Physician Clinical Performance Assessment: Prospects and Barriers

Bruce E. Landon, MD, MBA
Sharon-Lise T. Normand, PhD
David Blumenthal, MD, MPP
Jennifer Daley, MD

A number of constituencies are becoming increasingly interested in measuring the performance of physicians in their day-to-day clinical practices, especially since the Institute of Medicine's report suggested that the quality of care may often be less than optimal. Purchasers of health care services, for instance, are concerned about the effects of suboptimal care on workforce productivity and seek to maximize the quality of care provided. Consumers of care also want to be able to identify high-quality physicians and institutions but lack the effective means to do so. Although some groups have measured and reported quality of care for individual medical groups and physicians, these efforts have been limited. Both patients and health care purchasers desire more effective means of identifying excellent clinicians, and a number of organizations have begun discussing and implementing plans for assessing the performance of individual clinicians and the settings in which they choose to practice.

Physician clinical performance assessment (PCPA) can be defined as the quantitative assessment of physician performance based on the rates at which their patients experience certain outcomes of care and/or the rates at which physicians adhere to evidence-based processes of care during their actual practice of medicine. Theoretically, such assessments could serve a variety of purposes. First, PCPA could be useful for ensuring the competency of individual physicians, whether through credentialing, board certification, licensure, or other means. Physicians labeled as competent would have met specific quality-of-care standards that were defined for their specialty. Such a system would be useful for identifying high-quality physicians and also for protecting the public from physicians whose performance falls below accepted standards. Second, public release of PCPA information could be used to support health care choices by consumers (eg, patients or purchasers) or to reward individual physicians for excellent quality of care. Third, PCPA could be useful for quality improvement purposes on the part of physicians and/or the organizations in which they work. Although the requirements for effective PCPA would vary with these purposes, it should be most stringent for purposes of competency assessment to minimize the risks due to misclassification.

In 1998, the American Board of Medical Specialties—the professional association of 24 physician-certifying boards in the United States—created a task force on physician competence that concentrated its efforts on developing recommendations and methods to evaluate physician-specialists on an ongoing basis following their initial cert

For editorial comment see p 1210.

©2003 American Medical Association. All rights reserved.
Physician Evaluation: \[\text{continued}\]

served.” 4 In addition to PCPA, physicians who had been in practice for longer periods of time who might not be keeping abreast of the latest advances, and for whom the supervised setting of training and the comprehensive board examinations were in the more distant past. These concerns led the American Board of Medical Specialties to require lifelong maintenance of certification. Such certification may include demonstration of ongoing professional status, lifelong learning, cognitive expertise, and, in particular, evaluation of professional practice.

Working with the Council of Medical Specialty Societies—the entity to which 20 professional societies belong—the American Board of Medical Specialties established a joint planning committee on physician competence and a task force to develop policies and report on implementation of ongoing maintenance of certification. While progress has been made in most aspects of maintenance of certification, PCPA has been difficult to implement and has been met with resistance from the physician community.

To develop this program, the task force requested us to review the current state of PCPA. In this article, we focus on the usefulness of PCPA for the critical purpose of competency assessment. We describe desirable recommendations for implementing PCPA for this purpose and identify limitations in current techniques of performance assessment.

**Current Methods of Physician Evaluation**

Epstein and Hundert recently defined professional competence as “the habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values, and reflection in daily practice for the benefit of the individual and community being served.” 4 In addition to PCPA, physician competence can potentially be assessed through a variety of means. These methods include licensure and board certification, which are commonly required for practicing physicians, as well as less commonly used methods including additional cognitive testing, or assessments using standardized patients, peer review, or clinical vignettes. 5. 6

Currently, methods of competency assessment do not identify physicians with either inadequate or exceptional levels of competence because licensure and board certification are typically the only methods used. Licensure, for example, does not necessarily ensure minimal levels of competency. Most states require only the completion of requisite training by accredited educational institutions and passing of national examinations that test minimal general knowledge and that are not specialty specific. Thus, to become licensed, interns and orthopedists take the same general examination. Although there are some data from Canada that licensure examination scores are correlated with future performance, 7. 8 these studies have only examined specific areas of quality over a relatively short period of time after certification and, as one-time assessments, are not necessarily indicative of how well physicians keep up with advances in their specialties. In addition, these studies evaluated examinations that were designed for family practice only, and thus may not be generalizable to other specialties.

Similarly, board certification does not ensure exceptional competency in a specialty or subspecialty because the American Board of Medical Specialties now certifies most specialists. 9 Furthermore, board certification is generally a 1-time or infrequent (eg, every 10 years) occurrence, usually based on an assessment of knowledge and/or basic procedural skills.

Alternative methods for assessing competence such as standardized cases, peer review, and direct observation are rarely used and have been criticized because of their expense and lack of objectivity. While potentially useful, these methods have not been well studied and are subject to a number of limitations. 5 In addition, these approaches differ from PCPA in important ways. For instance, while clinical vignettes potentially can be used to measure adherence to processes of care, 9 they may not reflect accurately what physicians do in their day-to-day clinical practices. Furthermore, although direct observation, standardized patients, and peer review measure the activities of daily practice, these methods generally are not quantitative nor can they easily measure outcomes. Incorporating clinical performance assessment could potentially address some of these limitations, particularly if used in combination with other methods of assessment.

Common methods of competency assessment or accreditation for health care institutions such as hospitals or health plans are limited by analogous problems. For instance, Chen et al 10 found that while accreditation by the Joint Commission for the Accreditation of Healthcare Organizations was associated with quality of care for acute myocardial infarction, there was considerable variability within accreditation categories. In a similar analysis, they also found that their ability to discriminate between any 2 hospitals according to their ratings by a publicly released report card was also limited. 10, 11 Similarly, Schneider et al 12 found that Medicare beneficiaries’ assessments of quality using the Consumer Assessment of Health Plans survey instrument were minimally related to processes of care as measured by the National Committee for Quality Assurance’s (NCQA’s) Health Plan Employer Data and Information Set. Both of these measures are used by NCQA to accredit health plans.

**Current Uses of PCPA**

Currently, PCPA is used only rarely for competency assessment. Although some health plans and physician organizations profile physician use patterns to measure costs and efficiency, we believe that quality is conceptually distinct from efficiency and did not include these efforts in our review. Similarly, efforts of state licensing boards typically concern ethical behavior and substance abuse. The Joint Commission for the Accreditation of Healthcare...
Organizations and the NCQA require that hospitals and health plans review physicians’ practice patterns and quality indicator performance at the time of credentialing and recredentialing, but physicians are rarely if ever sanctioned and this information is not available publicly. In addition, it is not clear if the data that are reviewed are methodologically sound or meaningful.

Some professional specialty societies have begun encouraging physicians to measure their performance by offering increased recognition to those who participate in voluntary performance assessment. For example, the American Medical Association’s Consortium for Quality Improvement has developed and disseminated performance measurement sets for a number of conditions to promote the collection of uniform performance measures and improve the quality of care. The NCQA, in conjunction with the American Diabetes Association, has also developed a clinician recognition program that allows physicians providing care to diabetic patients to submit performance reports on how well they deliver different aspects of care appropriate for diabetic patients. The NCQA maintains a list of “recognized” clinicians on its Web site enabling patients to identify high-performing physicians in their state. In addition, some specialty societies have started programs that allow physicians to submit data voluntarily. These types of voluntary efforts, however, have met with inconsistent success, and because of selection effects cannot be used to systematically evaluate physician competency.

A number of firms are also compiling and releasing information on physician performance publicly on the Internet. These efforts, however, often post incomplete and sometimes inaccurate assessments of physicians and typically rely on readily accessible administrative data that are not necessarily reflective of physicians’ practices.

Finally, some states, including New York and Pennsylvania, sponsor PCPA programs that have been collecting and disseminating information about hospital and physician performance in cardiac surgery and other cardiology programs. Massachusetts has begun a similar data collection effort but it is not planning to publicly report on physician performance. While promising, these mandatory efforts are currently limited in scope, involve only 2 specialties (cardiac surgery and cardiology) and a single procedure for each specialty (either coronary artery bypass graft surgery or angioplasty).

PROPOSED STANDARDS FOR A CLINICAL PERFORMANCE MEASUREMENT SYSTEM FOR COMPETENCY ASSESSMENT

The use of performance measurement for competency assessment could have a considerable impact on practicing physicians. Sanctions for poor performance might include loss of board certification, suspension or loss of hospital privileges, decedentialing by health plans, or in the most extreme case, loss of licensure. Given the potential consequences for physicians, PCPA for the purposes of competency assessment necessitates a stringent set of requirements to be fair and equitable. We list a number of desirable qualities in the BOX. In addition to being evidence-based and valid, measures to assess physicians must also be feasible to collect, applicable to a large enough population of patients to allow assessment of individual physicians, attributable to the performance of an individual physician, and adjustable for patient level characteristics that might confound the measure. When used for credentialing, the collection of measures applicable to each particular specialty in medicine should be reasonably reflective of the scope of work performed by physicians in the field.

OBSTACLES TO IMPLEMENTING PHYSICIAN PERFORMANCE MEASUREMENT SYSTEMS

Lack of Evidence-Based Measures for Many Specialties

Ideally, for each medical specialty, evidence-based measures would exist either for outcomes of care or for clinical processes that have been linked to improved outcomes for patients and that are representative of the most important clinical activities of that specialty. These measures would serve as the basis of an objective, evidence-based performance assessment system. In fact, few medical specialties have an evidence base that is robust and comprehensive enough to support PCPA. Some specialties such as cardiology and endocrinology have some evidence-based process measures that have been definitively linked to improved patient outcomes. Other specialties, such as cardiology, have outcomes that have been studied, such as mortality in coronary artery bypass grafting. Outcome measures for other specialties, however, occur too infrequently or too long after care to make their collection feasible. Consequently, most specialties lack a comprehensive set of evidence-based measures.

Challenges in Defining Thresholds for Acceptable Care

A fundamental question for competency determination relates to the choice of standards or thresholds that delimit acceptable performance. This is an issue that is not unique to measuring physician performance but it takes on added importance in this case because of the potential problems that might result from misclassifying physicians. For instance, as noted above, physicians who are not deemed competent might be excluded from the networks of health plans or denied privileges at hospitals. For some performance measures, an absolute threshold will exist. With process-based measures, a natural optimal standard is 100%; for example, the goal is that all ideal candidates for β-blockers after a myocardial infarction receive the therapy. However, in the majority of performance areas, an absolute threshold may not exist, or even if such a bound does exist, it likely needs to be adjusted to account for other reasons for less than perfect performance such as idiosyncratic patient reactions, lack of insurance, or patient
Box. Characteristics of Ideal Performance Measures

Evidence-Based
The measure must be evidence-based and noncontroversial in the broad clinical community. Broad professional consensus is acceptable. Solid evidence consists ideally of US Preventive Task Force Level I evidence (from multiple randomized clinical trials) or overwhelming physician consensus.

Agreed-on Standards for Satisfactory Performance
The measure must have established standards for satisfactory performance.

Standardized Specifications
Ability to collect in a standardized and reliable way across multiple physicians and sites of care.

Adequate Sample Size for Reliable Estimates of Individual Physician Performance
The measure should be applicable to a group of patients of sufficient size to provide a reliable estimate of an individual physician's performance in caring for patients with that condition.

Adjustment for Confounding Patient Factors
The measure must account for patient factors that may confound individual physician performance assessment. Factors include differences among patients prior to medical diagnosis and treatment (ie, case-mix, severity of illness, comorbidity). Factors may also include sociodemographic characteristics that influence patient adherence to treatments.

Care Attributable to the Individual Physician
The measure must be attributable to the care rendered by the individual physician being assessed. Correct identification of the physician responsible for the care. Evidence that the process and/or outcome can be attributed to the individual physician as opposed to the system of care in which he or she practices.

Feasible to Collect
The information must be obtained from existing data sources in a reliable way. No potential for gaming of the system. Must be affordable to collect.

Representative of the Activities of the Specialty
The collection of measures applicable to a specific specialty should be reflective of the overall clinical practice of that specialty.

Sample Size and Other Statistical Considerations

In many instances, individual physicians do not have adequate numbers of patients with particular conditions to quantitatively assess their performance with respect to those clinical problems. Adequate sample size depends on baseline rates and sources of variation other than the direct actions of the physician and will thus vary from measure to measure. For instance, using process measures related to diabetes, Hofer et al30 estimated that a physician would need to have 100 patients with diabetes to achieve 80% reliability for most diabetes-related measures of quality. Most physicians in that particular study, however, had no more than 60 patients with diabetes.30 Although the NCQA diabetes program requires a minimum of 35 patients for each physician, we believe that this recommendation does not account for the likelihood that observations of patients within a single physician’s practice are not statistically independent. Lack of independence among observations increases the sample size needed to attain a particular level of precision in statistical tests.23,31,32 Not accounting for this statistical dependence has the effect of overstating the certainty with which physicians’ actual level of performance is measured. This lack of statistical independence increases the chance of wrongly classifying physicians either as adequate or inadequate in their performance.31,32

For process measures in particular, inclusion and exclusion criteria can often substantially reduce potential sample sizes, thus compounding problems related to the “tyranny of small numbers.”31,33 For instance, the proportion of patients considered ideal for a group of quality measures such as the use of reperfusion therapy or β-blockers ranged from 9% to 20% of the population with a myocardial infarction.34 These issues are more problematic for primary care physicians who treat patients with a wide variety of conditions. In addition, it is often not reasonable to combine different sets of choice. Finally, some formal mechanism must then be specified to determine how much deviation from this attainable standard is acceptable. For example, a physician may not achieve an acceptable standard but he or she may be in a satisfactory range of performance (eg, >80% adherence to a process measure but <100%).

One approach is to define outliers as the bottom or top percentiles of the distribution, so that a certain proportion of individuals will always fall beneath the cutoff simply by virtue of random variation.23,29 For competency assessment, however, it might be inappropriate to define some arbitrary statistical standard. This would be particularly true in cases in which performance across a set of physicians is tightly clustered, so that absolute difference between the bottom and top of the distribution is small. The lack of widely accepted, rational, and sensible thresholds for minimal performance for the many measures that may be used in PCPA constitutes another barrier to its widespread implementation at this time.
measures because different measures might not be substitutable for one another. Thus, it must be determined if sets of measures are capturing the same underlying property of the care provided prior to combining sets of measures. The sample size problem is less difficult in certain specialties, like cardiac surgery and interventional cardiology, in which physicians may perform a large number of a limited range of procedures. The proportion of all physicians for whom sample sizes are large enough to permit valid PCPA is unknown at this time.

Confounding in Process and Outcome Measurement

Another key challenge in assessing physician performance relates to confounding; because patients are not assigned at random to physicians, there are likely to be systematic differences across physicians in their patient populations. For instance, because of the nature of their practices, some physicians might tend to have better educated patients or healthier patients while others might tend to have more patients with inadequate insurance or poor health. Comparisons of performance across physicians must therefore consider adjusting for patient characteristics to avoid bias.

There are 2 ways to deal with this problem—risk adjustment and restriction. While there are many different methods currently available for risk adjustment of outcomes, these systems are imperfect and often require the collection of detailed clinical information that is uniquely specified for that particular condition. Such data collection is costly on a wide-scale basis and difficult to validate. Risk adjustment systems that rely on administrative data, while reasonable for research purposes and quality improvement, may not be precise enough to determine physician competence.

The second general approach for dealing with confounding usually applies to process measures (eg, receipt of a β-blocker after a myocardial infarction) and involves restricting the analysis to patients who are expected to benefit from a therapy or process of care. In contrast to the outcomes-based approach, this process-based approach implicitly makes risk adjustments by restricting the patient sample of each physician to only those patients who are known to benefit, eg, ideal candidates. This approach is costly because it also requires chart review to determine whether particular patients are eligible for the care in question. Another problem is that chart review may not reveal patient characteristics that influence adherence, although data suggest that these effects are small. Examples of typical process and outcome measures for several specialties are presented in the Table.

Representativeness

A fundamental limitation of PCPA for competency assessment is that the evidence concerning optimal processes usually exists for only a small fraction of the routine work of most specialties and subspecialties. For instance, while there are good measures for preventive services within primary care, these measures do not capture the full spectrum of work done by primary care physicians.

Thus, at the current time, physician performance measurement systems would need to use a limited set of potentially unrepresentative measures as indicators of a physician’s overall performance. Little evidence exists, however, that performance on a specific quality indicator is related to performance on other indicators for other conditions. Palmer et al found almost no association between clinician performance on different guidelines, such as those applying to the workup of an abnormal hematocrit or blood glucose level. The average correlation between a physician’s performances on these 2 guidelines was less than 0.02.

Correlations among Health Plan Employer Data and Information Set measures were typically no association between clinician performance on these 2 guidelines was less than 0.02. The average correlation between a physician’s performances on these 2 guidelines was less than 0.02. Correlations among Health Plan Employer Data and Information Set measures were typically

| Table. Assessing Physician Performance Using Outcomes- and Process-Based Measures |
|-----------------|-----------------|-----------------|
| Type of Physician | Outcomes-Based Measure | Process-Based Measure |
| Cardiac surgeon | 30-d Risk-adjusted mortality rate for all CABG patients treated in a given year | Annual rate of use of percutaneous β-blocker therapy |
| Patient | 30-d Death after CABG surgery | Received percutaneous β-blocker |
| Baseline rate, % | 2.5 | 60 |
| Fraction eligible, % | 100 | Unknown |
| Cardiologist | 9-mo Risk-adjusted restenosis rate for all patients receiving stenting within a given year | Hospital discharge prescription rates for indicated β-blockers for patients hospitalized annually |
| Patient | Restenosis within 9 mo after coronary artery stenting | Hospital discharge prescription for β-blockers in ideal candidates |
| Baseline rate, % | Approximately 10 | Approximately 90* |
| Fraction eligible, % | 100 | Approximately 80† |
| Obstetrician | In-hospital maternal death rate for all pregnancies in a given year | Annual rate of timely check-ups after delivery |
| Patient | In-hospital maternal death | Underwent a check-up 21-56 d after delivery |
| Baseline rate, % | 0.01 | 77* |
| Fraction eligible, % | 100 | 100 |
| Primary care | Risk-adjusted rate of annual hypertension patients achieving hypertension control | Annual mammogram rate among eligible women aged >50 y |
| Patient | Hypertension in control for hypertensive patients | Underwent mammogram if >50 y and eligible in the past year |
| Baseline rate, % | 55.4* | 76.6* |
| Fraction eligible, % | 100 | 100 |

Abbreviations: CABG, coronary artery bypass graft; NA, not applicable.
†Approximately 20% for ideal candidates.
sures, even at the health plan level, are also poor. Therefore, until a much broader set of measures is developed, PCPA could misrepresent the competence of physicians in the performance of the bulk of their work.

**Feasibility and Costs**

Measures applicable to any particular physician (or specialty) must be feasible to collect in an efficient, reliable, and reproducible manner. Generally, measures that rely on existing data sources such as administrative data or that can be collected with minimal additional effort within the flow of work will have the highest chance of effective implementation, but are not good indicators of actual performance. The costs of establishing performance measurement systems should also be considered. Large-scale, hospital-level quality initiatives that include clinical data for case-mix adjustment cost a minimum of $14 and $18 per discharge, which is a seemingly affordable sum. However, similar data collection in the outpatient setting would be substantially more expensive because of the multiple different locations and lack of funding mechanism to pay for this type of performance assessment activities. The increasing availability of electronic information, including claims data as well as laboratory results, will likely mitigate some of these expenses in the future.

Measures of functional status or longer-term outcomes, which would require additional data collection efforts, illustrate some of the problems that might raise the cost and limit the feasibility of collecting data necessary for PCPA. Several attempts have been made to measure these types of functional outcomes in a nonresearch setting and they have generally been met with limited success. The difficulty of collecting data on functional outcomes is likely to affect a number of specialties (eg, rheumatology, ophthalmology, and orthopedics) that have the primary purpose of improving patients’ functional status. A valid system would need to have a reliable method for collecting this information from patients at prespecified time intervals without losing a substantial portion of the patients to follow-up.

**EFFECTS ON QUALITY IMPROVEMENT AND OTHER CONSIDERATIONS**

One implication of quality improvement theory is that the use of PCPA data for other purposes may be detrimental to improving quality by engendering fear and suspicion among physicians. There is often a tension between performance measurement for the purposes of quality improvement and performance measurement to assess competence. The result of the latter could be to encourage gaming of the data or outright refusal to provide them, thus frustrating the goal of identifying and improving failing processes. These issues, however, may be resolvable in the right circumstances. Notable systems, such as the New York State Cardiac Surgery reporting system that was mentioned previously, have been able to achieve both goals simultaneously through rules requiring mandatory participation and careful auditing of data.

Additionally, clinical performance assessment efforts that only measure specific areas may impact other unmeasured aspects of care such as a physician’s time for developing personal relationships with patients. A common fear is that because resources are limited, physicians will concentrate their efforts on areas being assessed, to the detriment of other aspects of care. Farris recommended including sustained measurement of other valid quality indicators not targeted for improvement, and heeding physicians’ concerns about the effects of quality improvement efforts on overall care. A second strategy is to rotate through a set of quality measures. Although researchers almost universally caution that there may be unintended effects, little empirical information about the extent of the problem exists.

Finally, there are aspects of medicine that always will be difficult to measure adequately. Some physicians develop reputations as superb diagnosticians while others become expert at managing patients with multiple complicated issues or are superb communicators. Efforts to measure these aspects of medicine should also be pursued.

**RECOMMENDATIONS**

At the current time, given the state of technology and the existing infrastructure to support performance assessment, broad-based mandatory clinical performance assessment for individual physicians as a means of determining the competence of individual physicians, whether for board certification or other reasons, appears to be infeasible. The majority of specialties lack a sufficiently comprehensive set of measures that would be representative of the work done within the specialty. Still other specialties would need to rely on functional status or other patient-derived information that has rarely been collected in a valid representative way outside of purely research settings.

Nonetheless, assessing the competence of physicians remains a worthwhile goal that can be approached through careful, incremental steps. First, increased research and development are vital to identify valid, reliable, feasible, and acceptable approaches to physician performance assessment. These efforts should also focus on improving the information infrastructure to support such activities. Wider adoption of electronic medical records and the creation of data standards that would permit aggregation of data across multiple sites of care should also be promoted. In addition, research and development in other methods that can be used for competency assessment, including vignettes, peer review, and standardized patients also should continue.

Efforts to promote data collection and assessments for the purposes of quality improvement continue to be essential. Measurement systems for quality improvement, which need not meet all
of the criteria outlined above, would serve the dual purposes of providing data for quality improvement purposes while also stimulating development of an improved infrastructure that could be used for additional types of physician performance assessment.

Finally, competency assessment does not necessarily need to be a universal requirement for board certification or licensure. Some specialties, by the nature of their clinical content, will be more able to develop and implement measurement systems, which could potentially be implemented on a specialty-by-specialty basis as this technology matures. For instance, well-developed systems in cardiac surgery could serve as a basis for such a system and could be implemented relatively rapidly. In addition, measuring and reporting on the performance of groups of physicians may not face the limitations listed above (sample size in particular) and thus may be more feasible to implement in the short term. Patients and purchasers seek assurances of the competency of individual physicians in the day-to-day practice of medicine. Meeting these needs, however, will require considerable improvement in the technology of physician performance measurement.

Author Contributions: Study concept and design: Landon, Normand, Blumenthal, Daley. Acquisition of data: Landon. Analysis and interpretation of data: Landon, Blumenthal, Daley. Drafting of the manuscript: Landon, Normand, Blumenthal, Daley. Critical revision of the manuscript for important intellectual content: Normand, Blumenthal, Daley. Statistical expertise: Normand. Obtained funding: Blumenthal. Administrative, technical, or material support: Normand. Study supervision: Normand, Blumenthal, Daley.

Funding/Support: This work was supported in part by the United Health Care Foundation.

REFERENCES


