Journal of the American College of Cardiology
© 2010 by the American College of Cardiology Foundation and the American Heart Association, Inc.
Published by Elsevier Inc.

Vol. 55, No. 16, 2010
ISSN 0735-1097/10/$36.00
doi:10.1016/j.jacc.2010.02.016

PERFORMANCE MEASURES

ACCF/AHA 2010 Position Statement on Composite Measures for Healthcare Performance Assessment

A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Performance Measures (Writing Committee to Develop a Position Statement on Composite Measures)

WRITING COMMITTEE

Eric D. Peterson, MD, MPH, FACC, FAHA, Chair; Elizabeth R. DeLong, PhD; Frederick A. Masoudi, MD, MSPH, FACC, FAHA; Sean M. O'Brien, PhD; Pamela N. Peterson, MD, MSPH, FACC; John S. Rumsfeld, MD, PhD, FACC, FAHA; David M. Shahian, MD, FACC; Richard E. Shaw, PhD, MA, FACC, FAHA

ACCF/AHA TASK FORCE ON PERFORMANCE MEASURES

Frederick A. Masoudi, MD, MSPH, FACC, FAHA, Chair; Elizabeth R. DeLong, PhD; David C. Goff, Jr, MD, PhD, FAHA, FACP; Kathleen Grady, PhD, RN, FAHA, FAAN; Lee A. Green, MD, MPH; Kathy J. Jenkins, MD, MPH, FACC; Ann Loth, RN, MS, CNS; Eric D. Peterson, MD, MPH, FACC, FAHA; Martha J. Radford, MD, FACC, FAHA; John S. Rumsfeld, MD, PhD, FACC, FAHA; David M. Shahian, MD, FACC

TABLE OF CONTENTS

The American College of Cardiology Foundation and the American Heart Association make every effort to avoid any actual or potential conflicts of interest that may arise as a result of an outside relationship or a personal, professional, or business interest of a member of the writing panel. Specifically, all members of the writing group are required to complete and submit a Disclosure Questionnaire showing all such relationships that might be perceived as real or potential conflicts of interest.

This document was approved by the American College of Cardiology Foundation Board of Trustees in November 2009 and the American Heart Association Science Advisory and Coordinating Committee in October 2009.

The American College of Cardiology Foundation requests that this document be cited as follows: Peterson ED, DeLong ER, Masoudi FA, O'Brien SM, Peterson PN, Rumsfeld JS, Shahian DM, Shaw RE. ACCF/AHA 2010 position statement on composite measures for healthcare performance assessment: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Performance Measures (Writing Committee to Develop a Position Statement on Composite Measures). J Am Coll Cardiol 2010;55:1755–66.

This article has been copublished in *Circulation*.

JACC Vol. 55, No. 16, 2010
April 20, 2010:1755–66

1. Introduction

There is an increasing national focus on measuring, reporting, and rewarding the provider's quality of care based on objective measures of performance (1). The conceptual and methodological issues underlying the development of individual performance measures have previously been described (2), yet little has been written about the methods used to combine multiple individual metrics into summary or "composite" performance measures. The goals of this document are to 1) explore the uses of, and challenges associated with, composite performance measures in assessing healthcare quality; 2) discuss methods used in their creation; and 3) set forth some general principles for appropriate development, validation, application, and interpretation of composite measures.

2. Definition of Composite Performance Measures

A composite performance measure is the combination of 2 or more indicators into a single number to summarize multiple dimensions of provider performance and to facilitate comparisons. Composite measures are used in many areas, including assessment of intelligence such as intelligence quotient, product ratings (*Consumer Reports*), and stock market valuation (e.g., the Dow-Jones Industrial Average). The present document focuses specifically on the use of composite measures in healthcare provider performance assessment. These measures encapsulate relatively broad concepts, such as the overall quality of care, or may have a more focused perspective, such as adherence to a specific set of treatment guidelines. Examples include *US News & World Report* Annual Index of Hospital Quality (3), the Centers for Medicaid and Medicare Services Hospital Quality Incentive Demonstration composite quality score (4), and the Society of Thoracic Surgeons (STS) coronary artery bypass graft surgery composite performance measure (5).

In the present document, the term *provider* is used generically to refer to any of the various levels of the healthcare system whose performance may be evaluated, such as an individual practitioner, practice group, hospital, or healthcare plan. The term *developer* denotes the individual or group that develops the composite performance measure. The term *user* represents the intended consumers of this information, such as providers, payers, government regulators, or the public. The terms *measure*, *metric*, and *indicator* are used interchangeably in this document.

3. Rationale for Creating Composite Performance Measures

Composite performance measures have a variety of uses. Some important functions are:

Data reduction. A large and growing array of individual indicators makes it possible for users to become overloaded with data. A composite measure reduces the information burden by distilling the available indicators into a simple summary.

Scope expansion. The information in a composite measure is highly condensed, making it feasible to track a broader range of metrics than would be possible otherwise. Composite measures have been described as a tool for making provider assessments more comprehensive (1).

Provider performance valuation. Performance indicators are used for various decisions about providers, including the allocation of pay-for-performance incentives, designation of preferred provider status, and assignment of letter grades and star rating categories. If a decision is to be based on multiple indicators instead of a single indicator, a method of translating several variables into a single decision is needed. Composite measures serve this function by assigning providers to 1 position on a scale of better-to-worse performance.

4. Challenges Associated With Composite Performance Measures

Composite performance measures have many practical advantages, but numerous challenges are also associated with their implementation. Individual performance measures reflect specific aspects of a provider's quality of care, but this detailed information can be lost within a single summary measure. For example, a provider with an intermediate overall composite score rating relative to his or her peers may have had an intermediate performance on all of the measures or excellent performance on a select few and below-average performance on others. Distinguishing between these 2 scenarios requires knowledge of the provider's performance on each of the component measures. Likewise, as aggregate scores, composite measures may not provide clinicians or policymakers with clear actionable information from which to target or prioritize specific quality-

improvement efforts. The latter may be apparent only by examining individual performance measures.

Understandably, composite performance measures can also be perceived as "black boxes" if the measures and methods used in their creation are not transparent and easily understood. Thus, the inputs and rules used in the creation of composite performance measures should be clearly stated. Even so, their impact on the actual behavior of a composite score may be difficult to anticipate. For instance, Example A is a composite score calculated by assigning equal weight to each of 3 components (structure, outcomes, and reputation). Although each component is weighted equally mathematically, the composite score is largely determined by just 1 of the 3 components because the variances or spread of scores for the individual measures are different.

In addition, the weighting of items may not always reflect the unique interests, values, and preferences of all intended users. Outcomes and resource utilization might be 2 domains of a "value" composite. But patients and payers may place different relative weights on these 2 performance domains (6). Furthermore, as aggregate scores, composite measures may not provide clinicians or policymakers with clear actionable information with which to target or prioritize specific quality-improvement efforts. The latter may be apparent only by examining individual performance measures.

5. Development of Composite Performance Measures

The validity and usefulness of a composite performance measure are dependent on the quality of the individual performance measures they are based on, as well as the robustness of the underlying methodology used to combine these elements. The American College of Cardiology Foundation/American Heart Association have previously outlined important criteria for selection and creation of individual performance measures (2). The National Quality Forum has developed a framework to evaluate the validity and utility of composite measures (7). Creation of a composite performance measure involves (1) defining the purpose and theoretical framework for the composite performance measure (2), selecting individual component measures to be considered for inclusion in the composite performance measure, and (3) establishing rules for how these individual measures will be combined and weighted. In addition, the validity and operating characteristics of the composite performance measure must be carefully assessed (8).

6. Defining the Purpose and Conceptual Framework of Composite Performance Measures

The intended use of the composite performance measure should be clearly stated because it is the basis for selecting and combining indicators into a meaningful summary measure. Concepts such as quality of care are complex and may encompass a diverse array of possible meanings (1), depending on whether the consumer is a patient, provider, payer, or policymaker. Thus, when composite performance measures are being developed, it is important to state clearly which domains of quality are to be summarized. Furthermore, the interpretation and rationale for analyzing and reporting the chosen indicators should be delineated. The potential adverse consequences of inclusion (or exclusion) of given factors in a composite performance measure should also be considered. For example, to improve their scores, providers may focus their quality-improvement efforts on process measures that are included in a composite measure while ignoring care processes that are not included or not weighted heavily.

7. Criteria for Selecting and Evaluating Component Performance Measures

Types of measures. Performance measures are traditionally grouped into the categories of outcome, structure, and process (2). Structural measures describe components or characteristics of the care delivery system thought to have an influence on healthcare delivery or health-related outcomes (e.g., physical facilities, staff qualifications, case volume, or use of electronic health records). Process-of-care measures reflect what is actually done for a patient in terms of diagnosis, treatment, and other support services. Outcomes typically refer to clinical events such as mortality, morbidity, and quality of life. Beyond the triad of structure, process, and outcomes, metrics have been developed to measure patient satisfaction, appropriateness, and resources or costs associated with healthcare delivery.

Tradeoff between importance and feasibility. When choosing among indicators, there may be a tradeoff between what users would like to know about a provider and what can be measured reliably. Outcomes including mortality are often regarded as the gold standard in terms of relevance, but outcomes require risk adjustment and can be difficult to estimate, given the low frequency of adverse events in many situations (2,9). Structural measures such as case volume do not require risk adjustment and may be measured reliably, but these measures constitute indirect rather than direct indicators of quality. In other words, such measures are only interesting to the extent that they are surrogates for other events. Process measurement is feasible but requires careful consideration of patients with contraindications (2). Although statistical power and precision are potentially enhanced by incorporating process measures into assessments of providers, small sample sizes and ceiling effects (performance near 100%) are challenges for some process measures (10). To date, empirical studies have found only a modest association between nationally reported cardiovascular process measures and outcomes, including risk-adjusted mortality (11–13). Thus, a composite measure that has high accuracy for measuring adherence to evidence-based care practices may have limited accuracy for predicting an individual provider's outcomes.

Reliability. An important criterion for evaluating an individual indicator is the extent to which differences in the indicator between providers are explained by "true differences" (i.e., signal) versus "chance variation" (i.e., noise). Measures with a relatively high proportion of signal variance are said to be reliable and are useful because of their high power for discriminating among providers. In addition, large signal variation indicates a potential gap in quality and hence

an opportunity for improvement. Although highly variable measures are desirable from the standpoint of detecting statistically meaningful differences, the chosen end points must also be valid, collected reliably, and relevant to the intended users of the composite measure.

Considerations for process measure selection. Two considerations for process measure selection and evaluation are benefit and reliability. Ideally, the evidence for these should come from both clinical trials and observational studies (real-world settings), thus linking better performance on the measure with better patient outcomes. As mentioned previously, processes that are known to be effective from clinical trials and observational studies do not always exhibit a strong association with outcomes when measured at the provider level. Many of these evidence-based performance measures and their standardized definitions have previously been evaluated and endorsed by specialty societies, accrediting organizations, and federal agencies (2). Inclusion of endorsed measures is useful for ensuring a high level of evidence and promoting acceptance by stakeholders.

Internal consistency of indicators. When selecting indicators, items are frequently included or excluded on the basis of the extent to which the items correlate with each other. Generally, high internal consistency between indicators is regarded as evidence that they reflect a single underlying concept or domain (14). Indicators that correlate poorly with others may be questioned on the grounds that they appear to measure a different aspect or dimension. Although internal consistency is an important concept for some methods that have been developed in the fields of psychometrics and educational testing, the criterion of internal consistency has less relevance if the goal of the composite is to combine multiple distinct dimensions of quality as opposed to a single dimension. For example, timely reperfusion and use of secondary prevention discharge medications are both essential components of high-quality care for myocardial infarction (MI). In empirical studies, a weak or absent association between performance on these 2 domains would imply low internal consistency, yet both would still be needed for a complete assessment of a hospital's performance on MI care processes. When such dissimilar elements are grouped together into a composite, the ability to evaluate such composites based on standard psychometric criteria is limited.

8. Combining Items Into a Composite Performance Measure

Numerous methods have been used to combine individual measures or domains into composite scores, including linear combinations, latent trait modeling, opportunity scoring, and the creation of patient-level composite end points such as the all-or-none composite. These methods and related issues are discussed below.

8.1. Linear Combinations

Linear combinations are weighted sums of the form $w_1Y_1 + w_2Y_2 + \cdots + w_nY_n$, where Y_i denotes the value of the i-th indicator, n denotes the number of indicators, and w_i is the *weight* assigned to the i-th indicator. If the weight assigned to each indicator is the same ($w_i = 1/n$), then the items are weighted equally and the sum reduces to a simple average. Although linear combinations have the advantage of simplicity and transparency, the choice of weights is not straightforward. For example, when 3 individual scores are weighted equally within a linear combination, such a system is intuitive but does not account for potential differences in the validity, reliability, and importance of the different individual measures. For example, a score of 70% could be achieved by averaging 2 equally weighted scores of 70% or 1 score of 50% and 1 score of 90%. These may have far different implications for both consumers and providers.

Additionally, equal weighting may be undesirable if there is a considerable imbalance in the numbers of measures from different domains. For example, if 5 individual process measures are available for quality assessment of a disease but only 1 outcome is obtainable, then equal weighting of each indicator may result in a composite that is strongly reflective of care process rather than outcome performance. Even when equal weighting of measures is the composite methodology, the result of its application in real-world data may have unexpected results (10). Example B is a composite measure used by CMS. Although each individual performance measure is assigned equal weight mathematically, the overall composite is largely driven by the process measures.

Weights for a linear combination may sometimes come from other sources such as expert panels, literature, sample surveys, or discussions with stakeholders. In the STS composite performance measure (Example C), an expert panel of surgeons weighted individual items, relying on a review of the literature and a sample survey of the STS membership. In the absence of strong empirical evidence on which to base differential weighting, the developers of the measure weighted the 4 performance domains equally. However, subsequent evaluation of the score demonstrated that the relative impact of the 4 domains on the overall composite score was deemed to have face validity by the clinical experts.

Regardless of how weights are determined, it is important that they reflect the values and preferences of those using the measure. Healthcare quality measures may have a variety of user categories (e.g., patients, physicians, managed care organizations, employers). Because different stakeholders have different priorities, it is possible that more than 1 set of weights (e.g., multiple composites) will be needed to meet the needs of all potential users.

8.2. Regression-Based Composite Measures

If a certain outcome is regarded as a gold standard, the weighting of individual items may be determined empirically by optimizing the predictability of the gold standard end point. For example, acute MI performance measures might be weighted with the goal of predicting hospital-specific mortality rates. An appropriate statistical framework for this purpose is empirical Bayesian regression modeling (15). The weight assigned to each item is directly related to its reliability and the strength of its association with the gold standard end point. Although regression-based weighting may be appropriate for predicting specific end points of interest, such weighting may not be optimal for other objectives, such as motivating healthcare professionals to adhere to

specific treatment guidelines. Specifically, evidence-based process measures would not contribute to such a composite unless there was evidence of a strong empirical association between provider-level process performance and the criterion of interest.

8.3. Latent Trait Composite Measures

If multiple indicators are assumed to measure 1 dimension of care, the method of combining these indicators may be optimized by predicting 1 latent variable that reflects this dimension. This concept is the foundation of several psychometric methods, including versions of factor analysis, item-response modeling, and principal components analysis (16–19). To estimate a single latent trait, it is necessary that the various trait indicators pertain to a single dimension only rather than multiple distinct aspects of care. This distinction can be made from subject matter considerations and empirical testing. The assumption of unidimensionality may not be appropriate for a comprehensive measure that spans multiple domains of quality. For instance, a latent trait composite measure of acute MI might suppose that quality of care is consistent within a given hospital, although patients with acute MI receive care from multiple teams. It is quite possible that although the hospital's performance in acute process (e.g., provision of rapid reperfusion) may be excellent, its performance in provision of secondary prevention measures at hospital discharge may be poor, thus violating the single latent trait concept. However, it may be possible to identify clusters of correlated items so that the assumption of unidimensionality applies to all items in a single cluster. In this case, latent trait modeling may be used to combine items within clusters but not across clusters. The problem of how to weight fundamentally different dimensions of quality is beyond the scope of most psychometric methods.

Because psychometric methods may not be suitable in situations for which the unidimensionality assumption is untenable, in 1987 Feinstein introduced the "clinimetric" approach (20). This concept attempts to more specifically serve the aim of clinicians, "which is to choose and emphasize suitably the most important attributes to be included in the index, using multiple items which are not expected to be homogeneous because they indicate different aspects of a complex clinical phenomenon" (p 234) (21). The selection and weighting of items is based on deliberate judgment and is considered successful if the behavior of the composite score is consistent with the developer's intentions.

8.4. Opportunity Scoring

An opportunity-based score is an alternative to simple averaging often used for aggregating individual process measures. Opportunity scoring counts the number of times a given care process was actually performed (numerator), divided by the number of chances a provider had to give this care correctly (denominator). Unlike simple averaging, each item is implicitly weighted in proportion to the percentage of eligible patients, which may vary from provider to provider. This method has the advantage of increasing the number of observations per unit of measurement, consequently potentially increasing the stability of a composite estimate, particularly when the sample size for individual measures is not adequate. But these advantages can also be disadvantages because opportunity-based composite scores will inevitably be most influenced by the most common care processes, regardless of whether or not they are the most important methods.

8.5. All-or-None Scoring of Process Measures

In all-or-none scoring (also known as defect-free scoring), the patient is the unit of analysis. Only those patients who receive all indicated processes of care are counted as *successes*. Performance is defined by the proportion of patients receiving all of the specified care processes for which they were eligible. No credit is given for patients who receive some but not all required items. This method has been advocated on the grounds that it promotes a high standard of excellence (22). A counterargument is that dichotomizing outcomes as all or none wastes valuable information (23) and may weight common but less important processes more heavily than infrequent but important processes. All-or-none scoring gives the same credit for achieving none of 5 measures as it does for achieving 4 of 5 measures.

8.6. Any-or-None Scoring of Outcome Measures

Any-or-none scoring is an analogous method for combining outcomes rather than processes. In this method, a patient is counted as *failing* if he or she experiences at least 1 adverse outcome from a list of 2 or more adverse outcomes. Such end points are commonly used in randomized clinical trials of cardiovascular therapies when the analysis of mortality alone would require an extremely large sample size. As noted by many authors, such composite outcomes may be misleading if the component items occur with unequal frequency or vary in their importance (24–26). Any-or-none composite outcomes are particularly problematic when rare but important outcomes are mixed with common but relatively unimportant outcomes, because the composite is likely to be dominated by the outcome that occurs most frequently.

8.7. Incommensurable Measurement Scales

Performance measures are evaluated on various different scales and therefore must be transformed into a common scale before being combined. Some methods of rescaling include 1) dividing each measure by its standard deviation, 2) dividing each measure by its range, 3) assigning scores based on the provider's percentile ranking, and 4) assigning a specified number of points if the provider's score exceeds a certain threshold (e.g., above the mean or median) (8). Although different methods of rescaling may produce different results, these inconsistencies are often subtle, particularly when compared with the effect of no standardization at all (18,27). Example C illustrates rescaling used in the creation of the STS coronary artery bypass graft composite score methodology.

9. Statistical Models for Estimating Composite Performance Measures

Performance measures are susceptible to chance fluctuations, even when performance remains constant. Because these fluctuations are partly random, there is always some degree of uncertainty about a provider's true underlying performance. Statistical methods and models have been developed to

account for this source of uncertainty when estimating individual or composite measures used to compare provider performance.

Random effects modeling (also known as hierarchical modeling) is a commonly used statistical model for estimating provider performance. Unlike simple percentages and averages, a random effects model uses data from all analyzable providers to estimate performance measures for 1 provider. This "borrowing of information" across providers produces estimates with good statistical properties, including smaller standard errors than conventional estimates. The random effects model–based estimate can be viewed as a weighted average of a provider's actual score for a singular measure and the overall average score for all providers in the analysis database. The model weights an individual provider's own data more heavily when the denominator is large enough to be reliable and weights the overall average score more heavily when the provider's denominator is too small to support a reliable performance estimate.

Random effects models are commonly used to analyze a single outcome, such as mortality. But extensions to multiple outcomes are straightforward and have been used to measure the quality of adult cardiac surgery (15,27). When multiple outcomes are analyzed simultaneously, information is not only borrowed across providers but across outcomes. For example, information on a provider's mortality rate may help improve estimation of performance with respect to other nonfatal outcomes (e.g., stroke rate, infection rate) that are statistically associated with mortality.

Although model-based estimates are useful for estimating population parameters and predicting future outcomes, they are less transparent than simple unadjusted percentages and raw averages. Because random effects estimates are a combination of an individual provider's own data and data borrowed from other providers, the resulting estimate does not provide a simple summary of what actually occurred in the sample of patients treated by a single provider. Moreover, as a result of the "shrinkage" property of random effects estimates (i.e., moving a provider's measured performance closer to the overall population mean), for providers with a very small number of patients, the estimates tend to be close to the overall provider average, regardless of actual results. Although model-based approaches provide more reliable estimates, the construct validity of this approach may be more problematic for some stakeholders because a provider's estimate depends partly on the performance of other providers.

Although hierarchical model–based estimates are often preferred on the grounds that they are less variable than conventional estimates, they are still prone to large errors in the presence of small sample sizes, so these estimates should be accompanied by a measure of precision. In some contexts, such as public reporting, interval estimates, which consist of a range of plausible estimates, are more appropriate.

10. Reporting Issues Associated With Composite Performance Measures

10.1. Rescaling and Categorization

When a composite performance measure is formed by averaging items with different measurement scales (e.g., survival rates and staffing ratios), the scale of the resulting average composite will not be inherently meaningful or interpretable. In such cases, rescaling the composite to lie between 1 and 100 (or another familiar range) may enhance the perceived interpretability of the composite. When rescaling a composite measure, consideration should be given to the clinical and statistical significance of the observed variation in the component items. The scale should not create the false impression of large differences between providers if in fact these differences are negligible. In some cases, converting the composite performance measure to a small number of categories (e.g., 1 to 3 "stars") may enhance communication and graphic presentation. As with rescaling, in categorization there is a risk of exaggerating true differences between providers if the actual differences are statistically or clinically insignificant.

10.2. Measures of Precision

Reporting of composite measures should be accompanied by an estimate of the uncertainty caused by random statistical fluctuations. Ignoring this source of uncertainty may lead users to overestimate the reliability of the data and to draw false conclusions about performance. When comparing providers with each other or with a benchmark, reports should indicate whether the observed differences are within the bounds of normal sampling variation.

10.3. Reporting of Component Items

In addition to reporting the overall composite performance measure, the composite score should be capable of deconstruction. A composite score should be broken down into scores for each component domain or individual measure. Providing deconstructed detailed data are critical for 2 main reasons: 1) detailed data allow providers to focus their quality improvement efforts and 2) because the weighting of indicators may not optimally reflect an individual user's preferences, detailed data allow interested users to reconstruct and adjust the composite measure based on their own values or purpose. It is also suggested that composite performance measures include some indication of the number of cases in the numerator and denominator if relevant.

10.4. Dealing With Missing Data

The performance of a composite measure applied to real-world settings can be compromised by missing data. First and foremost, developers should have a defensible strategy for managing missing data. Case-wise deletion of records with missing data can be problematic because information is wasted when some but not all of the required information is missing. This may cause bias if data are not randomly missing. Imputation of missing data is generally preferred to case-wise deletion of records but may cause bias if the imputing method is not statistically rigorous. When feasible, techniques such as resampling and multiple imputation should be used. These techniques appropriately account for the uncertainty associated with missing data and produce valid point estimates and confidence intervals.

11. Evaluation of Composite Performance Measures

Although the selection of methodology and weights for a composite performance measure may be subjective, developers of composite measures must fully explore the implications of the chosen approach and verify that the resulting composite measure behaves in a manner that is intuitively acceptable and consistent with the intended purpose and interpretation.

11.1. Exploring Behavior of Composite Performance Measures

One way to evaluate the behavior of a composite performance measure is to determine the amount of improvement needed in 1 variable to offset worsening in another variable. A second method of elucidating the implications of weights involves calculating the correlation coefficient between each individual item and the overall composite. A very low correlation between 1 item and the composite suggest that the weight assigned to the item may be too low to substantially influence the overall composite. An exceptionally strong correlation between 1 item and a composite suggests that much of the variation in the composite can be explained by a single item.

11.2. Sensitivity Analysis

Often more than 1 method of creating a composite measure is reasonable and consistent with the intended purpose of the composite measure. In such cases, developers should explore a variety of these methods and document whether conclusions about provider performance differ substantially depending on the choice of method (6,8,10,28). If different versions of the composite measure are based on different weights and highly correlated, the choice between alternative weighting methods has relatively little practical importance. As a result, this issue can be safely ignored. On the other hand, if different weights produce widely divergent results, then this uncertainty should be considered by composite measure users, and conclusions about performance may be tempered.

12. Assessing the Validity of Composite Performance Measures

The validity of a composite performance measure depends on the purpose for which it is applied. For example, a composite performance measure that is composed of several process measures may have excellent validity for summarizing a hospital's performance on process measures but poor validity for predicting a hospital's outcomes. Documenting the intended purpose and interpretation of the composite score will help ensure that the composite score is evaluated for its intended purpose.

12.1. Construct Validity

A composite performance measure is said to have construct validity if it truly measures what it purports to measure. An important aspect of construct validity is content validity, which is defined as the presence of all important content aspects in the available indicators. If crucial items are missing, then the concept may need to be reformulated to reflect the available indicators.

12.2. Face Validity

A composite performance measure should also be tested for face validity, which is acceptance by stakeholders that the measure is useful and valid. Face validity is partly based on agreement that a composite measure has good construct validity. Acceptance of a composite performance measure is also enhanced by use of methods and a reporting format that are easily understood.

12.3. Criterion Validity

Criterion validity implies that the composite score correlates highly with an end point that may be regarded as a gold-standard method of measuring the concept of interest. If the goal of the composite performance measure is to identify hospitals with excellent outcomes, then criterion validity might be assessed by comparing predictions based on the composite performance measure with observed outcomes, including mortality. Yet criterion validity is difficult to establish for multidimensional composite scores, primarily because the concept being measured is ill-defined and thus lacks a criterion standard.

12.4. Precision/Reliability

A measure is said to be precise or reliable if the amount of variation in the composite that is caused by random statistical fluctuations is small, relative to variation caused by true differences between units. The acceptable level of precision depends on the context in which the measure is used. Measures dominated by chance variation may be harmful in that they can result in unfair conclusions about providers and mislead consumers.

13. Recommendations

The writing committee considers the following as necessary to the development and implementation of any composite performance measure used for public reporting or other forms of accountability. When used solely for the purposes of internal quality improvement, however, it may be reasonable to use criteria that are somewhat less strict.

1. The intended audience and purpose of the composite performance measure should be explicitly stated. This will provide the basis for selection of individual indicators and provide direction for methods of aggregation and weighting.
2. Decisions about what to measure should be based on the clinical importance associated with important patient outcomes and the reliability of individual performance measures. Assumptions underlying the choice of measures should be documented. Ideally, component measures should be tested before they are included in a composite performance measure.
3. Each individual component should be precisely defined to ensure consistent application in different settings.
4. The description of the methods used for weighting and combining individual measures into a composite performance measure should be transparent. The strengths and limitations of the selected method should be considered and discussed.

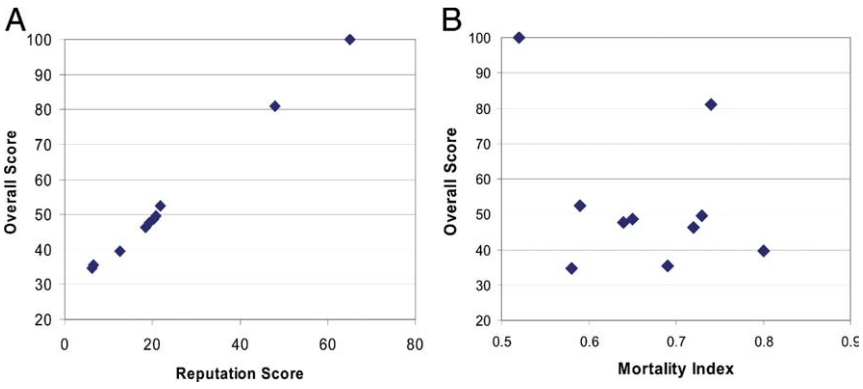

Figure 1. Index of hospital quality scores for the top 10 ranked heart and heart surgery hospitals in 2008 relative to their reputation and mortality scores.

5. Developers should explore a variety of alternative methods for combining measures and should document whether conclusions about provider performance differ with use of alternative methods.
6. Empirical testing should be performed to assess the properties of a composite measure score and to understand what is being measured. The considerations of validity and reliability typically are viewed as essential elements for determining the quality of any test.
7. Reporting of composite performance measures should be accompanied by detailed reporting of individual domains and components.
8. Reporting of composite performance measures should include a measure of the degree of uncertainty surrounding composite estimates for providers.
9. Composite performance measures based on scientific evidence must be reevaluated as that evidence changes. Additionally, the operating characteristics of a given composite performance measure should be periodically reevaluated for reliability and validity to ensure that they have not changed over time.

Example A. *US News & World Report* Index of Hospital Quality for Heart and Heart Surgery. The *US News & World Report* Index of Hospital Quality is the basis for evaluating hospitals in the magazine's annual "Best Hospitals" report (29). The Index of Hospital Quality for heart and heart surgery is a linear combination of 3 equally weighted components: reputation, risk-adjusted mortality, and structure. Although the 3 components are weighted equally, the recognition scores appear to be the most variable (ranging from 0 to 65 in 2008). Figure 1 shows the strong linear correlation between a hospital's reputation score and its overall Index of Hospital Quality score, suggesting that the reputation component is highly influential in determining the relative rank ordering of these 10 hospitals (Panel A). In comparison, the Mortality Index (Panel B) appears to have much less influence.

Example B. The Centers for Medicaid and Medicare Services Hospital Quality Incentive Demonstration. In the Centers for Medicaid and Medicare Services Hospital Quality Incentive Demonstration Project, a composite measure was used for assigning financial bonuses and penalties to hospitals in a pilot pay-for-performance project (4). For coronary artery bypass graft surgery, the composite quality score (CQS) was based on an equally weighted combination of 7 measures (4 process measures and 3 outcome measures). Thus, at first glance it might be thought that the overall composite would assign 4/7 of the weight to processes and 3/7 of the weight to outcomes. Indeed, this is true, based on a mathematical description of the aggregation method. But despite this equal mathematical weighting, the actual publicly reported data suggest that the CQS was more heavily influenced by process measures than would have been expected by the apparent 4:3 weighting. This point is illustrated in Figure 2, which depicts data from the top 50% of hospitals in year 1 of the project. As seen in Panel A, there is a near perfect increasing relationship between a given hospital's performance on process measures and a given hospital's decile ranking (based on overall CQS). In contrast, as seen in Panel B, there is almost no relationship between a given hospital's performance on outcome measures and a given hospital's ranking. The explanation is that compared with the process measures, the outcomes

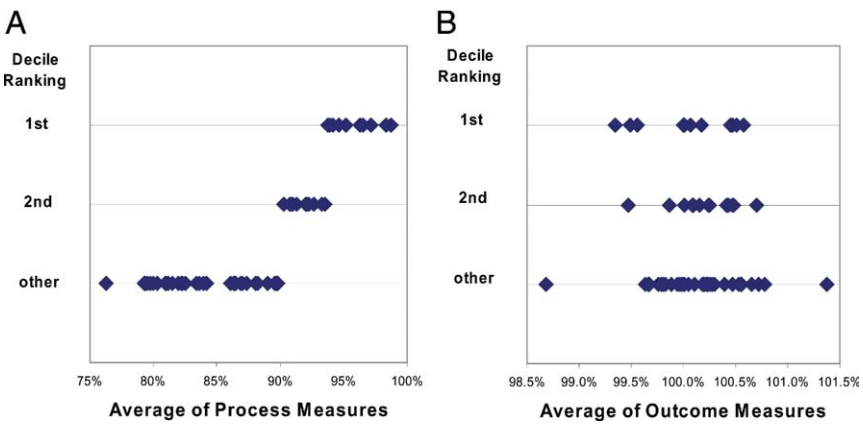

Figure 2. Data from top 50% of hospitals in year 1 of the Centers for Medicaid and Medicare Services Hospital Quality Incentive Demonstration reporting process measures versus outcome measures.

measures have a relatively small standard deviation. When items with a small standard deviation are averaged with items with a large deviation, items with the large standard deviation tend to dominate the average.

Example C. Society of Thoracic Surgeons Composite Measurement Methodology for Adult Cardiac Surgery. The STS Composite Measurement Methodology, published in 2005, is a linear combination of scores from 4 separate domains: risk-adjusted mortality, risk-adjusted morbidity, an intraoperative surgical process measure (use of the internal mammary artery [IMA]), and several adjuvant perioperative medications (5). To create the composite, the 4 subcomponent domain scores were first standardized by dividing their respective standard deviations and then averaged together with equal weighting of each component score.

The consequences of this standardization and weighting approach were explored analytically and empirically. From an analytic perspective, assigning equal weight to the standardized domain scores was shown to be mathematically equivalent to assigning unequal weights to the original domain scores, which were not standardized. The "weight" assigned to a given score was proportional to the reciprocal of its standard deviation. Thus, this approach reduced the relative weighting of items with larger standard deviations and increased the relative weighting of items with smaller standard deviations.

Pilot studies showed that the risk-adjusted mortality domain score had the smallest standard deviation, followed by risk-adjusted morbidity, IMA, and medications. After dividing each domain score by its standard deviation, it was shown that a 1-percentage point improvement in a provider's risk-adjusted survival rate would increase the overall composite score by the same amount as an 8.4-percentage point improvement in the morbidity rate, an 11.6-percentage point improvement in IMA usage, or a 28.6-percentage point improvement in medication usage. Additional empirical studies showed that no single item dominated the composite and that all 4 domains contributed substantial statistical variation. This behavior was deemed reasonable by the STS expert panel that developed the composite. The approach was endorsed by the STS Executive Committee before its public use.

Staff

American College of Cardiology Foundation

John C. Lewin, MD, Chief Executive Officer

Charlene L. May, Senior Director, Science and Clinical Policy

Melanie Shahriary, RN, BSN, Associate Director, Performance Measurement Policy

Erin A. Barrett, Senior Specialist, Science and Clinical Policy

Jensen S. Chiu, MHA, Specialist, Clinical Performance Measures

American Heart Association

Nancy Brown, Chief Executive Officer

Rose Marie Robertson, MD, FACC, FAHA, Chief Science Officer

Gayle R. Whitman, PhD, RN, FAHA, FAAN, Senior Vice President, Office of Science Operations

Nereida Crawford, MPH, Science and Medicine Advisor

Appendix 1. Author Relationships With Industry and Other Entities—ACCF/AHA 2010 Position Statement on Composite Measures for Healthcare Performance Assessment

Committee Member	Consultant	Speaker	Ownership/Partnership/ Principal	Research	Institutional, Organizational, or Other Financial Benefit	Expert Witness
Eric D. Peterson, *Chair*	None	None	None	• Bristol-Myers Squibb/sanofi-aventis* • Merck* • Schering Plough*	None	None
Elizabeth R. DeLong	None	None	None	None	None	None
Frederick A. Masoudi	• Amgen • Takeda • United Healthcare	None	• sanofi-aventis	• Amgen*	None	None
Sean M. O'Brien	None	None	None	None	None	None
Pamela N. Peterson	None	None	None	None	None	None
John S. Rumsfeld	• United Healthcare	None	None	None	None	None
David M. Shahian	None	None	None	None	None	None
Richard E. Shaw	None	None	None	None	None	None

This table represents the relevant relationships of authors with industry and other entities that were disclosed at the time of peer review. It does not necessarily reflect relationships with industry at the time of publication. A person is deemed to have a significant interest in a business if the interest represents ownership of 5% or more of the voting stock or share of the business entity, or ownership of $10 000 or more of the fair market value of the business entity, or if funds received by the person from the business entity exceed 5% of the person's gross income for the previous year. A relationship is considered to be modest if it is less than significant under the preceding definition. Relationships in this table are modest unless otherwise noted.

*Significant (greater than $10 000) relationship.

Appendix 2. Peer Reviewer Relationships With Industry and Other Entities—ACCF/AHA 2010 Position Statement on Composite Measures for Healthcare Performance Assessment

Peer Reviewer	Representation	Consultant	Speaker	Ownership/Partnership/ Principal	Research	Institutional, Organizational, or Other Financial Benefit	Expert Witness
Eric Bates	Official Reviewer—ACCF Board of Trustees	None	None	None	None	• ACCF Board of Trustees	None
Lee A. Fleisher	Official Reviewer—AHA	None	None	None	None	• Chair, American Society of Anesthesiologists Committee on Performance and Outcomes Measures • Member, Board of Directors of Accreditation Association for Ambulatory Health Care Quality Institute	None
Graham Nichol	Official Reviewer—AHA	• Innercool Therapies • Northfield Laboratories* • Paracor Medical Inc • Radiant Medical	None	None	• CIHR* • EMcools* • Laerdal Foundation* • Medic One Foundation* • Lifebridge* • NIH/National Heart, Lung, and Blood Institute*	• Channing-Bete* • Laerdal Inc. • Medic One Foundation • Physio-Control*	None
Michael Rossi	Official Reviewer—ACCF Board of Governors	None	None	None	None	None	None
John Brush	Content Reviewer—ACCF Board of Governors	None	None	None	None	None	None
Hector Bueno	Content Reviewer—AHA Council on Quality of Care and Outcomes Research	• Bayer	• Bristol-Myers Squibb	None	• Pfizer*	• Fondo de Investigaciones Sanitarias, Ministry of Health, Spain*	None
Pamela S. Douglas	Content Reviewer—ACCF Task Force on Appropriate Use Criteria	• Abingworth • BG Medicine • Medscape/WebMD • Northpoint Domain • Ortho Diagnostics • Pappas Ventures • Primera • Roche • Translational Research in Oncology (DSMC) • Xceed Molecular	None	• CardioDX • Expression Analysis • Northpoint Domain • Xceed Molecular	• Abingworth • Abiomed • Atritech • BG Medicine • CardioDX • Edwards • Expression Analysis • Medscape/WebMD • Northpoint Domain • Ortho Diagnostics • Osiris • Pappas Ventures • Primera • Roche • Viacor • Xceed Molecular	None	None
Joseph P. Drozda	Content Reviewer—ACC Clinical Quality Steering Committee	None	None	None	None	None	None
Gregg C. Fonarow	Content Reviewer—AHA Get With The Guidelines	• GlaxoSmithKline* • Novartis* • Medtronic* • Pfizer* • Merck* • sanofi-aventis*	None	None	• National Heart, Lung, and Blood Institute*	None	None
Jeffrey Geppert	Content Reviewer—statistical expertise	None	None	None	None	None	None
Robert C. Guyton	Content Reviewer—ACC/AHA Task Force on Practice Guidelines	None	None	None	• ACCF Board of Trustees	None	None
Robert C. Hendel	Content Reviewer—Task Force on Appropriate Use Criteria	• Astellas Pharma • PGx Health Care	• Astellas Pharma	None	None	None	None

(*Continued*)

Appendix 2. Continued

Peer Reviewer	Representation	Consultant	Speaker	Ownership/Partnership/ Principal	Research	Institutional, Organizational, or Other Financial Benefit	Expert Witness
Rosemarie B. King	Content Reviewer—AHA Stroke Council Quality and Outcomes Committee	None	None	None	• Principal Investigator, NIH NINR	None	None
David Nilasena	Content Reviewer—Centers for Medicare and Medicaid Services	None	None	None	None	• Spouse employed by Baylor Health System, Dallas, TX	None
Richard Nishimura	Content Reviewer—ACC/ AHA Task Force on Practice Guidelines	None	None	None	• ACCF Board of Trustees	None	None
Harry C. Odabashian, Jr	Content Reviewer—ACCF Board of Governors	None	• Merck • Schering Plough • Novartis	None	• Investigator, Improve IT	None	None
Ileana L. Piña	Content Reviewer—AHA Council on Quality of Care and Outcomes Research	• Food and Drug Administration	• AstraZeneca • Merck • Novartis • sanofi-aventis • Solvay	None	• NIH	None	None
Stephen Wallach	Content Reviewer—ACCF Board of Governors	None	None	None	• AstraZeneca-Charm • Roche	None	• Worker's compensation cases • Social Security reviews

ACCF indicates American College of Cardiology Foundation; AHA, American Heart Association; CIHR, Canadian Institutes of Health Research; DSMC, Data Safety Monitoring Committee; NIH, National Institutes of Health; and NINR, National Institute of Nursing Research.

This table represents the relevant relationships of reviewers with industry and other entities that were disclosed at the time of peer review. It does not necessarily reflect relationships with industry at the time of publication. Names are listed in alphabetical order within each category of review. Participation in the peer review process does not necessarily imply endorsement of this document. A person is deemed to have a significant interest in a business if the interest represents ownership of 5% or more of the voting stock or share of the business entity, or ownership of $10 000 or more of the fair market value of the business entity; or if funds received by the person from the business entity exceed 5% of the person's gross income for the previous year. A relationship is considered to be modest if it is less than significant under the preceding definition. Relationships in this table are modest unless otherwise noted.

*Significant (greater than $10 000) relationship.

References

1. Institute of Medicine. Performance measurement: accelerating improvement. December 2005. Available at: http://www.iom.edu/en/Reports/2005/Performance-Measurement-Accelerating-Improvement.aspx. Accessed December 7, 2009.
2. Spertus JA, Eagle KA, Krumholz HM, et al. American College of Cardiology and American Heart Association methodology for the selection and creation of performance measures for quantifying the quality of cardiovascular care. J Am Coll Cardiol. 2005;45:1147–56.
3. US News & World Report. 2008 Annual index of hospital quality. Available at: http://health.usnews.com/sections/health/best-hospitals/index.html. Accessed June 23, 2009.
4. Premier Inc. CMS/Premier Hospital Quality Incentive Demonstration (HQID): model hospital value-based purchasing program continues to improve patient outcomes. Available at: http://www.premierinc.com/quality-safety/tools-services/p4p/hqi/index.jsp. Accessed December 7, 2009.
5. Society of Thoracic Surgeons. The STS composite quality measurement methodology: executive summary. Available at: http://www.sts.org/documents/pdf/QualityExecutiveSummary-Final.pdf. Accessed May 27, 2009.
6. Rosenthal MB, Landrum MB, Meara E, et al. Using performance data to identify preferred hospitals. Health Serv Res. 2007;42:2109–19.
7. National Quality Forum. Composite measure evaluation framework and national voluntary consensus standards for mortality and safety—composite measures: a consensus report. Available at: http://qualityforum.org/Publications/Composite_Measures.aspx. Accessed December 2, 2009.
8. Nardo M, Saisana M, Saltelli A, et al. Handbook on constructing composite indicators: methodology and user guide. Organisation for Economic Co-operation and Development. Available at: http://www.olis.oecd.org/olis/2005doc.nsf/LinkTo/NT00002E4E/$FILE/JT00188147.PDF. Accessed February 6, 2009.
9. Dimick JB, Welch HG, Birkmeyer JD. Surgical mortality as an indicator of hospital quality: the problem with small sample size. JAMA. 2004;292:847–51.
10. O'Brien SM, DeLong ER, Dokholyan RS, et al. Exploring the behavior of hospital composite performance measures: an example from coronary artery bypass surgery. Circulation. 2007;116:2969–75.
11. Bradley EH, Herrin J, Elbel B, et al. Hospital quality for acute myocardial infarction: correlation among process measures and relationship with short-term mortality. JAMA. 2006;296:72–8.
12. Williams SC, Koss RG, Morton DJ, Loeb JM. Performance of top-ranked heart care hospitals on evidence-based process measures. Circulation. 2006;114:558–64.
13. Werner RM, Bradlow ET. Relationship between Medicare's hospital compare performance measures and mortality rates. JAMA. 2006;296:2694–702.
14. Cronbach L. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16:297–334.
15. Staiger DO, Dimick JB, Baser O, et al. Empirically derived composite measures of surgical performance. Med Care. 2009;47:226–33.
16. Nunnally JC, Bernstein IH. Psychometric Theory. 3rd ed. New York, NY: McGraw-Hill; 1994.
17. Crocker LM, Algina J. Introduction to Classical and Modern Test Theory. Mason, OH: Wadsworth Pub Co; 2006.

18. DeVellis RF. Scale Development: Theory and Applications. Thousand Oaks, Calif: Sage Publications Inc; 2003.
19. Kline P. Handbook of Psychological Testing. 2nd ed. New York, NY: Routledge; 1999.
20. Feinstein AR. Clinimetric perspectives. J Chronic Dis. 1987;40:635–40.
21. Fayers PM, Hand DJ. Causal variables, indicator variables and measurement scales: an example from quality of life. J Roy Stat Soc. 2002;165: 233–61.
22. Nolan T, Berwick DM. All-or-none measurement raises the bar on performance. JAMA. 2006;295:1168–70.
23. Hayward RA. All-or-nothing treatment targets make bad performance measures. Am J Manag Care. 2007;13:126–8.
24. Chi GY. Some issues with composite endpoints in clinical trials. Fundam Clin Pharmacol. 2005;19:609–19.
25. Ferreira-Gonzalez I, Busse JW, Heels-Ansdell D, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomized controlled trials. BMJ. 2007;334:786.
26. Freemantle N, Calvert M, Wood J, et al. Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA. 2003;289:2554–9.
27. O'Brien SM, Shahian DM, DeLong ER, et al. Quality measurement in adult cardiac surgery: part 2–Statistical considerations in composite measure scoring and provider rating. Ann Thorac Surg. 2007;83 Suppl 4:S13–26.
28. Jacobs R, Goddard M, Smith PC. How robust are hospital ranks based on composite performance measures? Med Care. 2005;43:1177–84.
29. US News & World Report. America's Best Hospitals. Available at: http://health.usnews.com/sections/health/best-hospitals/index.html. Accessed May 17, 2009.

KEY WORDS: ACCF/AHA Performance Measures ■ performance measures ■ composite measures ■ quality indicators ■ quality measurement.