different patient groups prior to widespread implementation of surveys.

**Administrative data.** In quality measurement, analysts commonly apply
two types of administrative data: enrollment records and insurance (billing)
claims data. Quality measurement also sometimes uses clinical data systems—
laboratory and pharmacy—although they more closely relate to EMRs while
often containing some administrative data.

Administrative data have the advantage of being a low-cost data source:
they are already stored in digital format for other purposes, so they are less
difficult to access and analyze. Researchers often apply administrative data
to identify denominator inclusions and exclusions for quality measures. For
example, quality measures for treatment of diabetics are often limited to
patients between the ages of 18 and 75 (NCQA, 2006), for whom age data
and ICD-9 diagnosis code data used to identify the denominator population
are often accessed through administrative data. A lack of detailed clinical
information, however, such as the results of laboratory tests, is a common
weakness of administrative data; in addition, diagnosis code data often need to
be screened or validated to ensure accuracy.

Enrollment data are useful for the basic demographic information needed
for both process and outcome indicators, such as age, gender, insurance
coverage, and death dates. These data are usually included in databases with
one record per patient; generally, they are easy to use for data analysis, but only
rarely do they provide all of the information needed for quality measures.

Claims data are useful for some types of process measures, in situations in
which the claims data are reasonably complete and provide sufficiently detailed
clinical information. Two good examples are indicators for pharmaceutical
utilization (e.g., whether MS immunomodulatory disease-modifying drugs
have been in continuous use) and laboratory test utilization (e.g., whether
patients taking interferons receive liver function tests and complete blood
counts with platelet counts every 6 months). However, in a recent study on
MS quality indicators (Trisolini et al., 2007), we found claims data to be
limited in their applicability for MS quality measurement in many ways, for
they did not have sufficiently detailed or consistent data on some types of
important diagnoses (e.g., urinary tract infections), important treatments
(e.g., intravenous corticosteroids), or episodes of illness (e.g., MS relapses). In
addition, claims data do not contain any information on a patient's course of
MS (i.e., relapsing-remitting, secondary progressive, primary progressive, or
Claims data do have several advantages. First, they are reasonably complete for the data they collect, because they are used primarily for billing purposes; health professionals and providers thus have a direct financial incentive to ensure that all bills are submitted for reimbursement. Second, they usually include data on all of the clinicians and provider organizations treating a patient and thus avoid one of the weaknesses of medical records data: patient records that may be fragmented across the different health professionals and facilities providing treatment. Third, they enable analysis of quality measures using large sample sizes, including up to thousands of patients at a time. The large numbers of enrollees that many private health insurance plans cover, and even larger numbers that public payers such as Medicare and Medicaid cover, make this possible.

Risk adjustment of quality measures can also use claims data, because they include variables such as age, gender, diagnoses, and others that risk adjustment models often apply. For example, the risk adjustment model applied for hospital mortality measures in Medicare's Hospital Compare Web site (CMS, 2009) uses claims data. With the advent of Medicare's requirement for POA coding of comorbid conditions, the potential for more accurate coding in claims data has increased considerably.

In sum, all three data sources have advantages and disadvantages for quality measurement. Efforts to measure quality indicators for P4P programs should consider all three options before selecting the most suitable source—or sources—for each indicator. A comprehensive set of quality indicators can include contributions from all three sources. The choice for each P4P program may depend on a range of factors, including budget constraints, preferences for the types of quality measures to be collected, and the need for patient surveys, if the program desires data on QOL or patient satisfaction outcomes.

**Number of Quality Measures**
P4P programs have included widely varying numbers of quality measures. The United Kingdom’s program includes 146 quality measures, far more than any of the P4P programs in the United States have used. In contrast, the Medicare Physician Group Practice Demonstration includes 32 quality measures, which were phased in over several years (Kautter et al., 2007). Private-sector P4P programs typically include fewer measures than those in the public sector.
How many quality measures to include in a P4P program depends on several considerations. Using a larger number of measures poses three risks: (1) increasing the administrative burden on both P4P program administration and on participating health professionals and provider organizations; (2) making the results more complex and cumbersome for health professionals and provider organization staff to interpret; and (3) requiring use of measures less closely linked to health outcomes or less well studied. Data verification and audit costs may increase greatly as the number of measures increases, although sampling providers or measures (or both) to be audited can reduce this burden. Physicians often express concern about the dozens of clinical guidelines and quality incentives they face, at the same time that they perceive themselves to be under pressure to see more patients and complete more paperwork. Under the circumstances, some quality measures may be ignored—especially in situations where P4P incentives for individual quality measures may affect only a small percentage of physician income or provider organization revenue. An advantage of including a larger number of quality measures is a more comprehensive evaluation of the care provided.

The pros and cons of including smaller numbers of quality measures in a P4P program are generally the converse of those for larger numbers of measures. The positives of fewer quality measures include less administrative burden, lower overall program costs, and easier interpretation of results. The negatives include the danger of focusing provider attention on a subset of the important clinical areas and the risk of financial incentives’ being focused on just a few measures. The financial incentives could motivate clinician or provider behavior that focuses too much on the clinical conditions included in the P4P program. Studies have found that high performers in some clinical areas are not necessarily high performers in other clinical areas (Sorbero et al., 2006). Although professional ethics and peer review may blunt the impact, inappropriate financial incentives nonetheless remain a risk. Given that hundreds of thousands of physicians practice in the United States, it is likely that some percentage will succumb to financial temptation. Even if this portion represents only 1 percent of all physicians, it would mean that thousands of physicians could be involved in such dubious financial and clinical practices.

Another potential advantage of including fewer structure or process measures is that researchers can focus on measures more closely linked with outcomes. For example, the Leapfrog Group initially focused on just three structure measures that had clear links to outcomes. Similarly, process
measures could focus more on immunization or pharmaceutical indicators that have more evidence for impacts on outcomes than on other measures that may be less closely associated with outcomes or that have less evidence to support the relationship.

Intermediate outcomes, such as blood pressure and HbA1c levels, could be substituted for process measures to provide closer links to final outcomes. For example, in its total of 146 quality measures, the UK P4P system included many structure measures, most of which have not been rigorously studied for impact on outcomes. The UK system will be a valuable test of a P4P program with a larger number of quality measures, but there is a need for a closer examination of the structure measures it used to reach its high number of quality measures.

Another issue is whether to weight all of the quality measures equally in calculating provider performance scores. Equal weighting makes it easier for health professionals, provider organizations, and policy makers to interpret the results but may not reflect the underlying value of the different measures or the underlying level of evidence supporting different measures. For example, HEDIS includes quality measures for treatment of diabetics that focus on both the frequency of HbA1c testing and the levels found in that testing. The quality measure focusing on the level of HbA1c could be weighted more heavily in calculating provider performance because it is more closely related to patient outcomes than is the frequency with which the HbA1c tests were conducted.

**Types of Quality Measures to Include**

The IOM's (2006) report on performance measurement criticized the focus of most current quality measures on specific types of health professionals, provider organizations, or settings of care, such as only on physicians, medical offices, or hospitals. That report recommended expanding quality measurement to include three other types of quality measures:

- **Composite measures**—documenting whether a patient has received all recommended services for a particular condition (and perhaps for multiple conditions). Composite measures of process and intermediate outcomes may show greater room for improvement than individual measures and may be more closely related to final outcomes than single measures are. In calculating composite measures, analysts can apply weighting schemes to give higher weight to quality measures identified
as more closely related to either final outcomes or cost savings. All-or-nothing measures may require success on each of a set of measures to be considered at the same time.

- **Population-based measures**—aggregating results for a given region or with breakdowns by population subgroup on socioeconomic status, race, or ethnicity to test for the presence of disparities. These aggregations can be done on several levels, such as groups of clinicians and provider organizations, delivery systems, a community, or a geographic region.

- **Systems-level measures**—analyzing performance across diseases, conditions, clinical specialties, or departments. Researchers and policy makers can define systems to include a continuum of care across ambulatory, inpatient, and long-term care services within a given community.

These alternate approaches to quality measurement have the potential to broaden the focus of current P4P programs, moving beyond the current emphasis on individual clinician, clinician group, or hospital accountability. For example, using these alternate types of quality measures could promote more shared accountability for quality performance across multiple health professionals and provider organizations, a goal that the IOM (2006) report highlighted for development of a national system of performance measurement. This approach may include rewarding the complete set of clinicians and providers included in the care of a patient, or participating in a system of care in a community. Such a step does open up the risk of “free riders,” however, in that some clinicians or provider organizations may not be fully motivated to improve quality, preferring to benefit from improvements in performance measures that result from the efforts of the other professionals or provider organizations being assessed with them. However, this broader approach is consistent with management literature that emphasizes the value of applying group incentives in addition to individual incentives (Packwood, 2008). P4P systems could also pursue such a strategy, for example, basing some bonus payments on physician group or provider organization incentives and some on incentives at the level of the community, region, or health care system. In that way, providers could earn bonus payments based on both their own work and their contributions as part of a broader community of professionals and provider organizations that are treating patients in a given region or system of care. This approach mitigates the risk of free riders by tying some
incentives directly to provider performance but also preserves some incentives for broader regional or systemwide performance results.

**Room for Improvement in Performance**
Another consideration for selecting quality measures is the degree to which there is room for improvement in performance on the measure. Ideally, P4P systems would select measures that have large opportunities for improvement, both because this represents good public health practice and because it enables health professionals and provider organizations to demonstrate improvement in quality. Conversely, if there is little room for improvement in a quality measure, where providers have already achieved high performance scores, then payers have less motivation to reward improvements, and providers have fewer opportunities to demonstrate improvement. For example, in recent years the NCQA removed one of its hospital quality measures—beta blocker treatment after myocardial infarction—because hospitals had improved their performance to a high average level, leaving little room for additional improvement.

**Cost Containment**
P4P programs have focused mainly on quality improvement, but both public and private payers have major concerns about cost containment as well. Quality improvement advocates have claimed that improving quality may in some situations also reduce costs, but evidence for that dual benefit is limited. In theory, better care for diabetics can reduce complications such as retinopathy, nephropathy, and neuropathy, thereby reducing or avoiding the future costs of treating those complications. However, many other factors can affect the actual levels of costs incurred by diabetic patients, such as age, comorbidities, and low-income status. Moreover, for most payers, the time horizon required to reap cost savings for reduced complications of diabetics is too long, meaning that they lack strong incentives to implement programs that address such complications.

Several types of quality improvements are fairly closely linked to cost savings, however. First, patient safety measures that improve quality by reducing adverse drug events, hospital-acquired infections, or surgical errors will directly affect costs by reducing hospital admissions, lengths of stay, or readmissions. Disease management programs that target heart failure patients for more intensive ambulatory care, case management, and nurse-
led home care can also reduce hospital admissions and result in cost savings (Anderson et al., 2005). Several chronic diseases known as ambulatory care sensitive conditions (ACSCs) provide opportunities for cost savings through reduced hospital admissions, and quality measures based on ACSCs have been published as Prevention Quality Indicators (PQIs; AHRQ, 2001). The essential idea of ACSCs is that through improved primary health care and preventive care, achieved by enhancing quality or access (or both), chronic disease patients will be less prone to complications or exacerbations of their illnesses that will result in hospitalizations. Given that hospital admissions are very high-cost events in health care, ACSCs have the benefit of linking quality improvement more directly to cost savings than many other types of quality measures, which may take many years to realize their cost impacts.

Although both public and private payers have goals to improve quality of care as an end in itself, both may sometimes opt to target quality measures for P4P programs that also have demonstrated cost savings. For private payers, such a strategy can help reduce the premiums they charge business customers, thus providing a competitive advantage. For public payers, limited governmental budget resources may lead to a dual focus on measures that can simultaneously promote both cost savings and quality improvement.

A related issue is how to fund the bonus payments to providers in P4P programs. Some programs require bonus payments to be funded by cost savings demonstrated by the participating health professionals or provider organization. This is the approach that Medicare’s Physician Group Practice Demonstration took (Kautter et al., 2007). Other P4P programs, such as that of the Integrated Healthcare Association, have provided “new money” for P4P bonus payments.

**Methods for Analyzing Quality Measures for P4P Programs**

**Risk Adjustment**

Ensuring fair performance assessments when using outcome measures often necessitates risk adjustment or stratification of performance results by population subgroups. For example, many factors apart from the quality of medical treatment affect outcome measures such as patient mortality (most notably, the patient’s age and the number and severity of diseases). As a result, when analysts or policy makers use mortality as a quality measure, comparing health professionals and provider organizations on raw mortality statistics can be misleading. At worst, those types of comparisons might encourage
clinicians and provider organizations to avoid treating older or sicker patients who most need their care, because such patients would adversely affect mortality performance measurements.

For example, the New York State cardiac surgery mortality report cards are based on data that are risk adjusted to better ensure fair performance assessment of surgeons (Jha & Epstein, 2006). Similarly, Medicare’s Dialysis Facility Compare Web site provides public reporting of mortality data associated with kidney dialysis facilities only after risk adjustment using a broad range of variables (Trisolini & Isenberg, 2007). At the same time, existing statistical models used for risk adjustment do not fully explain the range of factors affecting mortality outcomes. That is why researchers still prefer randomization of patients in evaluating outcomes from new pharmaceuticals in clinical trials (Palmer, 1995). Randomization controls for unmeasured and unknown factors affecting outcomes, whereas statistical models used for risk adjustment can only apply factors that can be measured. As a result, payers and policy makers have not yet been comfortable with moving from public reporting of risk-adjusted mortality outcomes to including mortality outcomes in P4P programs. Public reporting can include caveats, but bonus payments in P4P programs must be based on specific quantitative results, which leaves less opportunity to include qualifying statements regarding interpretation of the results.

Researchers and policy makers sometimes propose risk adjustment for process measures of quality (although in practice they are less often risk-adjusted). One rationale is that patient adherence to prescribed tests and pharmaceutical treatments may be lower for patients in lower socioeconomic groups or different racial or ethnic minority groups than in other populations. As a result, some health professionals and provider organizations argue that process measures such as HbA1c testing for diabetics or blood pressure levels should be risk-adjusted to account for patient factors affecting adherence. For example, Zaslavsky and Epstein (2005) found that racial, income, and education variables affected some HEDIS quality measure scores for health plans significantly, although the rates for most plans changed by fewer than 5 percentage points. Similarly, Mehta et al. (2008) found that patient characteristics (including age, body mass index, race, and type of insurance) and hospital characteristics significantly, but modestly, affected hospital process measures for treatment of acute myocardial infarction.
Nonetheless, a countervailing concern is that one could interpret risk adjustment for these types of factors as endorsing lower-quality care for low-income or minority patients. One method proposed to mitigate this concern is to stratify quality results for public reporting by patient-level factors, including insurance status, low income, and minority status. For example, NCQA requires that HEDIS quality measures for health plans be presented separately for different types of health insurance, including commercial insurance, Medicaid, and Medicare (Zaslavsky & Epstein, 2005). This approach could be extended to include other sociodemographic variables where sample sizes permit. P4P programs, however, may still face challenges of variable incentives for health professionals and provider organizations if the perception remains that avoiding treatment of certain population subgroups could improve performance scores and increase bonus payments. This problem could be mitigated if payers could provide higher P4P bonus payments for quality performance in treating patients in population subgroups known to be associated in the aggregate with worse outcomes or lower adherence to prescribed treatments.

Another approach that some quality measurement efforts use is for clinicians to document the prescription or recommendation for testing and to use that as the measure of quality, thus removing the effect of patient adherence from quality measurement. In most cases P4P program analysts will need medical record data for this measurement, because administrative claims data do not yet routinely capture this type of information. That drawback may change, however, if the new codes for the CPT-II system become more widely adopted; they allow coding for “patient reasons” (including refusal or nonadherence) why a given patient may not have undergone a particular test (American Medical Association, 2008). This new type of CPT coding reduces physicians’ incentive to avoid the more difficult patients who may adversely affect their measured quality performance. A potential risk is that physicians will overuse these codes for patient exclusions, and thereby game the performance assessment calculations to increase their bonus payments. Auditing patient records to verify the exclusions is one approach for mitigating this risk.

**Identifying High-Quality Providers**
P4P programs can take several different approaches to identifying high-quality health professionals and provider organizations that qualify for P4P bonus payments through meeting quality goals or targets. Three basic methods are
(1) threshold targets, (2) improvement-over-time targets, and (3) comparison with other providers.

**Threshold targets.** The most common method for identifying high-quality clinicians and provider organizations in P4P programs, the threshold approach mainly offers simplicity and ease of understanding for clinicians. For example, “For patients with diabetes, 75 percent will have an HbA1c test at least once per year.” The target is clear from the outset so practitioners and provider organizations know what specific number to aim for. A disadvantage is that providers starting at lower levels of quality may perceive thresholds as unattainable if the thresholds are set very high. Another disadvantage is the lack of incentives for further quality improvement above the threshold.

Results from the P4P program in the United Kingdom provide some evidence to support this latter point: an evaluation study found that initial gains in quality in the first 2 years of that program were significant, but gains slowed markedly in the third and fourth years when there were few additional financial incentives for further improvement (Campbell et al., 2009). These results must be interpreted cautiously, however, because there was no comparison group available for this P4P program, given that all UK family practitioners were included in it. As a result, the evaluators had to rely on an interrupted time-series analysis in their study design. Nonetheless, the results are consistent with the concern about lack of further incentives once threshold targets are achieved by providers in a P4P program that relies on that type of target.

One way to mitigate concerns of initially low-performing providers is to establish a series of thresholds, with successive incentives for higher levels of performance in a “stair step” model. For example, threshold-based P4P bonus payments could start at 40 percent performance (where 100 percent is perfect performance, with all denominator patients receiving the indicated numerator interventions) and increase with every 5 percentage points achieved, up to 80 percent. In this way, providers may be able to achieve the first two or three levels of incentives even if they cannot achieve all nine possible levels. They can then aim to achieve higher levels of incentives in future years of the P4P program as they are able to further improve quality performance. In this way, the threshold approach can motivate providers at lower levels of initial performance because they can earn some performance payments in even the first year of a P4P program.
P4P programs can apply several methods for setting specific performance levels for the threshold targets. For example, programs can use (1) consensus goals that P4P payers and participating health professionals and provider organizations have set through joint discussion and agreement, (2) levels set by payers to promote a “reasonable” degree of quality improvement, (3) target levels benchmarked to levels that other high-quality clinicians and provider organizations already achieve, and (4) comparison with other quality measurement programs to find targets these programs may have set for similar populations or similar quality measures.

**Improvement-over-time targets.** Improvement-over-time targets establish a baseline from a provider’s own prior performance level and then evaluate current period performance starting from that level. For example, “Providers should achieve at least a 5 percent increase in performance from the prior year.” An advantage of this approach is that providers starting from low levels of initial performance can view these targets as attainable. However, this approach has two disadvantages. First, providers at high levels of prior performance may find additional improvement difficult to achieve. For example, if a provider is already at 90 percent performance or above on a particular quality measure, then a 5 percent improvement may be difficult. Second, payers may object to rewarding providers at low levels of performance even if they are achieving improvements from even lower performance in the prior year. For example, if a provider improves from 10 percent to 15 percent from one year to the next, that 50 percent improvement may still represent a much lower absolute level of performance than that of all other providers in the P4P program.

P4P programs can set improvement-over-time targets in several ways. They can use percentage improvements (e.g., 5 percent), percentage-point improvements (e.g., 5 percentage points), or reductions in performance gaps (e.g., 10 percent reduction in the gap between 100 percent performance and the prior year’s performance level).

The Medicare Physician Group Practice Demonstration uses the reduction-in-performance-gaps approach (Kautter et al., 2007), which has the advantage of requiring larger percentage improvements at lower levels of initial performance and smaller improvements at higher levels of initial performance. For example, if the initial performance is 40 percent, then the gap from the perfect score of 100 percent is 60 percent, and the 10 percent improvement target represents a 6 percentage point improvement. As a result,
the target would be 46 percent performance in the year being assessed. In contrast, if the initial performance is 80 percent, then the gap is 20 percent and the target is just 2 percentage points’ improvement, or 82 percent. In this way the reduction-in-performance-gaps approach mitigates one disadvantage of improvement-over-time targets, by requiring more improvement from low performers and less improvement from high performers.

Another way to mitigate the disadvantages of both the threshold and improvement-over-time targets is to adopt a combined approach that includes both types of targets in one P4P program. The Medicare Physician Group Practice Demonstration adopted such an approach, which included both threshold targets and improvement-over-time targets (Kautter et al., 2007). Physician group practices participating in the demonstration can meet any of the targets to earn performance bonus payments. In this way, the program established positive incentives for physician group practices at both high and low initial levels of initial and ongoing performance.

**Comparison with other providers.** The third approach to identifying high-quality performance is to compare providers with one another. In this method, P4P programs consider only those who perform better than their peers to be high quality and deserving of P4P bonus payments (irrespective of their absolute levels of performance). For example, P4P programs could award incentive payments to the top 20 percent of providers. The Premier Hospital Quality Incentive Demonstration, which compared more than 200 hospitals using a range of different quality measures, used this approach (Lindenauer et al., 2007).

This approach contrasts with both the threshold and improvement-over-time approaches, in which P4P programs allow all providers the possibility of earning incentive payments. The comparison approach focuses on rewarding only the highest performers from among those participating in the P4P program.

The comparison approach has at least two disadvantages. First, even low absolute levels of performance may earn rewards, as long as any given provider’s performance is higher than that of the others. Second, providers do not know in advance what their goal is, because it depends on their peers’ performance levels. Some may consider themselves unlikely to perform in the highest 20 percent, and they may therefore lose their motivation to improve (at least by this incentive alone).
Another option with the comparison approach is to include penalties for low performers at the same time as providing rewards for high performers. This option may provide an additional (negative) incentive for those who do not think they have the potential to reach the top 20 percent.

**Statistical Analysis of Quality Improvement**

Statistical confidence in P4P results can be problematic when individual physicians or small physician practices are the units of accountability. In many of these situations, only small numbers of patients may be available for the denominator populations for some types of quality measures in any given practice; as a result, random statistical fluctuations may account for observed performance on quality measures. Minimum sample sizes per quality measure may need to be as high as 411 patients, a figure HEDIS used to indicate a sample size sufficient to provide confidence that the detectable difference in performance is 10 percentage points (NCQA, 2006).

Achieving sample sizes of 411 or more for diabetics, for example, may require a focus on large physician groups, hospitals, integrated delivery systems, combinations of smaller physician practices into networks or virtual groups, or a geographic area such as a city or county that contains a higher number of providers. It may be easier to achieve sufficient sample sizes for population-based quality measures that do not focus on patients with particular diseases such as diabetes. For example, quality measures for influenza vaccinations include all people ages 50 or older in the denominator population.

Analysts and policy makers sometimes consider smaller sample sizes acceptable if quality measurement can include the entire population of patients in a physician practice, rather than a sample, so that the observed number of patients can be considered the true number and not subject to random statistical fluctuation. However, a countervailing argument is that the observed patient population and quality performance levels may vary randomly over time, so, from that perspective, the population of patients a physician practice treats in any one year is still a sample of the patients treated over multiple years. From that perspective, application of statistical analysis and calculation of confidence intervals is still needed, and the intervals may be very wide when only small sample sizes are available.
Public Reporting of Quality Measures
Researchers and policy makers generally view public reporting of quality performance as a distinct approach to promoting quality improvement, separate from P4P programs. For example, the Medicare Web sites Hospital Compare, Nursing Home Compare, and Dialysis Facility Compare all provide online data that consumers and medical professionals can view to check on the quality-of-care performance of those types of provider facilities in all regions of the United States. Similarly, for many years the NCQA has provided comparative quality performance data on managed care organizations through its HEDIS program. These efforts and others aim solely to enable public reporting of quality-of-care performance data, unrelated to any direct financial incentives that P4P programs would include. Public reporting can provide indirect financial incentives, however, by potentially motivating patients to “vote with their feet” and thus increase utilization and revenue for higher quality providers. In most cases this is only a potential effect, however, and evidence of its impact on patient behavior is limited.

Despite the conceptual distinction between P4P and public reporting initiatives, several P4P efforts have integrated public reporting into their programs. Most notably, the IHA established public reporting of the quality performance results used in its P4P program as one of its program’s guiding principles to promote public transparency of P4P incentives.

Many P4P programs adopt a different strategy, releasing quality performance data only to participating clinicians and provider organizations—and not to the public. This is consistent with the methods that many continuous quality improvement (CQI) programs use, giving feedback of data only to the providers that the program is assessing; the aims are to preserve confidentiality of performance results and to promote providers’ willingness to participate in CQI initiatives. By avoiding public reporting, staff of these CQI programs argue that they are increasing provider participation, decreasing the risk of “defensive medicine,” such as avoiding sicker or more difficult patients, and forestalling efforts by providers to game the data collection efforts.

In general, it seems appropriate to limit public reporting to situations in which practitioners or provider organizations have developed a good level of experience with a set of P4P quality indicators and the methods for identifying performance targets. Especially in the early stages of P4P programs, many clinicians may be concerned about the fairness of quality measurement methods and performance assessments. They may prefer that public reporting
of results wait until the performance measurement system has been better tested, and better established through several years of measurement cycles, so that confidence in the accuracy and appropriateness of the quality data has become well established.

P4P programs might also consider several middle ground approaches that entail more limited public reporting. For example, public reporting could focus on aggregated results by region or for groups of providers rather than individual clinicians or provider organizations. In this way the public could view the P4P program's overall results, but those of individual clinicians and provider organizations would still remain confidential. In addition, P4P programs could present individual physician or physician group results while masking the names of the physicians or physician groups with code numbers to prevent performance results from being associated directly with them. Such middle ground public reporting efforts could facilitate some degree of public transparency while mitigating clinicians' and provider organizations' concerns.

References


