Delfini Group ™, LLC

Evidence- & Value-based Solutions For Health Care
Clinical Improvement Consults, Content, Seminars, Training & Tools

White Paper: Performance Measurement

Evidence-based Performance Measurement: Validity Issues & Avoiding Important Pitfalls

Short Version

Delfini Group ™, LLC

Michael Stuart, MD
President

Sheri Strite, Principal & Managing Partner

www.delfini.org

Our Mission –

- **To assist medical leaders, clinicians and other health care professionals by ~**
- Bringing science into medical practice in an **easy-to-understand** way.
- Using **simplified methods** to help navigate the complexities of such areas as evidence-based medicine and other topics.
- Building **competencies** and **confidence** in improving medical care through our well received consultations, educational programs and tools.
- Providing inspiration to others to **improve** medical care and help bring about needed change.

Delfini Group White Paper: Evidence-based Performance Measurement: Validity Issues and Avoiding Important Pitfalls — Short Version

Abstract

Background: Evidence-based performance measurement is an important component of quality improvement efforts in health care. Health care organizations are currently experiencing increasing external pressure to provide objective measures of performance to regulators, payers and patients who are interested in comparing the performance of individual physicians and health care organizations. This article provides a simple and powerful evidence-based approach for evaluating the strengths and weaknesses of performance measures.

Conclusions: Key issues with performance measures are the three validity considerations of a measure. Approaching performance measurement with these issues in mind has not been sufficiently addressed in the medical literature. The approach presented here provides clear and useful direction for organizations attempting to successfully integrate performance measurement into an evidence and value-based approach for improving health care. It also points out important problems or pitfalls frequently encountered when designing and using performance measures to evaluate quality in health care and competence in health care systems, specific care units and individual health care professionals.

Delfini Group White Paper: Evidence-based Performance Measurement: Validity Issues and Avoiding Important Pitfalls — Short Version

Introduction

Providing appropriate health care to their enrollees or patients is a major goal of health care organizations, and most organizations devote substantial resources to continuously improve their health care processes and patient outcomes. Yet, studies have pointed out that, compared to recommendations for care based on the best available evidence, at least 20 to 50 percent of all prescriptions, visits, procedures and hospitalizations in the United States are considered inappropriate as a result of overuse, underuse and mis-use of what has been demonstrated to be effective and beneficial care.[1,2,3] Performance measurement—an important component of quality improvement work– has received increasing attention in part due to the belief that performance measures are an accurate way of evaluating quality and encouraging appropriate care coupled with the belief that utilizing performance measures is an efficient way to achieve these goals. If performance measures can help us improve quality, they are well worth understanding – but it is imperative to understand not only their potential to provide effective solutions, but also to understand their potential to severely threaten quality if not utilized correctly. This paper provides a simple, yet powerful approach—based on three key validity considerations – for evaluating, designing and using evidence-based performance measures. It also emphasizes important considerations for avoiding the common pitfalls in quality comparisons, accountability and pay-for-performance.

Performance measurement is a quantitative way to measure quality. For purposes of this paper, we consider health care "quality" as encompassing effectiveness and efficiency as well as including other value considerations such as cost and utilization and the myriad of potentially important triangulation issues unique to each health care improvement endeavor such as legal considerations, public relations and community standards. Performance measurement can be defined as "whether or how often a process or outcome of care occurs".[4] Health care organizations expend great efforts to evaluate their processes and outcomes, in part, because of the increasing need to provide objective documentation of performance to regulators, payers and patients. Performance measurement can be of value in increasing attention to quality improvement efforts; however, it has the potential to mislead when designed poorly, applied incorrectly or when used to draw conclusions about the relative quality of individual clinicians, groups of physicians, units of care or health care organizations.

Defining Quality Improvement Through Performance Measurement

A performance measure consists of a numerator, a denominator and a frequency – and it is often expressed as a percentage or a rate. The text statement for a diabetic patient care performance measure might be, "The percent of patients with a diagnosis of diabetes mellitus receiving at least one hemoglobin A1c annually."

- To measure quality, the denominator specifies the "universe" of who or what ought to have had an occurrence (e.g., who should be treated with an ARB).
- The numerator is the count of what actually happened (e.g.., who actually got an ARB out of those who should have received an ARB).
- The frequency specifies how often it is supposed to happen.

Good performance measures must be quantifiable, valid and useful, and they need to fit neatly within the larger context of clinical quality improvement efforts. In determining what constitutes improvement, an organization might decide that quality improvement has been accomplished through reaching a –

- **Trend:** Are the results going in the desired direction?
- **Statistically Significant Change:** Have there been meaningful and sizable gains as demonstrated through measuring significant change?
- **Target:** Was a specific goal or range reached that has been defined as "improvement?"

Delfini Group White Paper: Evidence-based Performance Measurement: Validity Issues and Avoiding Important Pitfalls — Short Version

Most organizations set targets as goals for improvement by considering the performance of top-performing organizations, national or local averages and recommendations of groups such as Healthy People 2010[5]. Before setting targets it is important to be certain that the performance measure is valid, is associated with improved quality and that the data can be gathered appropriately (Table 1). These three validity considerations form the basis for evaluating, developing and implementing performance measures.

Table 1. Key Considerations in Assessing Validity of Performance Measure Components and the Data Gathering Process

Denominator Validity	A performance measure's denominator is the pool of eligibility, or the base number of units, from which measurements (or counts) are taken – the "pool." This pool or base may be determined by specific outcomes, characteristics or conditions. Usually this is a population such as adult patients with diabetes mellitus, or it can be a base of units such as the total number of adult patients receiving an appointment at an outpatient clinic, or the total number of patient charts in a specific clinic, for example. A valid denominator means that it specifies the right population or the right base number of units from which the count will be made. The denominator must have appropriate inclusion and exclusion criteria for the pool of who or what is eligible for measurement. For example, a denominator measuring clinical improvement for cervical pap smears which does not exclude women without a cervix would be invalid.
Numerator Validity	The numerator is a count taken from the denominator and measures the number of occurrences of the event of interest – the "what." Numerators generally measure events such as something that happens to a patient, something patients receive or something that is done to them – typically this is an outcome, an intervention, a service or a process. Examples are the number of patients with diabetes mellitus from the denominator (the eligible "pool") who have received an annual hemoglobin A1c, or the percent of charts available for patient appointments at an outpatient clinic. Numerators should be based on valid, useful and usable scientific evidence. That the numerator is a valid one means that it can actually measure improvement. For health status outcomes and interventions, this requires that we know what interventions are likely to result in improvement. Almost invariably this proof requires valid randomized, controlled trials (RCTs) — not database information or observational studies which can be misleading. Databases and clinical records are useful for measuring processes, but are not reliable for attempting to "prove" that a health status improvement was the result of an intervention. With rare exceptions, only well-designed and conducted RCTs can demonstrate cause and effect relationships, and only valid and useful information from RCT data should be used for interventions. Although RCT evidence regarding efficiencies, health care services and processes is often lacking, it should be sought. Only after searching and evaluating the best-available evidence should assumptions based on "common sense"[6] be substituted for valid, scientific evidence for these types of improvement.
Valid Time Intervals For Measurement	The frequency of measurement should consider health status and other factors such as cost, utilization, system impacts and patient inconvenience. Example: "The number of diabetic eye exams in adult, non-pregnant diabetics within the past year" contains a valid numerator and a valid denominator, but the time interval of every year is likely to increase costs and result in other problems without resulting in fewer patients with visual loss when compared to measuring every two years, according to the best available scientific evidence[7].
Data Gathering Validity	Data gathering validity pertains to how one actually obtains the data for numerators and denominators. Even with a valid performance measure, invalid results can occur during the data gathering process. Example: In a colon cancer screening quality improvement project, validity was threatened during data gathering because patients with exclusions (e.g., ineligible age, specific comorbidities, patients refusing

	screening, etc.) were inadvertently included in the denominator.[8] In this case, the hospital failed to meet the national target and was told that the failure could result in financial penalties. An audit revealed that forty-seven percent of the patients included in the denominator had declined screening, 12 percent failed to make their appointments for screening, 11 percent had chart documentation that screening was not indicated and 42 percent of the counted patients received diagnostic testing rather than screening (i.e., they had signs or symptoms of disease).

Performance Measures, Accountability and Pay-For-Performance—A Terrain with Pitfalls

Performance measures are an important way of focusing attention on quality improvement program priorities. The word "focus" is key – it may be that simply drawing attention to an effort or needed improvement can contribute importantly to making quality gains. Using performance measures for quality comparisons between organizations, clinical units or individuals, or for determining clinician compensation, however, raises major validity concerns and has the potential to lower quality.

Performance measurement for use beyond quality improvement is frequently required for accreditation or to create a "report card" for purchasers, consumers or an organization's marketing department. It is important to bear in mind that many of these end-users are unaware of the difficulties in comparing the quality of health care provided by different individuals, groups or systems.

When performance measures are used for quality comparisons and accountability, the requirement for using valid and useful information in the creation of the performance measure – while always important – becomes of paramount importance. For example, some organizations recommend routine screening of all newborns for hearing problems during postpartum hospitalization – this is even required by law in many states. There is, however, insufficient evidence to conclude that such testing leads to improved speech and language skills at three years of age[9]. It is also unclear from the best available evidence if the potential benefits outweigh the potential harms of false-positive tests that many low-risk infants and their parents might experience.

Risk Adjustment

Risk adjustment is a mechanism that is used in an attempt to level the playing field between organizations that are being compared when differences exist in their populations in health status or other patient characteristics, as examples. The goal is to avoid unfairly characterizing an organization as providing poorer quality care. However, there are many proprietary risk-adjustment programs available for organizations to choose from, some of whose formulas will result in better results on scorecards than others.[10] Using risk adjustment, a hospital can actually "game" the system by choosing a program which utilizes a formula which creates a large adjustment by placing as many patients as possible into a high risk category. This may result in an artificially low risk-adjusted outcome rate thereby "improving" the hospital's score. This is referred to as "coding creep" or "upcoding" of morbidities. Medical leaders, administrators and others should be aware of the significant potential for selection bias, observation bias, confounding and chance when relying on observational data—and performance measures are one form of observational data. These differences will confound performance comparisons between institutions, units and individuals. Trying to resolve this by adjusting for case mix is analogous to using models to adjust for patient differences in observational studies dealing with therapy – the potential to be mislead by confounding factors remains high.

At one hospital in New York, the proportion of patients undergoing cardiac surgery recorded preoperatively as having chronic obstructive pulmonary disease (COPD) increased from 1.8 percent in 1989 to 51.9 percent in 1995.[11] The net effect was a decrease in the hospital's risk-adjusted mortality rate resulting in a better cardiac surgery outcome on a scorecard. And what explains this dramatic improvement? It is possible that the result was due to a change in the definition of

COPD, which is equivalent to making a change in the population being measured. The reported improvement could have been due to the application of a different risk adjustment formula based on a more "lenient" definition of COPD.

Thomas and Hofer[12] modeled mortality outcomes, assuming no case mix or severity differences across hospitals, focusing on a hypothetical system of hospitals, a small portion of which delivered very poor quality care. The influences of case mix and severity differences among providers were removed through "perfect risk-adjustment." In their model they also assumed that all hospitals treated exactly the same number of patients. They reported that under optimal conditions – which they defined as "perfect risk adjustment" and no variation in patient volume among hospitals - - fewer than 12 percent of hospitals actually delivering poor quality care would be identified as high mortality rate outliers, and of the facilities identified as outliers, more than 62 percent would actually be good quality providers. They concluded that attempting to measure quality using risk-adjusted mortality rates will misinform the public about hospital performance. Similar findings have been reported by others[13], [14] while Dudley et al[15], using different assumptions, have reported more optimistic results. In summary, using performance measures to make any comparisons between organizations, units, individuals, etc., can result in erroneous conclusions about quality.

Clinician Quality

Using performance measures for comparison purposes to demonstrate which clinicians are delivering the highest quality care is problematic because these comparisons may lack validity due to the small sample sizes and other biases and confounders that come into play. Organizations and external reviewers understandably wish to have objective measures of clinician competence for privileging and credentialing purposes, and patients understandably wish to identify high quality physicians. Medical leaders are increasingly using financial incentives such as pay-for-performance strategies as powerful tools for changing physician behavior. But performance measures used as a part of a quality improvement process within an organization should not be confused with profiling individual clinicians. Performance measures relating to individual clinicians are designed to facilitate quality improvement at the point of care (internal measurement) through feedback to the individual clinician.[16] Feedback to an individual, through a performance report, may be of great value to the individual clinician as a way encouraging participation in quality improvement efforts and focusing attention on improving processes of care and attention to patients' needs.[17] However, due to significant validity and reliability problems inherent in observational data, it is altogether a different issue when an individual provider's performance is made available to others in the form of a performance "report card" or when an individual's income is based on a limited set of performance measures.

A physician may take appropriate actions to improve quality of care, but because of patient factors, systems factors or small sample size, the physician's action may not result in clinical improvement. An instructive example of how this might happen is provided by examining the use of profiling family physicians for glycemic control in their diabetic patients. In a typical family practice, only 4 percent or less of variance in hospitalization rates, visit rates, lab utilization and glycemic control (HbA1c) in diabetics was found to be attributable to differences in physician practice patterns.[18] For profiles of glycemic control, outlier physicians could dramatically improve their profiles by pruning their panels of one to three patients with the highest HbA1cs. This gaming could not be prevented by case-mix adjustment. This illustrates how patient characteristics and other factors can account for differences in performance and why performance measures may not be a valid way of evaluating an individual physician's competence. A key issue is how much of the outcome is under the control of the clinician or the health system and how much is not.

Small sample size is almost invariably a significant problem at the individual practice level and often results in validity problems due to chance or selection of patients who are not similar to "usual" patients or patients in other practices to which an individual physician is being compared. Small sample size results in insufficient power to compute statistically significant differences. In the Hofer study of HMO primary care physicians, the authors point out that to reliably distinguish physician rates in diabetes care, physicians would need to have more than 100 patients. In this study, 90

percent of primary care physicians had less than 60 patients, and none of the more than 250 physicians they studied had more than 85 diabetic patients.

Summary and Conclusions

Health care organizations are currently experiencing increasing external pressure to provide objective measures of performance to patients, insurers, accreditors, regulators, purchasers and others who are attempting to determine the quality of health care within organizations and compare the performance of individual physicians and health care organizations. Performance measurement is a quantitative way to document the results of health care interventions, services, processes or health status outcomes. Although performance measurement can play a significant role in overall quality improvement efforts, health care professionals need to understand and includes the three validity considerations when evaluating or developing performance measures (Table 1). It is equally important to avoid the pitfalls involved in the application of performance measures (Table 2.).

Table 2.

Common Pitfalls in Applying Performance Measures	
Challenge	**Cautions**
Inappropriate Time Intervals For Measurement	The frequency of measurement should consider health status and other time- and utilization-related factors such as cost, system impacts and patient inconvenience.
Using Performance Measures in Place of Clinical Guidelines	Clinical guidelines are designed to improve decision-making in health care. Guidelines are meant to be modified by users of the guidelines depending upon an individual patient's circumstances. Performance measures are designed for quantitative measurement of occurrences of health status outcomes, interventions, services or processes and, therefore, lack the flexibility of guidelines. Therefore, performance measures are much less flexible than guidelines and frequently require organizations to follow standardized and detailed steps in dealing with performance data. This rigidity can lead to lack of appropriate flexibility for clinicians, erroneous conclusions about appropriate care for individuals and potential waste and harm.
Drawing conclusions about the relative quality of individual clinicians	Feedback to an individual, through a performance report, may be of great value to the individual clinician as a way encouraging participation in quality improvement efforts and focusing attention on improving processes of care and attention to patients' needs. However, "physician profiling," i.e., holding a physician accountable for what happens to a group of patients by using a few performance measures, can be problematic. Caution is urged in implementing pay-for-performance initiatives without being sure that the performance measures are under the control of the clinician and that sufficient attention is given to all of the individual's circumstances of care. This requires a comprehensive evaluation process that involves both subjective and objective evaluation of individual clinicians.

References

1. Consensus Statement - September 16, 1998. The Urgent Need to Improve Health Care Quality Institute of Medicine National Roundtable on Health Care Quality *JAMA*.1998;280:1000-1005.

2. McGlynn EA, Asch SM, Adams J, et al. The quality of health care delivered to adults in the United States .*N Engl J Med*. 2003;348:2635-45.

3. Kerr EA, McGlynn EA, Adams J. Profiling The Quality Of Care In Twelve Communities: Results From The CQI Study**.** *Health Affairs*, May/June 2004; 23(3): 247-256.

4. Statement preceding each physician measurement set authored by AMA, JCAHO, and NCQA. Available at http://www.ama-assn.org/ama/pub/category/2946.html. Accessed June, 2005.

5. Avalilable from the National Center for Health Statistics http://www.cdc.gov/nchs/hphome.htm. Accessed June, 2005.

6. Shojania KG, Duncan BW, McDonald KM, Wachter RM. Safe but sound: patient safety meets evidence-based medicine. *JAMA*. 2002 Jul 24-31;288(4):508-13.

7. Klein R, Klein BE, Moss SE**,** Davis MD, DeMets DL. The Wisconsin Epidemiologic Study of Diabetic Retinopathy: IX Four-year incidence and progression of diabetic retinopathy when age at diagnosis is less than 30 years. *Arch Ophthalmol.* 1989;107:237–243.

8. Walker LC, Davidowitz NP, Heineken PA et al. Pitfalls of Converting Practice Guidelines into Quality Measures: Lessons Learned from a VA Performance Measure. *JAMA* 2004.291:2466-2470.

9. Thompson DC, McPhillips H, Davis RL, Lieu TL, Homer CJ, Helfand M. Universal newborn hearing screening: A summary of the evidence. *JAMA* 2001 Oct 24/31;286(16):2000-10.

10. Health care reform: 'report cards' are useful but significant issues need to be addressed. Washington, D.C.: General Accounting Office, 1994. (GAO/HEHS-94-*219.)*

11. Green J, Wintfeld N. Report cards on cardiac surgeons: assessing New York state's approach. *N Engl J Med* 1995.332: 1229-32.

12. Thomas JW, Hofer TP. Accuracy of risk-adjusted mortality rate as a measure of hospital quality of care. *Medical Care*. January 1999; 281:2098-2105.

14. Park RE, Brook RH, Kosecoff J, et al. Explaining variations in hospital death rates: Randomness, severity of illness, quality of care. JAMA 1990;264:484.

15. Hofer TP, Hayward RA. Identify poor-quality hospitals: Can hospital mortality rates detect quality problems for medical diagnoses? *Med Care* 1996;34:737.

16. Dudley RA, Frolich A, Robinowitz DL, et al.. Strategies to Support Quality-based Purchasing: A Review of the Evidence. Technical Review 10. (Prepared by the Stanford-University of California San Francisco

17. AMA Physician Performance Measurement Sets. Available at http://www.ama-assn.org/ama/pub/physician-resources/clinical-practice-improvement/clinical-quality/physician-consortium-performance-improvement/pcpi-measures.shtml. Accessed August, 2009.

18. Endsley S. Putting Measurement into Practice with a Clinical Instrument Panel. *Family Practice Management*. 2003:43-48.

19. Hofer TP, Hayward RA, Greenfield S, Wagner EH, Kaplan SH, Manning WG.. The Unreliability of Individual Physician "Report Cards" for Assessing The Costs and Quality of A Chronic Disease. *JAMA* 1999.281: 2098-2105.

