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

Long 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.

Evidence-based Performance Measurement: Considerations and Pitfalls Made Comprehensible

Acknowledgements: The authors would like to thank Brent C. James, MD, Executive Director for Intermountain Health Care at the Institute for Healthcare Delivery Research and Adjunct Professor of Family and Preventive Medicine at the University of Utah School of Medicine, and Bruce Bagley, MD, Medical Director for Quality Improvement, American Academy of Family Physicians, for their reviews of this article and suggestions.

Abstract

Evidence-based performance measurement is an important component of quality improvement efforts in health care. Performance measurement can be defined as a quantitative way to measure the care patients receive. Health care organizations are currently experiencing 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 describes the strengths and weaknesses of performance measures and how organizations can successfully find, evaluate, select and document performance measures and integrate performance measurement into an evidence and value-based approach to improving health care. It also describes the pitfalls of using performance measures to evaluate quality and competence, whether it be for health care systems, specific care units or individual health care professionals.

Background

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 outcomes within the constraints of limitations on those resources. Performance measurement is an important component of quality improvement work. Over the past several decades, performance measurement has received increasing attention which may be due to the belief that performance measures are an efficient way of evaluating quality and encouraging appropriate care. 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, non-use and mis-use of what has been demonstrated to be effective and beneficial care, such as the use of beta blockers after myocardial infarction. [1], [2],[3] A simple calculation demonstrates that inappropriate care in the United States translates into hundreds of billions of dollars of waste annually and may result in patient harms – including deaths. If performance measures can help us, 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*.

Performance measurement can be defined as a quantitative way to measure the care patients receive. The Physician Consortium for Performance Improvement defines performance measurement as "whether or how often a process or outcome of care occurs".[4] Health care organizations are utilizing performance measures, in part, because of the increasing need to provide objective documentation of performance to regulators, payers and patients. As we shall see, performance measures can be of value in increasing attention to quality improvement efforts, but have 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.

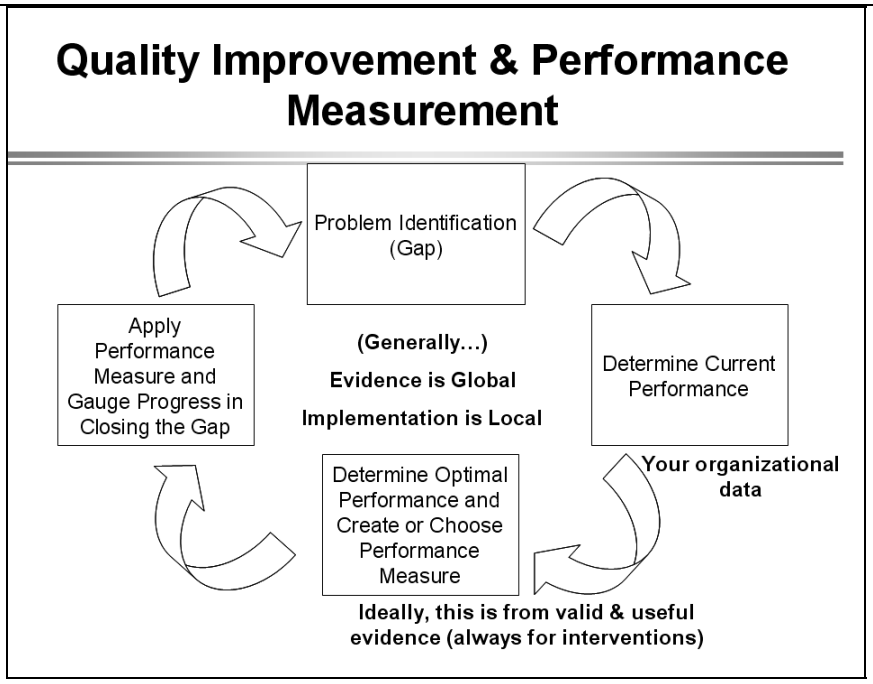
Each health care organization is capable of continuously improving quality by paying close attention to its own system of care in order to achieve desirable outcomes.[5] It is our perspective that solutions to both the quality and cost problems in health care are likely to be most effectively realized by using an evidence and value-based approach to determining

appropriateness of care and effectively implementing quality improvements. Performance measurement can contribute to more appropriate care, when utilized correctly.

The Quality Improvement Cycle and Steps for Performance Measure-based Quality Improvement

The four phases of quality improvement and performance measurement are shown below in Figure 1. For purposes of this paper, we consider "quality" as encompassing effectiveness and efficiency as well as including value considerations such as utilization, triangulation issues (e.g., legal issues, public relations, etc), and cost. We intend for this paper to be useful to people who are doing efficiency improvements – however, our primary work and the focus of this paper is on evidence-based quality improvements aimed at improving patients' health status.

Figure 1



Quality Improvement & Performance Measurement—4 Phases

1. Identify problem (i.e., conduct gap analysis)
2. Determine current performance
3. Determine optimal performance and create, or choose, performance measure
4. Apply performance measure and gauge progress in closing the gap

Within this cycle, there are eight steps that need to be considered when designing quality improvement projects (Table 1).

Table 1. 8 Steps for Quality Improvement Project Design

Steps I. through IV. Help for Selecting Good Projects

Step I. Do you have a gap between current & optimal performance? Examples include areas of uncertainty, gaps in clinical care affecting health status, satisfaction or cost.

Step II. What will close the gap and improve quality? Search for valid and useful evidence for quality improvement effort. IF THERE IS NO VALID, USEFUL SCIENTIFIC EVIDENCE AND THIS IS FOR A CLINICAL INTERVENTION = STOP. Do not do an "improvement." If this is for efficiency, satisfaction, a service or a process, ideally you want to use valid and useful scientific evidence. If none is available, assumptions will need to be substituted for evidence.

Step III. Is attempting the improvement feasible in your environment?

- Are you going to be able to successfully make clinical practice change happen?
- Are resources available to support the initiative?

Step IV. Can you measure it?

- Is your measure quantifiable?
- Is your measure valid, accurate and dependable?
- Is the measure useful and usable?
  - Is it comprehensible?
  - Will it assist with QI projects?
- Is measurement achievable in your local circumstances?

Measure Name / Descriptor / Validity Consideration	Example
Numerator = the occurrence you are counting. Validity is ideally based on valid, useful evidence. (It must always be based on valid, usable evidence for interventions.) The evidence is directly applied or inferred from measuring success of implementation.	An Rx for an ACEI or an ARB
Denominator = the pool eligible for the count. Validity is based on inclusions and exclusions to achieve who or what is eligible for measurement.	In patients with heart failure, who are admitted to a hospital, and who do not meet exclusion criteria
Frequency = time interval for the numerator (e.g., the performance or process). Validity is ideally based on valid, useful evidence.	By the time of hospital discharge

Steps V. through VIII. Help for Applying Performance Measures

Step V. How are you going to gather the data to measure the improvement?

Step VI. What is the meaning of your measurement, (i.e., what goal will you set to define "improvement" such as a trend, a target or a statistically significant change)?

Step VII. How are you going to report it and to whom?

Step VIII. What is your process for updating your improvement?

What Good Performance Measures Look Like

The purpose of a performance measure is to count specific events or elements of interest that occur within a given time frame. Performance measures count such things as health status and health care outcomes; diagnostic, therapeutic or monitoring interventions; health care services or processes (see Table 3. for a more detailed list). Thus good performance measures must be quantifiable, valid, useful and usable, and fit neatly within the larger context of clinical quality improvement efforts, including selecting meaningful projects that have a high likelihood of success. A performance measure consists of a numerator, a denominator and a frequency – and it is often expressed as a percentage or 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." The "formula" for this performance measure is shown below (Table 2):

Table 2. "Formula" For A Performance Measure	
Numerator (including frequency) =	**Patients from denominator receiving an annual hemoglobin A1c**
Denominator =	**Patients with diabetes mellitus (meeting the inclusion criteria and not excluded using the exclusion criteria)**

- 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.

Validity refers to "closeness to truth" – meaning we have measures that will give us accurate and dependable information. Useful means that the measure correlates with something that will be meaningful to us in our efforts to improve our performance. For health status outcomes and interventions, we want performance measures that address effectiveness, are beneficial and relevant to our population, and which patients and clinicians will accept and apply appropriately. Usability is distinguishable from usefulness. A fire hose is useful, but not very usable to individuals who are not trained as fire fighters. To be usable, performance measures must be comprehensible and actionable.

The denominator can be thought of as the pool of eligibility or base number of units from which measurements are taken – the "who." 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. It also can be the total number of patient charts, visits, etc.

The numerator is a count taken from the denominator, which expresses the number of occurrences of the event of interest – the "what." Examples would include the number of patients with diabetes mellitus from the denominator (the eligible "pool") who have received an annual hemoglobin A1c, or the number of patients from the denominator whose wait time is ten minutes or less for their appointment at an outpatient clinic.

Numerators generally measure something patients receive or something that is done to a patient. The categories of numerators, along with examples are listed in Table 3.

Table 3. Examples of Topics (Numerators) for Performance Measurement		
	Category	**Examples**
OUTCOME-RELATED OCCURRENCES	**Health status outcomes**	Morbidity, mortality, symptom relief, functioning and health-related quality of life **Caution:** Unless you are doing a high quality research study, health outcomes are **highly susceptible to confounding** which can result in misleading conclusions
	Health care outcomes	Satisfaction scores, behaviors, costs, etc. **Caution**: Satisfaction and behaviors are highly susceptible to confounding.
PERFORMANCE-RELATED OCCURRENCES	**Interventions**	Diagnostic, therapeutic and monitoring interventions such as procedures, lab tests, imaging tests, referrals, prescriptions, surgical or non-surgical therapeutic interventions, counseling and advice given, etc.
PROCESS-RELATED OCCURRENCES	**Services delivered to patients**	Appointments, visits, referrals, phone calls, providing information, etc
	Processes	Those activities which may be "invisible" to patients, but affect care such as hours worked, daily calibration of equipment or availability of x-rays to doctors

Frequency reflects the intervals of time that occurrences should take place (e.g., **annual** diabetic foot exam).

Validity Issues

Validity—closeness to truth– as it pertains to performance measures includes several considerations. First and foremost, if performance measures are to lead to improvements in care – and not increased waste and harms – then performance measures must account for performance-based occurrences such as interventions, services and processes which are based on valid and useful information indicating that those performances will truly improve the quality of care. Secondly, the measures, themselves, must be valid. A valid measure is one that "measures what is purports to represent," meaning that it is true and can be relied upon to measure actual clinical, service or cost improvement, as examples. Simply put, the measure must measure what it is intended to measure. Thirdly, how we go about accomplishing the measurement must be done in a valid way. Validity for a performance measure is summarized below (Table 4).

Table 4. Performance Measurement and Validity Considerations	
1. Performance Measure Validity Considerations	
Measures chosen to document quality are valid as determined by validity of the denominator, the numerator, the frequency and the measure itself.	
Denominator Validity	Means that the denominator has the right inclusions and exclusions to identify the right pool for measurement.
Numerator Validity	1. If you have valid, usable evidence demonstrating a cause and effect relationship for outcomes, you will wish to count instances supporting that the evidence has been effectively applied. This may be a direct count of the intervention. If a direct count is not possible, you will wish to count a performance or process that is associated with the intervention and from which you can reasonably infer that the intervention happened. 2. If you do not have valid, usable evidence from which you can conclude cause and effect, there is no assurance that numerators for outcomes will be valid. When constructing numerators in this situation, assumptions and judgments must be substituted for evidence. IF THERE IS NO VALID, USEFUL SCIENTIFIC EVIDENCE AND THIS IS FOR A CLINICAL INTERVENTION = STOP. Do not do an "improvement."
Frequency Validity	Requires appropriate time intervals for the events of interest. Ideally, these are evidence-based.
Measure Validity	Measures need to be accurate and dependable. ▪ **Accuracy** is the ability of the performance measure to correctly identify the events of interest that it is designed to identify. ▪ **Dependability** is the ability to provide consistent measurement on repeated testing.
2. Data Gathering Validity Considerations	
Data gathering validity pertains to the obtaining of numerators and denominators. Even with a valid performance measure, invalid results can occur if appropriate data are not obtained or gathered correctly. For example, invalid results can occur when the numerator includes denominator exclusions, such as counting pregnant women with diabetes when they are an exclusion from the denominator. As another example, gathering only inpatient data may miss diagnostic tests performed in the outpatient setting, such as for left ventricular function.	

Valid Denominators

A valid denominator means that it specifies the right population or the right base number of units from which the measurement will be made, (e.g., the "pool.") The "right population" or the "right base number of units" means that the denominator has appropriate inclusion and exclusion criteria for the "eligible for measurement" pool. For example, a denominator measuring clinical improvement for PAP smears which does not exclude women without a cervix would be invalid. A performance measure targeting a specific drug without accounting for drugs that patients may have acquired over the counter (e.g., aspirin) would be invalid.

Choosing an appropriate denominator is more complicated than it might seem. Special care must be taken to ensure that the denominator includes all of those patients or units – and **only** those patients or units – that should be counted. This requires close attention to the inclusion and exclusion criteria for the denominator.

Some key considerations in constructing a denominator include:

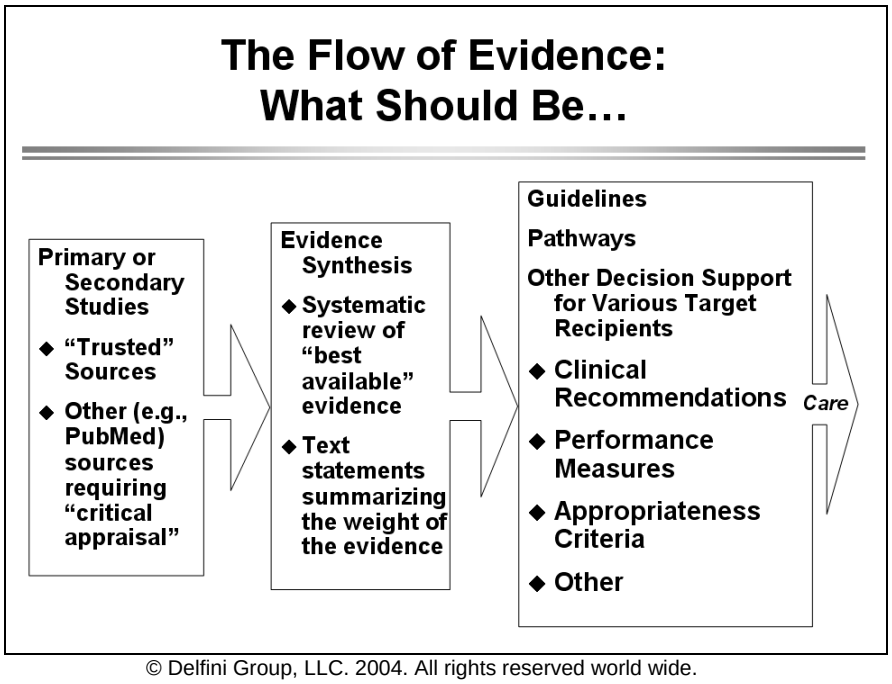
- **Appropriateness**: What makes a patient not only appropriate – but also inappropriate – for inclusion in the denominator? Does the patient actually have the diagnosis? Does the clinician deem the patient an appropriate candidate for care – or is there documented clinical judgment that the patient should be excluded from consideration for the improvement? Are patients with terminal conditions or severe comorbidities excluded?
- **Control**: Do exclusions account for patient refusals, no-shows, non-adherence or the possibility of being screened or cared for elsewhere? Do exclusions account for system barriers such as access problems or cost, if these are areas that are currently not your focal point for improvement?

Valid Numerators

The numerator counts occurrences of events such as outcomes, interventions, services or processes. The numerator is a subset of the denominator in that it is a count taken from the denominator in an attempt to quantify whether an improvement has taken place.

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 can result in improvement. To do this, performance measures need to be based on valid, useful and usable scientific evidence. What we do not want to measure is whether or not there has been a health status improvement unless we are doing a high quality, randomized, controlled trial, because we run the risk of obtaining misleading results due to bias, confounding and chance. Therefore what we want to measure when assessing health status outcomes is our performance or our processes that we know from valid, scientific evidence lead to improved health status. Almost invariably this requires randomized, controlled trials—not observational studies. In order to critically appraise studies for validity, usefulness and usability, skills in all the processes for effectively searching, evaluating scientific validity, assessing the usefulness of results from valid studies, creating evidence syntheses, constructing clinical recommendations, projecting impacts of practice change, implementing change and re-measuring performance are required. This "flow" of evidence into clinical care is illustrated below (Figure 2).

Figure 2



An example of a numerator in this instance might be counting prescriptions for ACE inhibitors. If we do not have strong evidence of benefit, we should not do the "improvement," since we do not actually know if we will help or harm the patient and/or waste our limited resources.

Although evidence regarding efficiencies is often lacking, it should be sought. Sometimes there is very helpful evidence for efficiency. Only after searching and evaluating the literature should assumptions based on "common sense"[6] be applied.

How Evidence Should Flow into Clinical Recommendations as the Basis for Performance Measures

Take, for example, a performance measure with an invalid numerator which counts something with no demonstrated association with true clinical improvement. If we choose as a measure of quality the readmission rate to an intensive care unit (ICU) within 30 days of discharge from the unit, and the readmission rate for a particular intensive care unit is high, the intensive care unit may falsely appear to be performing sub-optimally. There may be many other known and unknown confounders –- meaning, associated causes – that could result in readmission. The reason for readmission might be a result of poor care delivered on the general medical ward, for example. In this case, the numerator will not measure quality provided within the ICU.

Accuracy and Dependability of the Performance Measure

Performance measures should be accurate, meaning they need to correctly identify the occurrences such as the interventions, services or processes that they are designed to identify. They must also be dependable, meaning upon repeated testing, the measures provide consistent results. For example, when weighing patients we want the scales to yield measurements that are accurate and consistent.

Valid Frequencies

For frequencies, validity means that the intervals of the occurrences – such as annually performed interventions, services or processes – are appropriate. For instance, 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 frequency of every year is likely to increase costs without resulting in fewer patients with visual loss when compared to a frequency of every two years, according to the best available scientific evidence[7].

Data Gathering Validity Considerations

Data gathering validity pertains to how one actually obtains the data for numerators and denominators. Effective data gathering requires an approach for collecting data, i.e., for measuring or counting the number of instances of an improvement. This may be a direct method such as measuring the direct application of evidence which has been shown to result in improved outcomes (e.g., the number of patients receiving beta-blockers after MI) or measuring an intermediate marker that has evidence of a chain of events leading to a benefit (e.g., improving blood pressure where there is strong evidence of reduced risk of cardiovascular events). Or an indirect method may be used to measure whether an improvement was effectively implemented – such as using a proxy to determine if an evidence-based improvement has occurred (e.g., the percentage of diabetics who received eye exams or fundal photographs).

Even with a valid performance measure, invalid results can occur during the data gathering process. It is in the data gathering efforts where the greatest challenges, obstacles and opportunities for error are to be found. Considerations include actually getting to the right people or getting the right units from which to count and ensuring that the count is made correctly.

Imagine that an organization has selected colon cancer screening as a quality improvement project and has selected a valid performance measure. Validity is threatened if, during data gathering, patients with exclusions (e.g., ineligible age, specific comorbidities, patients refusing screening) are inadvertently included in the denominator. Similar errors occurred during the data gathering phase of a colon cancer screening performance measure project at the Veterans Affairs hospital in San Francisco.[8] In this case, the VA 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 in the denominator had declined screening, 12 percent failed to show up 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).

Data gathering validity needs to take into account the sources for data and how the data are collected. Some considerations include whether the sources for data seem reasonable, whether items of interest are sufficiently comprehensive and defined (e.g., diagnosis and treatment), appropriate responses for variables, whether bias might be introduced through data collection and ensuring that data is collected at the right point to give sufficient time for recording the occurrence of interest.

Performance Measures and Clinical Practice Guidelines

Over the past three decades, clinical practice guidelines have been increasingly used by many health care organizations as important quality improvement tools. On the face of it, many performance measures have the appearance of clinical recommendations extracted from guidelines. However, several important differences between clinical guidelines and performance measures should be noted because understanding the differences can decrease the likelihood of misapplying performance measures.

Firstly, guidelines are not rules or standards, but statements providing guidance, advice or options in a general way for clinicians and patients to use in making decisions. Good clinical guidelines assist clinicians and patients in making appropriate health care decisions. They are based on the best available evidence, clinical judgment and patient

preferences. Guidelines can be very helpful in closing quality, satisfaction, cost and uncertainty gaps. The clinical recommendations contained within guidelines may be modified by users of the guidelines depending upon an individual patient's circumstances.

Conversely, performance measures are designed for quantitative measurement of occurrences of health status outcomes, interventions, services or processes. The two main uses for performance measures are for quality improvement (i.e., for focusing attention and resources on closing quality gaps through quality improvement projects), and to document quality in an attempt to demonstrate accountability.

Performance measures are frequently used to compare the performance of one health care organization with another or to compare an organization's performance to local or national targets or outcomes. Therefore, although performance measures may be developed from valid clinical guidelines, 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 serve as a trap which can lead to mistaken conclusions and potential waste and harm if one isn't sufficiently knowledgeable about both the potential utility and the potential pitfalls of performance measures.

This lack of flexibility makes it critical to ensure that the denominator is a valid one and that the numerator truly measures an event that results in quality. An invalid denominator that does not exclude patients for whom an intervention is contraindicated penalizes clinicians who may be delivering quality care and puts patients at risk of harm. A numerator that is not associated with quality may mislead, cause waste and drive up costs – and may cause harm to patients as well.

On the other hand, well-crafted performance measures can have some advantages over guidelines. They can be simpler and easier to recall – and thus, potentially easier to implement, especially for busy clinicians who value quick, point-of-care information tools. Some organizations may find it easier to emphasize performance measures rather than guidelines since there may be numerous – and potentially conflicting – guidelines from numerous sources with differing recommendations for a single clinical condition. With a good performance measure it is less important that a physician utilize guideline A or guideline B as long as the performance measure fits with a valid clinical guideline.

Selecting Good Quality Improvement Projects

Performance measures are frequently used to determine how successful an organization's quality improvement efforts have been. Evidence-based clinical improvement efforts close gaps between an organization's current practice and optimal practice as determined by the best available evidence. For interventions, it is imperative that valid and useful scientific evidence be used. Selecting good quality improvement projects entails a minimum of four considerations: 1) the importance of the area for clinical improvement, 2) the existence of useful and usable evidence upon which quality improvements can be based, 3) feasibility, meaning the ability of the organization to carry out the improvement, and 4) measurability of the quality improvement.

Consideration 1: Importance of the area for clinical improvement

In selecting a work area for quality improvement, the first question for a group to answer is, "Do we have a gap between our current care and what we consider to be optimal care, or do we have an area of significant uncertainty that requires attention?" Gaps may be present in health status or health care outcomes, interventions, services or processes. In order to identify gaps it is useful to look for inappropriate variation, clinical uncertainty, sub-optimal performance in the areas of health status (mortality, morbidity, symptom relief, functional status or health-related quality of life), patient satisfaction, staff satisfaction and value considerations such as liability, cost and inappropriate utilization.

Consideration 2: Useful and usable evidence upon which quality improvements can be based

To more accurately predict what will occur with a change in clinical practice, it is important to have evidence. For interventions, it is imperative that we have valid and useful evidence to inform quality improvement efforts about the effects – that is., the benefits and risks of potential harms – of an intervention on various outcomes. Useful evidence is scientific evidence that is valid and results in a net benefit for the target population. The changes in care should be acceptable to clinicians and patients, and they should be actionable.

Consideration 3: Feasibility —the organization's ability to carry out the improvement

To successfully carry out a quality improvement initiative, an organization should be sure that the change in clinical care is beneficial, acceptable, achievable and measurable. It is important to know if the change is required by regulators. Organizational leadership support is imperative as are the needed structures, processes, tools and available skilled staff in the right roles for carrying out the improvement. A useful way of assessing the feasibility of a quality improvement project is to evaluate and find means to address the driving and restraining forces which will assist with or hinder successful implementation of the clinical practice change within the organization.

An organization must also identify possible implementation strategies and ensure that means are available to carry out those strategies. Successful implementation of change is much more likely if a "champion" or lead is available to guide the quality improvement project. Educational programs providing training in evidence-based quality improvement basics which target all medical and management leaders and clinical and non-clinical staff team members are worthwhile because such programs can create individual or work-unit awareness, drawing attention to the effort along with providing needed skills. Plus interested individuals can form a pool from which committee members and work group participants can be drawn. These strategies help an organization create an evidence-based culture which makes implementation of change all the easier.

A specific implementation plan needs to be created, and following these considerations, it is important to assess potential success and to estimate the effort, expense and other resources that will be required to close the gap, and to assess all anticipated impacts of change, in order to decide if the quality improvement justifies the efforts and resources required and if it justifies all the possible untoward outcomes which could result from the change.

Consideration 4: Measurability

Measurability refers to the achievability of measurement within specified settings. A performance measure should be –

- **Quantifiable:** Translatable into a numerator and a denominator and be associated with a frequency for the occurrence to be measured.
- **Valid:** Denominator, numerator and frequency are all valid and will measure what they purport to measure.
- **Accurate**: Correctly identify the occurrences it is designed to identify.
- **Dependable**: With repeated testing measurement should give the same results.
- **Useful and usable:** The measure should be easily understood by all users. To be useful, a measure should assist with quality improvement efforts and, in some instances, have the potential for risk stratification or risk adjustment. Risk stratification takes into account the characteristics of individual patients and is a process by which an individual's risk is calculated by considering various factors. Risk stratification is extremely useful because it provides information that is specific to each individual. For example we stratify an individual's cardiovascular risk by considering blood pressure, lipid levels, smoking history and history of diabetes. Risk adjustment takes into account the health status of a population being measured as a whole and is a process by which an actual rate of events is adjusted for various factors that may confound the results which would otherwise over- or understate the magnitude of those results. For example, adjusting a surgical mortality rate

for a comorbidity such as chronic obstructive pulmonary disease (COPD) will result in a lowered mortality rate. This is called an adjusted mortality rate. While risk adjustment is an attempt to decrease reporting bias to adjust for confounders, there remains an enormous potential for misleading results which we will discuss this later in the paper. Risk adjustment cannot be relied upon to remove bias, confounding and chance – and frequently, it will add another element of bias making results even more misleading.

- **Achievable**: Achievability means that the measurement can be accomplished in the setting intended, and that data will be available and collectable within a reasonable time frame. When considering achievability it is useful to ask the following questions:
- Does the source for data collected seem reasonable? For example, if looking for a diagnosis of left ventricular dysfunction, were both inpatient and outpatient sources utilized?
- Are items of interest sufficiently comprehensive and defined (e.g., diagnosis and treatment)? An example would be defining the age and precise laboratory cut-off values for a diabetes performance measure.
- What are the possible responses for clinician and patient variables (e.g., documentation that the intervention is inappropriate or contraindicated, or that the patient declined the intervention)?
- Might bias be introduced through any data collection process, instrument or survey tool or through the means of its administration? Example 1: A nurse asking patients questions about satisfaction following an appointment may create more bias than a mailed survey. Example 2: Analysts including invalid inclusions or exclusions in numerator or denominator will invalidate the results.
- Is the data collection timeframe sufficient to get appropriately meaningful results given the area of interest? Consider: Primary effect, side effects, symptoms, regression, remission, recurrence, survival, etc.

Finding and Evaluating Performance Measures

Some groups will develop their own performance measures. Most, however, will utilize performance measures developed by other institutions and will integrate performance measures developed by others into their own quality improvement projects.

It should be kept in mind that many performance measures (such as clinical guidelines and other secondary clinical sources) are not evidence-based – and often, even when they use the "evidence-based" label. Therefore workgroups should include individuals with the skills to evaluate and apply medical evidence and evaluate clinical guidelines. This is imperative for performance measures which address health status outcomes and interventions. An example of an organization with the mission of improving American healthcare through endorsement of consensus-based national standards for measurement and public reporting of healthcare performance data is the National Quality Forum (NQF). NQF is a

private, not-for-profit membership organization created to develop and implement a national strategy for healthcare quality measurement and reporting. "Seed" performance measures for hospital care, safe practices, children's health care, diabetes care and other practices are available from NQF.[9] It should be remembered, however, that measures developed by NQF are developed through a consensus process, and it will be necessary to evaluate them for validity and usefulness.

The following are sites are sources for performance measures. Measures adapted from these sites should be evaluated using criteria for validity and usefulness:

1. **National Quality Measures Clearing House**

   http://www.qualitymeasures.ahrq.gov/

2. **National Committee for Quality Assurance (NCQA)**

   HEDIS Effectiveness of Care Measures (2 URLS)

   a. http://www.ncqa.org/Programs/HEDIS/

   b. http://www.ncqa.org/Communications/State percent20Of percent20Managed percent20Care/SOHCREPORT2003.pdf

3. **National Quality Measures Clearing House**

   http://www.qualitymeasures.ahrq.gov/

4. **National Quality Forum**

   http://www.qualityforum.org/home.htm

5. **Joint Commission on Accreditation of Healthcare Organizations (JCAHO)**

   http://www.jcaho.org/pms/index.htm

6. **Physician Consortium for Performance Improvement**

   http://www.ama-assn.org/ama/pub/category/2946.html

7. **American Academy of Family Physicians (AAFP) Performance Measurement**

   http://www.aafp.org/x18919.xml

8. **Ambulatory Quality Alliance**

Founded by AAFP, ACP, AHIP and AHRQ. Endorsed 26 performance measures as starter set. Goal is to focus on agreed-to measures accepted by health plans to eliminate tracking different measures for different initiatives.

   http://www.aqaalliance.org/

   Quality Measures

   http://www.aqaalliance.org/performancewg.htm

   Individual practice internal improvement demo flow sheets:

   http://www.ama-assn.org/ama/pub/category/4837.html

   Registry

   http://www.aafp.org/fpm/20060400/diabetesregistry.xls

9. **National Guideline Clearinghouse**

   www.guidelines.gov

10. PubMed

www.pubmed.gov

Baseline Measurement and Goal Setting

Once an organization has determined what constitutes effective improvement and has evaluated and selected a performance measure, the organization needs to determine its baseline, and it needs to determine its goals and how it will define "improvement." 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 significant gains in improvement demonstrated through measuring significant change?
- **Target:** Was a specific goal or range reached that we have defined as "improvement?"

Statistically significant change may offer some advantages over setting a target in that it may be most reflective of true improvement. Imagine a process improvement for appointment waiting times. A target set at ten minutes maximum wait per patient may not demonstrate true improvement for a unit that is already operating near the target level as compared to a unit that has been operating at an average of twenty minutes and improves to an average of eleven minutes.

However, most organizations set targets as goals for improvement. Targets are frequently set by considering the performance of top-performing organizations, national or local averages and recommendations of groups such as Healthy People 2010[10]. Targets may also be set by considering baseline data and targeting statistical improvement. It is useful to determine where the organization should be within a given time period.

It is important that an organization not set its targets unrealistically high. Constructing denominators with the right exclusions can help significantly. If you have not accounted for the following in your denominator, then – before setting your targets – you will need to consider issues affecting control and appropriateness such as patients not consenting; potential for non-adherence; system problems (e.g., costs and difficulties which will prevent some patients from adherence); patients getting care outside the system; inappropriate patients for intervention due to comorbidities or other reasons; and, patients likely to die within the next several months. Then, compare your proposed target or other definition of "improvement" to your current performance to do a rough reality check before implementing your measure. Many organizations use statistical process control (SPC) charts (line graphs) to graphically display changes in performance. SPC charts allow readers to quickly see a visual tracking of changes in performance from previous levels.

Performance Measures and Accountability

Performance measures can be utilized at any level within an organization, such as at the individual level, the clinical department, the clinic or hospital unit – or the hospital or the organization itself. Performance measures are most useful – and least problematic – when used internally by a health care organization as part of an evidence-based clinical quality improvement effort. In these instances, performance measures help to determine whether the quality improvement effort has been effectively implemented. Also they 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, however, raises major validity concerns.

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. Two of the largest accrediting bodies, the Joint Commission on Accreditation of Health Care Organizations (JCAHO) and the National Committee on

Quality Assurance (NCQA) attempt to standardize the performance measurement process to obtain valid measurements. These organizations make performance measurement sets available to purchasers, payers and patients.

JCAHO evaluates hospitals, ambulatory care centers, long term care facilities and health care networks. JCAHO provides a web site[11] that allows evaluators to see a hospital or facility's scorecard of their performance in the following areas:

- Patient Rights and Organizational Ethics
- Assessment of Patients (e.g., lab services)
- Care of Patients (e.g., medication use)
- Education
- Continuity of Care
- Improving Organizational Performance (e.g., measurement of processes and outcomes)
- Leadership
- Management of the Environment of Care (e.g., safety plans)
- Management of Human Resources (e.g., education of staff)
- Management of Information
- Infection Control
- Governance
- Management
- Medical Staff (e.g., credentialing)
- Nursing

NCQA evaluates managed care organizations in the key areas of quality of care, access to care and member satisfaction with the health system and doctors. NCQA encourages the use of its report card[12] by employers, consumers and others in order to compare health care systems in the following areas:

- Access and Service
- Qualified Providers
- Staying Healthy
- Getting Better
- Living with Illness

When performance measures are used for accountability and quality comparisons, the requirement for using valid and useful information in the creation of the performance measure – while always important – becomes of paramount importance. Imagine the creation of a performance measure for an intervention that is based on no scientific evidence of benefit and actually results in harms and tremendous costs – and which gets applied nationally as a quality indicator. The potential for inappropriate care is huge.

And yet, this happens. 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[13]. It is 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 following universal hearing screening of newborns.

Using performance measures for comparison purposes to demonstrate which organization or clinician is delivering the highest quality care is extremely problematic because these comparisons may lack validity due to the small sample sizes and numerous biases and confounders that come into play. This can be most easily illustrated by examining the problems of observational studies published weekly in the medical literature and understanding how the same problems with observational data occur when measuring performance within an organization. Medical leaders, administrators and others should be aware of the significant potential for selection bias, observation bias, confounding and chance in

observational data—and performance measures are one form of observational data. For example, some hospitals will attract patients with multiple and more complex problems than other hospitals. Differences in patient characteristics (e.g., comorbidities, differences in severity of illness, likelihood of adhering to treatment plans and countless unique, individual variations) will always be an issue when comparing different health care systems. 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 published observational studies dealing with prevention, screening and therapy.

A classic illustration of this is that numerous well-done observational studies with well-done adjustments repeatedly resulted in erroneously reporting that hormone replacement therapy (HRT) was effective in reducing cardiovascular events in postmenopausal women with coronary artery disease (CAD).[14] In these studies, women chose to take HRT. They were compared to "matched" women who chose not to take HRT. It is very likely that women who chose to take HRT were significantly different (e.g., they were more health-conscious in diet and exercise) from women who chose not to take HRT. Therefore, when the incidence of CAD was compared, there was a much lower incidence of CAD in women choosing to take HRT. Recently, randomized controlled trials reported no cardiovascular protection and revealed increased risks such as breast cancer and thromboembolic complications. [15], [16], [17] In the observational studies, there was an association between taking HRT and a lower incidence of CAD, but the association was due to a selection bias—i.e., healthier women, leading healthier lives represented a "selection bias."

Drawing conclusions about health status outcomes requires the use of high quality research studies – meaning valid randomized controlled trials (RCTs) with useful results. This type of data is not available from health care systems or hospitals delivering care in the community – or individual health care providers – unless obtained through a high quality RCT involving those care entities. For example, even when organizations have valid numerators and denominators and have done their data collection well, their populations very likely differ in ways that make comparisons between organizations problematic.

Comparing organizations on the basis of their care processes carries the same risks as comparing health status outcomes. For example, when comparing processes of hospital A with those of hospital B, what if the hospitals care for different populations or use different risk adjustment formulas?[18] In comparing care processes between organizations it is impossible to be sure that the patients, locations, care experiences, interventions and analyses are similar enough to "rule out" bias, confounding or chance as possible explanations for the differences between the organizations. All observational comparisons are confounded and cannot be used to draw cause and effect conclusions.

Risk Adjustment

Risk adjustment is a mechanism that is used in an attempt to level the playing field between organizations when they may have differences in their populations arising from differences in health status or other patient characteristics. The goal is to avoid unfairly characterizing an organization as providing poorer quality care. Yet validity issues may arise when one hospital expends great effort to adjust for the disease severity of a population while another does not. Further, using risk adjustment, a hospital can actually "game" the system by choosing 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" their score. This is referred to as "coding creep" or "upcoding" of morbidities. 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.[19]

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.[20] The net effect was a decrease in the hospital's risk-adjusted mortality resulting in a better cardiac surgery outcome on a scorecard. And what explains this dramatic improvement in the adjusted mortality rate? 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[21] 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% 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% 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[22, 23] while Dudley et al[24], 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.

Special Considerations – Using Performance Measures for Accountability of Individual Clinicians and Payment for Performance

Extra caution is also urged when considering applying performance measures to evaluate individuals such as physicians or other health care professionals.

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. The Physician Consortium for Performance Improvement has clearly stated that 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.[25]

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.[26] 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 performance measures.

Because of small sample size and multiple confounders it is, impossible to reliably establish an individual's competence through use of a few performance measures. Performance measures can, in some instances, signal a need to look closely at an individual's circumstances of care, practice pattern and patient population. This, however, is a far cry from "physician profiling" i.e., holding a physician accountable for what happens to a group of patients by using a few performance measures.

A physician may take appropriate actions to improve quality of care, but because of patient 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 regarding 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, glycemic control (HbA1c) in diabetics, was found to be attributable to differences in physician practice patterns.[27] 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. We can, therefore, see how patient characteristics and other factors account for difference 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 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 frequently results in validity problems due to selection bias, i.e., the patients are not similar to "usual" patients or patients in other practices to which an individual physician is being compared. Or the small sample size results in insufficient power to compute statistically significant differences, thereby increasing the likelihood that any differences observed are merely due to chance. In one study of HMO primary care physicians, Hofer et al[25] 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.

The simplest way to summarize the issue is this – only valid and relevant randomized controlled trials can sufficiently eliminate bias and confounding and thereby allow cause and effect conclusions about treatments to be reliably drawn. Except as part of a research study, patients are rarely, if ever, randomized to physicians' practices. Thus, a single physician's practice report card is equivalent to a case series or observational study. And worse, case series and observational studies can result in highly misleading – even harmful – conclusions. Physicians should understand these pitfalls in order to make informed decisions before agreeing to participate in various reward or pay-for-performance initiatives, and medical directors should realize the potential for rewarding poorer quality performance and penalizing good performance by instituting this kind of system.

Summary

Health care organizations are currently experiencing external pressure to provide objective measures of performance to patients, insurers, accreditors, regulators, purchasers and others to evaluate the quality of health care within organizations and compare the performance of individual physicians and health care organizations. Performance measurement is a way to document, in a quantitative way, the results of health care interventions, services, processes or health status outcomes. Health care professionals need to understand the strengths and weaknesses of performance measurement and how performance measures can play a significant role in their overall quality improvement efforts. Core competencies include how to find, develop, evaluate and implement performance measures. Application of these competencies can help achieve truly evidence- and value-based clinical quality improvement and help avoid some serious pitfalls which could result in waste, increased costs, misuse of organizational resources and patient harms.

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 August, 2004.

5. Strite, SA, Stuart ME. "The Evidence-based Organization: Part I. Introduction to the Evidence and Value-Based Health Care System." *The Physician Executive, Journal of Medical Management*. In press.

6. Shojania KG, Duncan BW, McDonald KM, et al. 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**,** et al. 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. National Quality Forum
http://www.qualityforum.org/publications.html. Accessed November, 2004.

10. Avalilable from the National Center for Health Statistics http://www.cdc.gov/nchs/hphome.htm. Accessed August, 2004.

11. Joint Commission on Accreditation of Healthcare Organizations. http://www.jcaho.org/quality+check/index.htm. Accessed August, 2004.

12. NCQA's Health Plan Report Card. http://hprc.ncqa.org/frameset.asp. Accessed August, 2004.

13. Thompson DC, McPhillips H, Davis RL, et al. Universal newborn hearing screening: A summary of the evidence. *JAMA* 2001 Oct 24/31;286(16):2000-10.

14. Sullivan JM, Vander Zwaag R, Hughes JP, et al. Estrogen replacement and coronary artery disease. Effect on survival in postmenopausal women. *Arch Intern Med*. 1990;150:2557–2562.

15. Hulley S, Grady D, Bush T, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research Group. *JAMA*. 1998 Aug 19;280(7):605-13.

16. Herrington DM, Reboussin DM, Brosnihan KB, et al.. Effects of estrogen replacement on the progression of coronary-artery atherosclerosis. *N Engl J Med* 2000;343:522-9.

17. Risks and benefits of estrogen plus progestin in healthy postmenopausal women. Authors: Writing Group for the Women's Health Initiative Journal: *JAMA* 2002;288:321–333.

18. Iezzoni LI. The risks of risk adjustment. *JAMA* 1997.278: 1600-7.

19. 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.)*

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

21. 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.

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

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

24. 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

25. 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.

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

27. Hofer TP, Hayward RA, Greenfield S et al. The Unreliability of Individual Physician "Report Cards" for Assessing The Costs and Quality of A Chronic Disease. *JAMA* 1999.281: 2098-2105.

