ACCF/AHA New Insights Into the Methodology of Performance Measurement

A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Performance Measures

WRITING COMMITTEE MEMBERS

John A. Spertus, MD, MPH, FACC, FAHA, Chair; Robert O. Bonow, MD, MACC, FAHA; Paul Chan, MD, MSc, FACC; George A. Diamond, MD, FACC; Joseph P. Drozda, Jr, MD, FACC; Sanjay Kaul, MBBS, FACC; Harlan M. Krumholz, MD, SM, FACC, FAHA; Frederick A. Masoudi, MD, MSPH, FACC, FAHA; Sharon-Lise T. Normand, PhD, MSc, FACC, FAHA; Eric D. Peterson, MD, MPH, FACC, FAHA; Martha J. Radford, MD, FACC, FAHA; John S. Rumsfeld, MD, PhD, FACC, FAHA

ACCF/AHA TASK FORCE ON PERFORMANCE MEASURES

Frederick A. Masoudi, MD, MSPH, FACC, FAHA, Chair; Elizabeth DeLong, PhD; John P. Erwin III, MD, FACC; David C. Goff, Jr, MD, PhD, FAHA; Kathleen Grady, PhD, RN, FAHA, FAAN; Lee A. Green, MD, MPH; Paul A. Heidenreich, MD, FACC; Kathy J. Jenkins, MD, MPH, FACC; Ann R. Loth, RN, MS, CNS; Eric D. Peterson, MD, MPH, FACC, FAHA; David M. Shahian, MD, FACC

Downloaded from http://circ.ahajournals.org/ by guest on March 1, 2017

Table of Contents

1. Introduction 2091
2. New Insights Into the Selection of Possible Performance Measures 2092
   2.1. Abbreviations Used Throughout the Report . .2092
   2.2. Strength of Evidence 2093
   2.3. Costs and Performance Measures 2094
3. New Insights Into the Construction of Performance Measures 2095
   3.1. Use of Exceptions in Performance Measures. .2095
   3.2. Considerations in the Use of Outcomes Measures. .2096
   3.3. Numbers of Measures 2097
   3.4. Modification and Retirement of Measures . . .2098
4. New Insights Into the Implementation of Performance Measures 2098
5. New Insights Into the Analysis and Interpretation of Performance Measures 2099
   5.1. Composite Measures. 2099
   5.2. Attribution 2100
6. Conclusion. 2102
Appendix A. Author Relationships With Industry and Other Entities–ACCF/AHA New Insights Into the Methodology of Performance Measurement 2102
Appendix B. Peer Reviewer Relationships With Industry and Other Entities–ACCF/AHA New Insights Into the Methodology of Performance Measurement 2103

Introduction

Since the publication of the initial American College of Cardiology (ACC)/American Heart Association (AHA) Methodology for the Selection and Creation of Performance

This document was approved by the American College of Cardiology Foundation Board of Trustees in June 2010 and the American Heart Association Science Advisory and Coordinating Committee in June 2010.

The American Heart Association requests that this document be cited as follows: Spertus JA, Bonow RO, Chan P, Diamond GA, Drozda JP Jr, Kaul S, Krumholz HM, Masoudi FA, Normand S-LT, Peterson ED, Radford MJ, Rumsfeld JS. ACCF/AHA new insights into the methodology of performance measurement: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Performance Measures. *Circulation*. 2010;122:2091–2106.

This article has been copublished in the *Journal of the American College of Cardiology*.

The online-only Data Supplement is available at http://circ.ahajournals.org/cgi/content/full/CIR.0b013e3181f7d78c/DC1.

Copies: This document is available on the World Wide Web sites of the American College of Cardiology (www.cardiosource.org) and the American Heart Association (my.americanheart.org). A copy of the document is also available at http://www.americanheart.org/presenter.jhtml?identifier=3003999 by selecting either the "topic list" link or the "chronological list" link (No. KB-0075). To purchase additional reprints, call 843-216-2533 or e-mail kelle.ramsay@wolterskluwer.com.

Expert peer review of AHA Scientific Statements is conducted at the AHA National Center. For more on AHA statements and guidelines development, visit http://www.americanheart.org/presenter.jhtml?identifier=3023366.

Permissions: Multiple copies, modification, alteration, enhancement, and/or distribution of this document are not permitted without the express permission of the American Heart Association. Instructions for obtaining permission are located at http://www.americanheart.org/presenter.jhtml?identifier=4431. A link to the "Permission Request Form" appears on the right side of the page.

(*Circulation*. 2010;122:2091-2106.)

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

***Circulation* is available at http://circ.ahajournals.org** **DOI: 10.1161/CIR.0b013e3181f7d78c**

Measures,[1] there has been an explosion in the development and application of performance measures. Although initially envisioned as a means for physician-led quality-improvement efforts, performance measures have been primarily used as tools for accountability and performance-based reimbursement instead. Given the centrality of and experience with performance measures for quantifying healthcare quality, the American College of Cardiology Foundation (ACCF)/AHA Task Force on Performance Measures sought to update its methodology so that ongoing efforts to measure performance could benefit from emerging insights. The original methodology, proposed in 2005,[1] remains the foundation for developing process performance measures. The principal recommendations of the 2005 report are summarized in Table 1. The 2010 report does not address detailed issues of analysis,[3] pay for performance,[4] or nonfinancial rewards for better performance[5] because these topics have been addressed in other statements. The focus of the 2010 report is to provide a state-of-the-art perspective on the construction, collection, and emerging directions of performance measurement as a means to improve healthcare quality.

It is important to contextualize the creation of performance measures in light of the new and existing metrics for evaluating health care. Clinical practice guidelines, such as those developed by the ACCF/AHA Task Force on Practice Guidelines,[6] remain the primary activity through which the rapidly evolving clinical literature is evaluated and synthesized. These guidelines provide an evidentiary review and recommendations that support patient care. Through direct linkages to scientific evidence, guidelines are the foundation underpinning most efforts to improve the quality of care. Appropriate use criteria identify common, prototypical patient subgroups for which expert clinicians, using available evidence from the guidelines and medical literature, assess the benefits and risks of a test or procedure on patient outcomes. Appropriate use criteria are a framework with which to examine the rational use of diagnostic and therapeutic procedures to support a more efficient use of medical resources, a major goal of the US healthcare system.[7] Such criteria arose from the observation that use of procedures varied across the nation, and it was not clear whether some patients were undertreated or others overtreated. The primary goals of appropriate use criteria are to identify overuse (patients who received unnecessary therapy) to improve the safety and cost-effectiveness of care.

Table 1. ACCF/AHA Attributes of Performance Measures

Choosing Performance Measures	
Selection Factors	Considerations
Adherence to the potential performance measure results in meaningful improvements in clinically important outcomes	Evidence-based trials, strong clinical practice guideline recommendations for (Class I, Level of Evidence: A) or against (Class III, Level of Evidence: A) the measure
Broad sampling from multiple domains associated with the process of medical care (see Figure 1)	Measures should be distributed across the domains of diagnosis, patient education, treatment, patient self-management, and serial monitoring of success of treatment
Attributes of Selected Measures	
Measure Characteristics	Relevant Attributes
Useful in improving patient outcomes	Interpretable Actionable
Measure design	Denominator precisely defined
	Numerator precisely defined
	Established types of validity
	• Face
	• Content
	• Construct
	Established reliability
Measure implementation	Feasibility
	• Reasonable effort
	• Reasonable cost
	• Reasonable time period for collection
Overall assessment by Performance Measures Writing Group	Overall assessment of measure by explicit, predefined criteria for inclusion in measurement set

Adapted from Normand SL et al.[2]

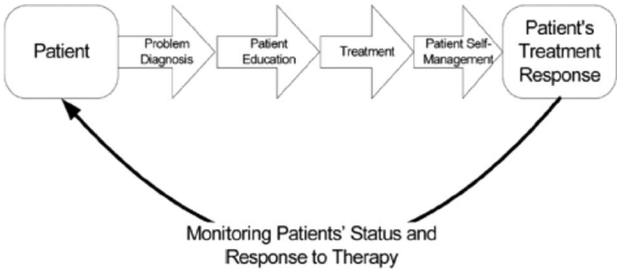

Figure 1. An overview of the steps in providing care by domain. Reprinted from Spertus et al.[1]

Performance measures that articulate discrete processes of care, as opposed to structural aspects of care or outcomes, are distinctly different from both clinical practice guidelines and appropriate use criteria because they represent a subset of the clinical guidelines for which the evidence is sufficiently strong: typically where the highest-quality evidence of benefit unequivocally exceeds risk (Class I recommendation, Level of Evidence: A),[8] failure to provide the therapy to an eligible patient meaningfully reduces the likelihood that the patient will experience the best possible outcome. In this report, the writing committee, commissioned by the Task Force on Performance Measures, discusses new insights into the selection of performance measures, including the strength of evidence needed to consider creating a performance measure from a clinical guideline and the role of costs in considering selection of a performance measure for an expensive technology. The committee then describes new insights into the construction of performance measures (creation of exclusions, use of outcomes measures, numbers of measures, and modification/retirement of measures). Finally, the implementation and analysis of performance measures are discussed, including the use of composite measures and attribution as the foundation for a continually growing and improving healthcare system.

2. New Insights Into the Selection of Possible Performance Measures

2.1. Abbreviations Used Throughout the Report

AMI, acute myocardial infarction
CMS, Centers for Medicare and Medicaid Services

Downloaded from http://circ.ahajournals.org/ by guest on March 1, 2017

D2B, American College of Cardiology Foundation Door-to-Balloon Alliance
EMRs, electronic medical records
ICER, incremental cost-effectiveness ratio
INR, international normalized ratio
IOM, Institute of Medicine
MI, myocardial infarction
NCQA, National Committee for Quality Assurance
NQF, National Quality Forum
PCI, percutaneous coronary intervention
PCPI, American Medical Association–Physician Consortium for Performance Improvement

2.2. Strength of Evidence

Initial recommendations for construction of a performance measurement set[1] involved 1) evaluating the strength of evidence supporting a potential performance measure, 2) defining the clinical significance of the outcome most likely to be achieved by adherence to a performance measure, and 3) assessing the magnitude of the association between adherence to the potential performance measure and a clinically important outcome. Because there can be strong financial incentives for a manufacturer to have its diagnostic or therapeutic products included in a performance measure, a clearly articulated approach to the selection of performance measures is needed. The writing committee reviewed current perspectives to determine how to select performance measures that could improve patients' health in clinically meaningful ways.

An important concept in the selection of performance measures is confidence that the selected measures will meaningfully improve the health—either survival or health status (patients' symptoms, function, and quality of life)—of the population to whom the measures are applied. For performance measures other than outcomes, writing committees should clearly establish that the selected process or structural performance measures have a strong association with clinically meaningful outcomes. The strength of this association can be measured in a number of different ways through a qualitative or a quantitative assessment of the likely benefit and the range of uncertainty about the size of that benefit. Under the current paradigm, clinical practice guideline writing committees classify information obtained from mixtures of randomized clinical trials, nonrandomized studies, expert panel consensus, and case studies to create a hierarchical grading system. This system integrates the methodological quality of the underlying evidence (level of evidence, from the highest [A] to the lowest [C]) and the trade-off between benefit and risk (class of recommendation, from the highest [I] to the lowest [III]).[8] In comparing risks and benefits, writing committees ultimately develop a qualitative sense that the benefits outweigh the risks. However, this qualitative assessment is usually based on the number and type of supportive studies rather than the clinical importance of the observed differences in outcome.[6,8] By design this approach elevates ratings to those studies where statistically significant differences in outcomes are replicated in several randomized clinical trials. Reliance on statistically significant differences indicates that there was some benefit from the intervention, however. In an era where many studies use combined end points, though, rather than relying solely on mortality or quality of life, less-important outcomes (including surrogates) may drive the statistical significance of a trial. In fact, because industry-funded clinical trials are primarily designed to provide data to support regulatory approval of a novel treatment and because the US Food and Drug Administration often requires several supportive trials to grant approval, industry trials are often large and replicated. In contrast, nonindustry-sponsored trials, such as those sponsored by the National Institutes of Health or the Veterans Affairs healthcare system, are rarely repeated. Although replication of scientific findings is a key tenet in assessing cause and effect, the strength, accuracy, and clinical importance of the findings must also be weighed. The writing committee believes that an enhanced system for selecting potential performance measures that provides quantitative summaries of the impact on outcomes from adherence to the measure is needed.

Translation of clinical evidence into quantitative summaries for use in the development of performance measurement is challenging but feasible (see Online Appendix C). In particular, explicit assessments of the clinical importance of an observed finding, such as whether 1) the outcome was important and 2) whether the range of possible "true" differences between the treatment groups represents a clinically important difference in outcomes (see example E in Figure 1 of Online Appendix C) will enhance the understanding of the benefit. The writing committee believes that no hard and fast rules of minimal clinically important differences can be created outside the context of a particular intervention and outcome. But converting both survival and health status (eg, being asymptomatic, having a clinically important improvement in function or health-related quality of life) benefits of treatment into meaningful summary metrics, such as number needed to treat with corresponding measures of uncertainty, could help writing committees establish a standard for their use in creating a performance measurement set. Reporting quantitative measures of comparative evidence, such as Bayes factors, will also help writing committees in their decision-making. Regardless of the approach used, writing committees should be explicit as to which outcomes and benefits were considered clinically important in recommending that an intervention be developed into a performance measure. This process would be much more straightforward if clinical trialists, when designing their studies, explicitly stated what defined a clinically important difference in outcomes for each of the end points assessed in the trial. Such routine reporting would markedly simplify the incorporation of study results into guidelines and performance measures.

The Task Force on Performance Measures recommends examining the evidence and range of clinically important benefits to provide quantitative evidence with which to assess the potential benefit of a proposed performance measure. A particular advantage to this approach is its formal specification of clinical benefit and the ability to systematically incorporate the range of available clinical evidence into a transparent analysis demonstrating the confidence with which a benefit of a certain magnitude might be gained from widespread adoption of the clinical practice. An explicit

Downloaded from http://circ.ahajournals.org/ by guest on March 1, 2017

delineation of the clinical logic used to create a performance measure should be disclosed, and a formal process for evaluating existing evidence should be developed with the goal of different performance measures writing committees likely selecting similar processes of care from which to create performance measures. Such a process would have the added advantage of minimizing potential conflicts of interest among members of a performance measures writing committee. The Task Force on Performance Measures recognizes that formally integrating available evidence into a framework to define the clinical significance of a benefit is labor-intensive. Ideally, this would be done by guidelines writing committees, but this is not always the case.[9] Although it is not to be implied that it is the role of performance measures writing committees to conduct such analyses, it is important that explicit articulation of the clinically meaningful benefit of introducing a performance measure be demonstrated and referenced before the measure is created and selected. The ACCF/AHA Task Force on Practice Guidelines is examining alternative approaches to grading clinical evidence. On completion of this process, a more standardized approach can be developed.

2.3. Costs and Performance Measures

The creation of a performance measure implies that all eligible patients (see Section 3.1) for that measure should receive, or at least be considered for, the therapy. The writing committee believes that it is important to consider both the cost-effectiveness and total cost burden of potential performance measures before selection. Although these may change over time, explicitly quantifying the cost-effectiveness of treatments at the time that performance measures are created is aligned with the Institute of Medicine (IOM) goal for a more efficient healthcare system and will minimize the likelihood that unintended economic consequences for society and hospitals emerge from adopting a measure.[7] It is not necessarily the role of performance measure writing committees to conduct formal cost-effectiveness analyses, but the writing committee believes that it is important to consider such analyses during selection of performance measures so that the societal outcomes, including financial outcomes, of implementing performance measures can be transparent. Cost-effectiveness analysis should occur before or be concurrent with performance measure recommendations, should be conducted by parties free of conflicts of interest, and should preferentially use the societal perspective in defining cost-effectiveness.[10] In some situations, therapies are both more effective and less costly than the standard of care (ie, dominant treatments). When this occurs, there is strong justification to promote the intervention to a performance measure because it is likely to both improve care and lower costs. Although other issues may preclude the selection of a dominant treatment as a performance measure (eg, feasibility of implementation), such treatments represent an ideal opportunity for creating performance measures. In most circumstances, however, effective therapies are also associated with increased costs. This creates a need to balance costs against benefits attained, especially because there are competing demands for the limited resources available to governments and societies for improving the health of populations.

There is no consensus on how cost considerations should be integrated into decisions about performance measures. Traditionally, value has been defined as the absolute effectiveness of a given therapy compared with an alternative, conditional on the cost of that therapy (ie, the incremental cost-effectiveness ratio [ICER]).[11] Unfortunately, although most cost-effectiveness studies have been conducted from a societal (ie, population-wide) perspective, significant heterogeneity remains in study designs (eg, in-trial analyses versus Markov models), costing methods (eg, microcosting versus macrocosting), measures of effectiveness, assumptions, and time horizons (eg, 3 years versus lifelong), and there is no consensus as to what ICER threshold (if one should even be put forth) would be considered cost-effective.[10,12] Because of these considerations, there are significant limitations in comparing ICERs across studies. For example, the ICER for implantable cardioverter-defibrillator therapy for primary prevention varies from $34 000 to $235 000 per quality-adjusted life-year across different cost-effectiveness studies, depending on which patient subpopulations are considered.[13–15] Moreover, a cost-effectiveness analysis is not sufficient to fully appreciate issues of cost because it does not provide a transparent reporting of the total cost burden of the intervention to society, which is determined by the cost of the therapy and the prevalence of the condition for which the therapy is indicated. As such, 2 therapies could have identical ICER estimates but vastly different impacts on a healthcare budget with competing demands.

There are other cost considerations for performance measures. In some cases a therapy may be more effective and less costly from a societal perspective, but its implementation may financially penalize clinicians or hospitals (eg, if the therapy prevents hospital readmissions or is poorly reimbursed by payers, such as higher nurse–patient ratios). In these circumstances, when a patient's benefit is expected to increase and the total costs to society should decrease, realignment of dysfunctional economic reimbursements is needed to better align the financial paradigm so that the performance measures can be implemented without disadvantage to a particular component of the healthcare system.

Finally, providing incentives through pay for performance for physician compliance with performance measures may also have unintended consequences on cost-effectiveness. Such incentives may lead to physicians "gaming" the system. By using strategies such as aggressive screening and overdiagnosis, clinicians can appear to achieve better performance with some performance measures (eg, achieving higher rates of hemoglobin A1C [HbA1C] of <7.0 for patients with diabetes mellitus and blood pressure control for patients with hypertension) because their sicker patient population is "diluted" with patients having an early stage or milder forms of a condition.[16] Such efforts only lead to increased population costs for treatment, decreased average net effectiveness of treatment, and, consequently, decreased cost-effectiveness (ie, higher ICERs). Moreover, the use of artificial thresholds to warrant payments may present a problem by rewarding a practice that achieves that threshold (eg, lowering HbA1C from

Downloaded from http://circ.ahajournals.org/ by guest on March 1, 2017

7.1 to 6.9) rather than making a more substantive improvement in patient management (eg, lowering HbA1C from 12.0 to 7.5).

Recognizing both the responsibilities of advancing a more efficient healthcare system and the existing limitations and lack of standardized methods in assessing costs and cost-effectiveness, the Task Force on Performance Measures believes that a working committee should be created to develop recommendations and standards for applying considerations of cost and cost-effectiveness to the creation of performance measures, including any potential medicolegal consequences of explicitly considering cost considerations in performance measures. Because all performance measures writing groups will confront the challenge of having to integrate costs into their selection process, an overarching strategy needs to be developed and implemented.

3. New Insights Into the Construction of Performance Measures

3.1. Use of Exceptions in Performance Measures

One area of performance measure creation that has garnered significant attention in the past few years is the subject of exclusions. As noted in the initial methodology report,[1] "Occasionally the denominator will exclude subsets of patients within the target population and the dimension of care for the performance measure" (p 1153). These exclusions might more accurately be termed exceptions because the data from these patients should still be captured for purposes of internal quality improvement analyses, even though the data may not be included in performance measurement reports. This also implies that the performance measure was at least considered for each potentially eligible patient, a primary goal of performance measures and the quality improvement that they are intended to facilitate. Provisions for exceptions should be made in most process and outcome measures that are used for accountability, including both provider compensation (pay for performance) and public reporting purposes. It is less critical to provide for exceptions when measures are used solely for internal quality improvement, although collection and use of these measures could still be useful to physicians in analyzing practice patterns. A detailed discussion of the logic for this perspective is provided in Online Appendix D.

According to a useful construct developed by the American Medical Association–Physician Consortium for Performance Improvement (PCPI), exceptions to the use of process-based performance measures can be documented on the basis of medical, patient, or system reasons.[17] These major categories are further delineated into subcategories:

Medical Reasons

- Contraindicated (patient history of allergy, potential adverse drug reaction, other)
- Not indicated (already received/performed, not likely to benefit, other)
- Intolerant (therapy tried and patient could not tolerate it)
- Other medical reason(s)

Patient Reasons

- Patient preference
- Social reason(s)
- Religious reason(s)
- Economic reason(s)*
- Other patient reason(s)

System Reasons

- Resources to perform services not available
- Insurance coverage/payer-related limitations
- Service/treatment to be provided by another physician
- Other reasons attributable to healthcare delivery system

The principal advantage of such a categorization is the decreased burden of data collection. Rather than listing and collecting data on each potential contraindication—and given the virtually infinite number of unique situations that likely exist to justify when a performance measure should be responsibly withheld from a potentially eligible patient—it is now possible to merely select the category for which the performance measure is not appropriate. The writing committee continues to support this framing, given its advantages in improving the feasibility of performance measurement, but recognizes that there are potential problems with both the reproducibility of assigning a specific contraindication into the correct category and the possibility of "gaming" the performance assessment efforts by incorrectly excluding potentially eligible patients. To correct misclassification, either additional staff training or more accurate coding of electronic medical records (EMRs) is likely to be needed. Professional ethics and a structured audit system are the 2 most effective means of minimizing intentional manipulation of patient data to achieve artificially better performance reports. Selective auditing of practices with a large proportion of potentially eligible patients with exclusions might be one way of ensuring more accurate categorization of exclusions. Importantly, the writing committee believes that if a patient has a potential exclusion but receives the treatment, then that patient should be included in both the numerator and denominator of the measure. Finally, it is recognized that the use of patient-level exclusions has the potential to sustain or exacerbate disparities in care. In its report, "Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care," the IOM acknowledges that poor (or culturally insensitive) presentations of treatment recommendations may potentially influence patient decisions.[18] To the extent that the quality of medical explanations presented to different racial, ethnic, sex, or age groups varies, then patients may refuse treatment—a "patient-centered" exclusion—even though a better or more culturally sensitive explanation might have led the patient to

*Although the PCPI has placed economic reasons under System Reasons for exclusion because insurers can greatly influence both the tier and level of copayments that patients are required to pay, the writing committee thought that the choice to buy a medication is ultimately one made by the patient, and if the patient chose not to buy an expensive medication, it was appropriate to include this exclusion within the Patient Reasons category. Because the exclusion of a patient from the denominator of a performance measure is not influenced by the category of the exclusion, this should not alter current estimates of performance.

Downloaded from http://circ.ahajournals.org/ by guest on March 1, 2017

accept the therapy and receive it. This underscores the need to examine not only actual performance rates, but also the proportion of providers' populations excluded so that outliers (eg, those with large proportions of potentially eligible patients excluded) can be identified for further investigation.

In summary, the Task Force on Performance Measures supports the application of exclusions by removal of patients from the denominator. If a patient with a potential exclusion does in fact receive treatment, then the patient should be counted in both the numerator and denominator. This approach recognizes that some contraindications are relative, and if a clinician believes that the benefits of a treatment outweigh its risks, then the clinician should receive credit for successfully fulfilling the performance measure.

3.2. Considerations in the Use of Outcomes Measures

Outcomes measures are emerging as a critical component of the measurement portfolio. In 2007, the Centers for Medicare and Medicaid Services (CMS) began to publicly report the results of its National Quality Forum (NQF)–approved 30-day mortality measures for patients admitted with heart failure and acute myocardial infarction (AMI), respectively. In 2008, a 30-day mortality measure for pneumonia was added, and in 2009, 30-day readmission data for these conditions were added.[19] Additional organizations, such as the Society for Thoracic Surgeons, have NQF-approved mortality measures. These measures are understood as complementing the process measures because they provide a broader perspective on the quality of care provided.[20]

The Task Force on Performance Measures recognizes significant strengths and limitations in the use of outcomes as performance measures. First, there is no debate as to the importance of clinically meaningful outcomes, such as mortality and health status.[21] It therefore follows that the most interpretable and potentially important performance measures to patients are outcomes measures. Potential limitations include the fact that some patients are more likely to have adverse outcomes regardless of the quality of care received, and that the system should encourage, not discourage, care for such high-risk patients. One method for preventing negative consequences from treating the sickest patients is the use of risk adjustment to "level the playing field." But even the best risk-adjustment models can explain only a modest proportion of the observed variation in outcomes. This may be a limitation if other unmeasured patient characteristics account for differences in outcomes or a strength if the unmeasured variance is due to differences in quality. In addition, for many outcomes, no well-validated risk-adjustment models exist.

The Task Force on Performance Measures recognizes that debate continues regarding the inclusion of specific risk-adjustment factors. A particularly controversial issue is whether race and socioeconomic status should be included in models. Some experts argue that these variables can carry important prognostic information that can improve the performance of the models and perhaps serve as surrogates for chronic diseases or poorly managed conditions before hospital admission. Others argue that these characteristics may be associated with the quality of care and that their inclusion may "adjust away" quality differences among providers, making it possible that those caring for vulnerable populations who perform poorly will not be identified. Finally, the clinical interpretation of models that adjust for race or sex would suggest that different outcomes for blacks or women are acceptable and could undermine the goal of equity in US health care.[7] Consequently, the CMS measures do not adjust for race because of the concern of creating different standards of care by the use of these variables.

In 2006, the AHA published a consensus statement with the endorsement of the ACCF that articulated the key attributes of outcomes measures suitable for public reporting.[22] In developing the statement, the writing group, which included clinicians, quality experts, a statistician, and policymakers, identified the following 7 preferred attributes:

1) clear and explicit definition of an appropriate patient sample
2) clinical coherence of model variables
3) sufficiently high-quality and timely data
4) designation of an appropriate reference time before which covariates are derived and after which outcomes are measured
5) use of an appropriate outcome and a standardized period of outcome assessment
6) application of an analytic approach that takes into account the multilevel organization of data
7) disclosure of the methods used to compare outcomes, including disclosure of performance of risk-adjustment methodology in derivation and validation samples

The Task Force on Performance Measures supports these standards in developing valid and useful outcomes-based performance measures, although several clear challenges exist in their application.

An example of such a challenge is in the use of outcomes to evaluate the quality of coronary revascularization. Current risk-adjustment methods are useful for comparing one provider's performance against all other studied providers who have performed that procedure. But these same risk-adjusted results may not be appropriate for comparing one hospital with another, leading to errors in interpretation. Current outcomes-based mortality performance measures exist for both bypass surgery and percutaneous coronary intervention (PCI). Because these treatments are handled separately, it is possible that 2 institutions with identical patient populations and performances may look very different. Online Appendix E outlines such a potential scenario for both periprocedural mortality outcomes and efficiency when bypass surgery and PCI are examined independently or together. To facilitate these comparisons between hospitals, it would be more appropriate to redefine the population of analyzed patients as those with significant obstructive coronary artery disease rather than create one stratum for those undergoing PCI and another for those being treated with bypass surgery. Consequently, the writing committee favors, wherever possible, using a clinical condition and state rather than a procedure as the basis for applying an outcomes-based performance measure. Nevertheless, all of the domains articulated by this

Downloaded from http://circ.ahajournals.org/ by guest on March 1, 2017

report are similarly complex, and transparency by writing committees is needed to support and promulgate outcomes-based performance measures, a clear priority for performance measurement development.

Another outcome measure relates to patient health status. Several performance measurement sets include the assessment of patients' symptoms and function—a process—as measures of healthcare quality meeting the dimension of care associated with serial monitoring of patients.[23,24] The results of these assessments, although not currently reported, would be a clinically important outcome measure and could provide quantitative information on the variability in symptom control and quality of life of outpatients with coronary disease or heart failure. A recent national study of primary care clinics in Australia examined the proportion of each clinic's patients with coronary disease who had weekly or more frequent episodes of angina. The results showed that although 14% of clinics had no patients with weekly angina, in 18% of clinics, more than half of patients had weekly angina (weekly episodes of angina across the 207 clinics ranged from 0% to 100%).[25] However, until robust risk-adjustment models are developed, these outcomes are likely better used as tools for quality improvement than accountability.

The use of outcomes measures as indicators of quality are currently best understood as tools to assist hospitals and healthcare professionals to understand their performance. Because not all adverse events represent a failure of quality, given that some events cannot be averted even with the highest-quality care, the goal of outcomes-based assessments is to show relative differences in performance across delivery systems so that those with the worst performance may reflect on ways to improve care or to learn from those with the best performance.

3.3. Numbers of Measures

The proliferation of agencies and entities developing performance measures, often with different methodological rigor, goals, and perspectives, is creating an unmanageable burden for providers that threatens to undermine the stated goal of performance measurement: to improve the quality of health care. Not only do professional organizations, such as the ACCF, the AHA, and the PCPI, create performance measures within a given disease, but payers (eg, CMS, United Healthcare, Blue Cross/Blue Shield) and other accrediting bodies (eg, the National Committee for Quality Assurance [NCQA]) also create unique measures for the same condition(s). Consequently, even when the same process of care is being evaluated, subtle differences in measure specifications can lead to a marked administrative burden in properly providing the requisite data, as well as differences in results. It is therefore possible that different assessors examining the same patients may reach different conclusions about a provider's performance. Because these differences can be attributable to the method of assessment rather than the quality of care provided, they have the potential to undermine trust and confidence in the system and can impair the capability of performance measurements to be used to improve care. The Task Force on Performance Measures strongly supports the need to attain national consensus on a limited number of measures that are universally accepted by all who are interested in performance assessment. Toward that end, the task force has been actively engaged with other professional organizations (eg, the PCPI) and payers (eg, the Joint Commission, CMS, NCQA) to achieve consensus on definitions of these measures. Yet the different perspectives of these different bodies sometimes make reconciliation difficult. These differences are being negotiated to achieve balance between the available clinical evidence, the need to have clinical rather than administrative data, and feasibility. By using the same measures with the same definitions and a reasonable number of requisite data elements, consistent collection and benchmarking is likely to be far more feasible.

The Task Force on Performance Measures recommends 2 additional strategies to limit the number of measures. First, the NQF has emerged as a national clearinghouse for vetting and approving measures. On the one hand, this provides a valuable validation of the methodology used to create measures and should theoretically elevate the quality of a performance measurement set. On the other hand, it also accepts measures from multiple different entities and has the potential to include measures that are built from data systems (eg, administrative data, proprietary date ranges surrounding a seminal event) that are opaque to clinicians, create administrative costs to understand and contest, and are not sufficiently actionable so that they cannot be used for quality improvement, the ultimate purpose of any quality-assessment program. Over time, a vetting process such as that provided by the NQF should limit the number of measures to only those that have the greatest potential to achieve the goal of improving healthcare quality. Currently, CMS is reporting only measures that have been endorsed by the NQF.

The other promising approach is the creation of measure sets for a given disease and rotation of selected measures over time. By selecting a subset of measures to be used, a much more practical data collection effort that can be more easily accomplished by a larger number of practices and institutions can be undertaken and linked to meaningful quality improvement. The national Door-to-Balloon (D2B) Alliance, which reduced delays in performance of primary PCI, is a notable example of this approach.[26,27] What is needed is a national consensus on which measures should be used over what period of time. Ideally such a decision-making body would include payers (for their pay-for-performance programs), regulators (for accreditation and public reporting), methodologists, and clinician representatives of those who care for patients with the disease under assessment. Over time, some measures would be retired and others introduced. Ideally, when the subset of measures to be selected is defined, measures from all of the multiple dimensions of care would be chosen so that a more comprehensive assessment of quality health care could be attained. Those measures that are not actively being used to quantify performance, either because there is inadequate variability in care, difficulties in collection, or insufficient data to support their elevation as performance measures could still be used as quality metrics. A recent statement by the ACCF and AHA delineates the differences between these 2 types of measures.[28]

Downloaded from http://circ.ahajournals.org/ by guest on March 1, 2017

3.4. Modification and Retirement of Measures

To date there has been a strong push to expand the number and diversity of the performance measures portfolio to provide a more complete assessment of quality. However, there is also a need to periodically reconsider whether previously established performance measures should be modified or retired. This can occur for several reasons. First, new scientific evidence may come to light that changes the previous consensus views regarding a measure. An example of such a change is the use of early beta-blocker therapy for patients with AMI.[25] This performance measure was originally based on older trials that found that acute beta-blocker therapy reduced postinfarction angina, arrhythmias, and reinfarction risks. More contemporary trials, however, that found acute beta-blocker therapy had no net impact on mortality. Although such therapy reduced deaths from arrhythmias, it also increased risks for cardiogenic shock in certain subpopulations.[29] On the basis of changes in guideline recommendations, the ST-Elevation and Non–ST-Elevation Myocardial Infarction Performance Measures Writing Committee determined that early beta-blocker therapy should be dropped from the measure set, citing the complexity required to distinguish patients who benefit from this therapy from those who may be harmed.[25] Similarly, because cigarette smoking is known to have a detrimental impact on cardiovascular health and there is evidence that high-intensity behavioral and pharmacological therapies can help patients quit,[30] smoking cessation counseling was developed as a performance measure for several conditions.[24,25] But recent studies found a striking discordance between hospital performance on this measure and the rates at which patients actually quit smoking after myocardial infarction (MI).[31,32] These data suggest the need for reevaluation of the smoking cessation measure.

A second reason for modifying or dropping a performance measure is that collection of the data necessary to calculate the measure is prohibitively complex or expensive. In some cases these issues can be corrected with clearer instructions, more training, or minor changes in the numerator or denominator of the measure.

A third reason for considering revision or retirement of a measure is if its use has unintended adverse consequences. An example of this was recently raised regarding a performance measure to give intravenous antibiotics for community-acquired pneumonia within 4 hours of diagnosis. Although rapid administration of antibiotics is beneficial for patients with pneumonia, the metric has been criticized because it may pressure clinicians to administer antibiotics despite diagnostic uncertainty and may lead to overtreatment.[33] Similar concerns have been raised that the current ACCF/AHA performance measure for D2B within 90 minutes may lead to an increase in false-positive activation of cardiac catheterization for patients with suspected ST-elevation MI.[34] Although, in this case, the net benefits of the D2B measure likely outweigh the risks, such examples highlight the need to carefully study the real-world impact of performance measures on provider care and patient outcomes to minimize unintended consequences.[35]

A final reason for retiring a measure is when there is limited to no room for further improvement in performance and clinical practice reaches near-perfection. Currently, several MI performance measures are achieving asymptotic "ceilings" of performance, including aspirin at arrival and discharge, as well as beta-blocker use at discharge for patients with AMI.[36] This achievement of near-perfection in performance should be seen as a celebration for the field and a mark of success of the performance measure and quality-improvement cycle. Yet, ever conscious of the burden of data collection on the provider, some have argued for consideration of retiring these metrics.[37] Retirement of a performance measure because of its success, however, should also be carefully monitored, because there is a risk that ending active measurement may lead to provider complacency and ultimately a regression in performance. As noted in Section 3.3, recycling measures after a period of dormancy can both assess the sustainability of the original performance assessment effort and reinforce the need for this process of care.

4. New Insights Into the Implementation of Performance Measures

Although the initial publication addressing the methodology for measures creation and selection[1] explicitly called for feasibility testing before endorsement of performance measures, this has rarely been done. The writing committee wanted to emphasize the importance of preliminary testing of proposed measures in local, regional, or national projects before application for purposes of accountability. Congruent with this perspective, the NQF has begun issuing only time-limited endorsements of proposed measures pending demonstration of their feasibility.[38] A number of potential barriers exist that could render an otherwise valuable potential performance measure impractical to collect in clinical practice.

The burden of data collection has emerged as a primary challenge to implementation of performance measures. Not only do multiple performance measures often exist for a particular condition, but patients also have multiple diseases, so that it is not practical or possible to collect all measures for all patients. This situation is compounded, in particular, by the superiority of clinical data over administrative data one encounters when seeking to quantify the quality of health care.[39] The concepts described in this report, including elevating the evidentiary threshold for endorsing a performance measure, simplifying the inclusion/exclusion criteria, limiting the number of measures, and retiring measures, may all lead to a more parsimonious, feasible measurement set for quality improvement and accountability.

A second critical aspect of collecting performance data is the integration of data collection through the process of providing care. To the extent that extra work is needed to provide the data required for performance assessment, the more unsustainable such a program will become. An important responsibility of performance measures writing committees is to consider how data elements can be acquired throughout the transactions of a clinical encounter without requiring the collection and recording of additional data at a clinical visit. The challenge for the Task Force on Performance Measures is to consider how multiple measurement sets for different conditions, the similarity of measure

Downloaded from http://circ.ahajournals.org/ by guest on March 1, 2017

construction across diseases, and the totality of ACCF/AHA-approved measures might affect a clinical practice or institution.

Although EMRs would seem to offer a potential solution, this is not currently the case. Many systems are unable to export the collected data to other entities for performance assessment and improvement, have data definitions that are not congruent with those used by the developers of performance measures (eg, the ACCF/AHA data standards[40]) and may require "pop-ups" and other prompts that are increasingly ignored by practitioners frustrated by the perception that these aids are interfering with efficient patient care. It is important that the effort to create exporting functions from EMR systems in standard formats be accelerated so that those who use EMRs can more efficiently participate in quality assessment and improvement efforts. Although alternatives, such as the patient flow sheets proposed by the PCPI and prior ACCF/AHA performance measures writing committees, still have some potential to help in performance measurement, a range of strategies for data collection needs to be considered by writing committees. In addition, it would be valuable for experts in medical informatics to participate in such writing groups, given the unique perspective and knowledge required to convert clinical logic into code.

Beyond the challenges of data collection, other insights have emerged over the past several years, including the need to develop "windows" around timeframes for performance. For example, although it is reasonable to state that cholesterol levels should be assessed every year in a patient with chronic stable coronary artery disease,[23] a patient assessed in the last week of December one year and the first week of January in the following year, 12½ months later, would not meet the measure. Even more challenging is the current requirement for a patient with atrial fibrillation to have an international normalized ratio (INR) measurement every month.[41] Even if a patient had 10 to 12 INR assessments per year, which would generally be considered high-quality anticoagulation management, many of these assessments might be within the same month, whereas in other months there might be none. The increasing use of home INR monitoring[42] further compounds the problem. Although there may be no solution to handling the example of serial cholesterol screening that falls just outside a reporting window, the case of atrial fibrillation might be better handled with a range of possible assessments over the entire reporting window (eg, ≥10 assessments within the reporting year) to minimize the challenges in accurately representing the quality of care being provided.

A final lesson learned from early experiences in performance measurement is the limited reproducibility of some measures. For example, measures that encourage counseling represent an important dimension of care, but the quality of delivering counseling is difficult to quantify and yet may have profound influence on the ability of the activity to achieve its desired outcomes. Also, as noted above, smoking cessation counseling at the time of an MI has long been endorsed as a performance measure,[43] yet recent data suggest that there is no overall association between smoking cessation counseling and quit rates among smokers.[32] In these studies, however, there was an association between the presence of an inpatient smoking cessation program with patients stopping smoking after discharge.[31] Although not definitive, these findings suggest that the quality of counseling achieved may influence behavior and that failure to quantify or specify the quality of counseling efforts may lead to a measure that is not associated with a clinically meaningful outcome. From this experience, future performance measurement writing groups need to be confident that proposed performance measures can be adequately quantified so that the expected benefits from adherence to the measure can deliver the expected benefits in outcome.

5. New Insights Into the Analysis and Interpretation of Performance Measures

5.1. Composite Measures

Combining measures or indicators of performance is a relatively new consideration for assessing the quality of medical care.[44–46] The proliferation of efforts to measure, publicly report, or reward healthcare providers has focused attention on the need to ensure that performance measures comprehensively represent the quality of care, including sampling from among the multiple dimensions of care identified in the original methods report.[1] Composite quality of care measures are increasingly being developed and deployed. A composite measure is a single measure of a construct that is defined in terms of ≥2 individual measures. Although composite measures have many advantages, their construction, development, and validation require more attention than that needed for individual performance measures for a number of reasons.[47–50]

Several lessons have been learned with the construction of composite quality measures. First, standard psychometric properties of composites, such as reliability, accuracy, and predictive validity, may be difficult to demonstrate. In particular, there may be no universal standard for some composite measures. Consequently, other quality or health measures should be shown to be related to the composite. Individual measures that make up the composite performance measure should contribute unique information to the underlying construct but at the same time should not differ from the other components of the composite.

Second, the scoring methods used to create the composite measure deserve serious consideration. A scoring method is the rule used to combine the individual components of the composite. Common methods of combining individual components include *all-or-none* rules, where a success is declared only if all the individual components are met (conjunctive scoring); *any* rules, where a success is declared if any of the specified components are met (compensatory scoring); and *empirically weighted* rules, where a number is produced using the variability in the data to determine the weight of each specific component (factor analytically derived or item response theory derived). It is important to note, however, that these methods can lead to different conclusions.

Third, although missing data always pose a problem in any analysis, the extent and impact of missing data can be hidden depending on the scoring rule. Moreover, the scoring strategy

Downloaded from http://circ.ahajournals.org/ by guest on March 1, 2017

may affect how missing data are handled. For *all-or-none* rules, if a single component of the composite is missing, then the composite is missing; however, for *any* rules, as long as one component is observed to have met success, the composite is *observed.* Strategies for handling missing data in the scoring rule must be transparent and valid.

Fourth, because some individual performance measures may be continuous (health status) and some may be binary (within-range blood pressure), statistically combining such measures requires some thought. Most applied researchers will try to solve this problem by converting all individual components into the same scale. Although this is an easy solution, it is associated with a loss of information.

To address these challenges, the ACCF/AHA Task Force on Performance Measures has developed a position statement on composite measures.[51] In addition, several professional organizations have developed recommendations for the development of valid composite measures. The NQF has created a consensus report that has outlined a composite evaluation framework.[52] Only those composite measurements aligned with these recommendations will be considered as potential performance measures. The Task Force on Performance Measures recommends that composite performance measures follow the criteria described by the ACCF/AHA 2010 Position Statement on Composite Measures for Healthcare Performance Assessment[51] as outlined in Table 2.

5.2. Attribution

The majority of patients who have cardiovascular disease have multiple comorbidities and hence often have multiple healthcare providers within a single system of care or among different systems of care. The complexity of measuring the quality of coordinated cardiovascular care across multiple healthcare professionals and multiple settings is compounded by the difficulty in establishing the appropriate individual, institution, or healthcare system to which to assign attribution or accountability. Although some aspects of care and care coordination are suitable for measurement at the level of the individual and appropriate accountability lies with the individual provider, others are more appropriate for measurement at the group, institutional, or system level.

The IOM has called for measurement approaches that foster shared accountability.[7] In such measures, all members of the healthcare team(s) are held accountable for quality and efficiency of care. The IOM has identified gaps in current performance measurement sets, including too few measures of patient-centered care, too few focusing on more than a narrow time window, and too few with more than a narrow focus of accountability beyond individual provider actions.

The NQF has endorsed measures for efficiency of episodes of care across the continuum of care[53] that focus on quality and efficiency of care as perceived by the patient rather than by the healthcare provider or institution. In this construct the framework for efficiency measurement addresses all levels of the healthcare system, including individual providers, provider organizations, and communities. From the patient's perspective, an episode of care is not a discrete encounter or hospitalization but a longitudinal experience that may last months to years or even an entire lifetime. For example, the patient's experience of an AMI does not begin and end with the D2B time but encompasses the full spectrum from onset of chest discomfort through activation of an emergency medical system, the hospital experience, discharge planning, and return to long-term outpatient care and rehabilitation. In some cases this experience also entails end-of-life planning and palliative care. There are multiple real and potential gaps in care related to the many transitions in this definition of an episode of care. For a hospital discharge, this would include, among others, medication reconciliation, transmission of the discharge record, timeliness of postdischarge tests and services, and patient understanding of the discharge plan and care needed. New longitudinal measures need to be developed to fill gaps in the episode-of-care framework related to transitions from inpatient to outpatient settings (and vice versa), transitions among health systems, changes in the plan of care, and transitions and hand-offs among multiple providers. Care coordination is essential because these transitions can be disconnected, uncoordinated, and unsafe. Assigning attribution in this framework is difficult and can only be accomplished by assuming that there is shared accountability for the quality of care provided across all providers, institutions, and systems involved in the episode of care.

The concept of an accountable care organization, based on the local delivery system (eg, a multispecialty group practice or hospital and extended professional staff) from which patients receive the majority of their care, rather than the individual practitioner, has been proposed as a first step in creating the sense of shared accountability.[54–56] The patient-centered medical home also promotes shared accountability, because members of the team are equally responsible for satisfactory delivery of the care plan.[57] When patients transition among different delivery systems, however, measures also need to aggregate care and care coordination across sites and over time to operationalize the concept of shared accountability.

Thus, the concept of shared accountability may be effective in well-organized systems of care that exist for some patients, such as those enrolled in the Veterans Affairs system, but would be associated with pitfalls and unintended consequences if applied at the provider or institutional level for the majority of patients who transition over time among providers and care sites. For example, length of stay will be prolonged if an institution transfers patients only to preferred nursing homes (and a bed is not available at that facility), and acute cardiac care will be diminished if an institution refuses transfer of patients from another institution considered to be of poorer quality. Conversely, current performance measures may exclude patients who are transferred from one institution (where a procedural complication may have occurred) to another institution, so that neither institution may have the adverse outcome attributed to their system. Some aspects of quality of care of acute coronary syndromes are related to the emergency preparedness of the community, such as 9-1-1 responsiveness and availability of rapid transfer between institutions, and this may be beyond the ability of the medical profession to control. Finally, there would also be unique methodological issues in implementing a system

Downloaded from http://circ.ahajournals.org/ by guest on March 1, 2017

Table 2. Choosing Performance Measures

Choosing Performance Measures	
Selection Factors	New Considerations
Confidence that adherence to a potential performance measure would result in meaningful improvements in clinically important outcomes	The methodology and logic by which a performance measure was selected, with a clear description of anticipated benefits on meaningful clinical outcomes, should be disclosed by the writing committee.
Costs of measure	Explicit demonstration that application of the performance measure to patients is associated with an acceptable ICER.
	An estimate of the societal burden of more complete adherence to the performance measure.
Outcomes measures	The outcomes to be considered must be clinically relevant, including mortality, irreversible morbidity, and health status (symptoms, function, and quality of life), and surrogate outcomes should be avoided.
	Previously published recommendations for publicly reported outcomes should be followed.[22]
	Risk adjustment, with carefully selected clinical variables and explicit consideration of demographic characteristics, must be available to render observations interpretable.
	Where possible the population should represent a disease state rather than a procedure applied to a subset of that population.
Measure Characteristics	New Considerations
Use of exceptions	Exclusions of patients from the denominator of a performance measure are reasonable and should be broadly grouped into
	• medical
	• patient
	• system-based reasons for why the patient was not eligible
No. of measures	To minimize the number of measures, efforts at creating a national consensus, involving all stakeholders, on measures to be used for a specified period of time for accountability, pay for performance, and quality-improvement efforts should be developed.
	Although performance measure writing committees should create a full complement of measures for a disease, the NQF should select only a subset of these for use at any particular time.
	The subset should include measures from multiple dimensions of care to facilitate a more complete assessment of quality.
	Measures should be retired when new evidence questions the association of those measures with clinically meaningful outcomes or performance is so high that there is little room for future improvement.
	Retired measures should be considered for reassessment in future years.
Feasibility of data collection	Data collection should occur prospectively through routine transactions of medical care because retrospective collection of data is not sustainable.
	EMR companies need to create and support export of data using standardized formats so that a greater number of providers can participate in national quality-assessment programs.
	Measures need to be developed in a way that recognizes the longitudinal patient care experience and creates “windows” for capturing performance that are practical and clinically interpretable.
Composite measures	The psychometric properties of these measures, including reliability, accuracy, and predictive validity, should be demonstrated.
	The purpose, intended audience, and scope of a composite measure should be explicitly stated.
	The individual measures used to create a composite measure should be evidence-based and reliable.
	The methodology used for weighting and combining individual measures into a composite performance measure should be transparent and empirically tested.
	Composite performance measure reporting by providers should include a measure of the degree of uncertainty surrounding composite estimates.
Attribution	Accountability is an important opportunity to improve practice.
	It is essential that those held accountable have the processes of care being assessed under their locus of control.
	More methodological work is needed for promotion of the concept of shared accountability for evaluating transitions in care.

EMR indicates electronic medical record; ICER, incremental cost-effectiveness ratio; and NQF, National Quality Forum.

of shared attribution related to feasibility and determination of sample sizes needed for measurement. The Task Force on Performance Measures recommends that the concept of shared accountability undergo appropriate field testing before there is further consideration of implementation. However, it would be prudent for hospitals to examine and reengineer their processes of discharge planning, patient education/self-management, and communication with community physicians and nurses if they wish to improve hand-offs and decrease 30-day mortality and readmission rates. This could be an important first step toward shared accountability.

Downloaded from http://circ.ahajournals.org/ by guest on March 1, 2017

6. Conclusion

This update to the methodology of performance measure selection and creation seeks to clarify key challenges and opportunities to elevate the science of quality assessment and improvement. Experience since the publication of the initial methodology report has identified critical opportunities to improve the selection, construction, implementation, and interpretation of performance measures (Table 2). With respect to the selection of potential measures, there is a pressing need to elevate the transparency and rigor by which the evidence supporting a performance measure is synthesized, including a focus on clinically meaningful outcomes, and the need to express the costs, both incremental cost-effectiveness and overall societal costs, associated with a potential performance measure. With respect to the construction of performance measures, refinement of patient eligibility, considerations in the use of outcomes, and the number of measures and their retirement have all emerged as important opportunities to improve the process of performance measure creation. With respect to implementation of performance measures, challenges have emerged that require ever-greater scrutiny of the importance of potential performance measures to quality improvement and the need to create measures that can be feasibly collected. Finally, with respect to the analysis and interpretation of performance measures, careful attention to and testing of composite measures and the attribution of performance measures to appropriately accountable units need to be tested before implementation. Although these recommendations substantially increase the complexity and work involved in creating performance measures, the Task Force on Performance Measures believes that following these processes will elevate the consistency and quality of new measures and improve the processes of quality improvement so that patients and society may benefit from higher-quality cardiovascular care.

Staff

American College of Cardiology Foundation

John C. Lewin, MD, Chief Executive Officer

Charlene May, Senior Director, Clinical Policy and Documents

Melanie Shahriary, RN, BSN, Associate Director, Performance Measures and Data Standards

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

Erin A. Barrett, MPS, Senior Specialist, Clinical Policy and Documents

American Heart Association

Nancy Brown, Chief Executive Officer

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

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

Nereida Crawford, MPH, Science and Medicine Advisor

Appendixes

Appendix A. Author Relationships With Industry and Other Entities–ACCF/AHA New Insights Into the Methodology of Performance Measurement

Committee Name	Employment	Consultant	Speaker	Ownership/Partnership/Principal	Research	Institutional, Organizational, or Other Financial Benefit	Expert Witness
John A. Spertus	Mid America Heart Institute–Director, Outcomes Research	Amgen Novartis St. Jude Medical United Healthcare	None	Copyright to Seattle Angina Questionnaire, Kansas City Cardiomyopathy Questionnaire, and Peripheral Artery Questionnaire–Health Outcomes Sciences	Amgen* Bristol-Myers Squibb/ Sanofi-aventis* Eli Lilly* Medtronic	None	None
Robert O. Bonow	Northwestern University Feinberg School of Medicine; Goldberg Distinguished Professor	None	None	None	None	None	None
Paul Chan	Mid America Heart Institute–Division of Cardiology; Quality Care Researcher	None	None	None	None	None	None
George A. Diamond	Cedars-Sinai Medical Center–Senior Research Scientist	None	None	None	None	None	None
Joseph P. Drozda, Jr	Sisters of Mercy Health System–Director, Comparative Effectiveness Research	None	None	None	None	None	None
Sanjay Kaul	Cedars-Sinai Medical Center–Director, Cardiology Fellowship Training Program	None	None	None	None	None	None
Harlan M. Krumholz	Yale University School of Medicine–Professor of Medicine, Epidemiology, and Public Health	Alere Amgen United Health* VHA, Inc*	None	None	None	None	Vioxx litigation for plaintiff
Frederick A. Masoudi	Denver Health Medical Center, Division of Cardiology–Associate Professor of Medicine	None	Amgen Takeda United Health	None	Amgen	None	None

(*Continued*)

Downloaded from http://circ.ahajournals.org/ by guest on March 1, 2017

Appendix A. Continued

Committee Name	Employment	Consultant	Speaker	Ownership/Partnership/Principal	Research	Institutional, Organizational, or Other Financial Benefit	Expert Witness
Sharon-Lise T. Normand	Harvard Medical School–Associate Professor of Biostatistics	None	None	None	None	None	None
Eric D. Peterson	Duke Clinical Research Institute–Professor of Medicine; Director, Cardiovascular Outcomes	None	None	None	Bristol-Myers Squibb/Sanofi-aventis Corgentech CV Therapeutics Merck Schering-Plough	None	None
Martha J. Radford	NYU Hospital Center–Professor of Medicine; Chief of Clinical Quality	None	None	None	None	None	None
John S. Rumsfeld	University of Colorado, Denver VA Medical Center–Associate Professor of Medicine; Clinical Coordinator, VA Ischemic Heart Disease	United Healthcare	None	None	None	None	None

NYU indicates New York University, and VA, Veterans Affairs.

This table represents the relevant relationships of committee members with industry and other entities that were determined to be relevant to this document. These relationships were reviewed and updated in conjunction with all meetings and/or conference calls of the writing committee during the document development process. The table 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 relationship.

Appendix B. Peer Reviewer Relationships With Industry and Other Entities–ACCF/AHA New Insights Into the Methodology of Performance Measurement

Peer Reviewer	Representation	Consultant	Speaker	Ownership/Partnership/Principal	Research	Institutional, Organizational, or Other Financial Benefit	Expert Witness
Paul N. Casale	Official Reviewer—ACCF Board of Governors	None	None	None	None	None	None
Lee Fleisher	Official Reviewer—AHA	None	None	None	NIH	Accreditation Association for Ambulatory Care Quality Institute AstraZeneca*	2009 Defendant, perioperative stroke
Edward Havranek	Official Reviewer—AHA	None	None	None	NHLBI*	None	None
Steve Nissen	Official Reviewer—ACCF Board of Trustees	Alnylam Pharmaceuticals GlaxoSmithKline Hollis-Eden Karo Bio Neptune Novo Nordisk Sanofi-aventis	None	None	AstraZeneca* Daiichi-Sankyo* Eli Lilly* Novartis* Pfizer* Resverlogix* Roche* Sanofi-aventis* Takeda*	None	None
H. Vernon Anderson	Content Reviewer—ACCF–NCDR Carotid Artery Revascularization and Endarterectomy Registry Subcommittee	Watermark Research Partners	Bristol-Myers-Squibb Pharmaceuticals Sanofi-aventis Pharmaceuticals	None	None	None	None
Chris Cannon	Content Reviewer—ACCF–NCDR Take ACTION Campaign Planning Committee	Automedics Medical Systems Bristol-Myers-Squibb	None	None	Accumetrics* AstraZeneca* GlaxoSmithKline Interkin Kai Pharmaceuticals Merck* Sanofi-aventis/Bristol-Myers Squibb Partnership* Takeda	None	2009 Defendant, postoperative myocardial infarction
Elizabeth DeLong	Content Reviewer—ACCF/AHA Task Force on Performance Measures	None	None	None	None	None	None

(*Continued*)

Downloaded from http://circ.ahajournals.org/ by guest on March 1, 2017

Appendix B. Continued

Peer Reviewer	Representation	Consultant	Speaker	Ownership/ Partnership/ Principal	Research	Institutional, Organizational, or Other Financial Benefit	Expert Witness
John Dent	Content Reviewer—ACCF Board of Governors	None	None	None	None	None	None
Kathleen Grady	Content Reviewer—ACCF/AHA Task Force on Performance Measures	Eli Lilly	None	None	Bristol-Myers Squibb /Sanofi-aventis	None	None
Lee A. Green	Content Reviewer—ACCF/AHA Task Force on Performance Measures	None	None	None	Clinical and Translational Science Award	None	None
Robert Hendel	Content Reviewer—ACCF Task Force on Appropriate Use Criteria	Phx Health United Healthcare	Astellas Pharm*	None	GE Healthcare	IACNL	None
Richard Kovacs	Content Reviewer—ACCF Board of Governors	Abbott Biomedical Systems Cook Group Incorporated ECG Scanning Services* Eli Lilly* Endocyte Essentialis MED Institute* Xenoport	None	None	None	None	None
Ronald Krone	Content Reviewer—ACCF–NCDR Scientific Oversight Committee	None	None	None	None	None	2009 Defendant, atherosclerosis 2009 Defendant, sudden death 2009 Plaintiff, peripartum cardiomyopathy 2005 Defendant, smoking
Eldrin Lewis	Content Reviewer—Individual	Amgen Medtronic	None	None	None	None	None
Ileana Pina	Content Reviewer—AHA QCOR Steering Committee	FDA Sanofi-aventis Solvay	Merck Novartis	None	None	None	None
Leslee Shaw	Content Reviewer—ACCF/AHA Task Force on Practice Guidelines	None	None	None	Astellas Healthcare* Bracco Diagnostic* GE Healthcare*	None	None
Allen Taylor	Content Reviewer—Individual	None	None	None	None	None	None

ACCF indicates American College of Cardiology Foundation; AHA, American Heart Association; FDA, US Food and Drug Administration; IACNL, Intersocietal Commission for the Accreditation of Nuclear Medicine Laboratories; NCDR, National Cardiovascular Data Registry; NIH, National Institutes of Health; NHLBI, National Heart, Lung, and Blood Institute; QCOR, Quality of Care and Outcomes Research; and UCSF, University of California, San Francisco.

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. 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. Names are listed in alphabetical order within each category of review.

*Significant relationship.

References

1. 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.
2. Normand SL, McNeil BJ, Peterson LE, Palmer RH. Eliciting expert opinion using the Delphi technique: identifying performance indicators for cardiovascular disease. Int J Qual Health Care. 1998;10:247–60.
3. Krumholz HM, Keenan PS, Brush JE Jr, et al. Standards for measures used for public reporting of efficiency in health care: a scientific statement from the American Heart Association Interdisciplinary Council on Quality of Care and Outcomes research and the American College of Cardiology Foundation. J Am Coll Cardiol. 2008;52:1518–26.
4. Brush JE Jr, Krumholz HM, Wright JS, et al. American College of Cardiology 2006 principles to guide physician pay-for-performance programs: a report of the American College of Cardiology Work Group on Pay for Performance (A Joint Working Group of the ACC Quality Strategic Direction Committee and the ACC Advocacy Committee). J Am Coll Cardiol. 2006;48:2603–9.
5. Bufalino V, Peterson ED, Krumholz HM, et al. Nonfinancial incentives for quality: a policy statement from the American Heart Association. Circulation. 2007;115:398–401.
6. Gibbons RJ, Smith S, Antman E, et al. Part I: Where do they come from? Circulation. 2003;107:2979–86.
7. Committee on Quality of Health Care in America, Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: National Academy Press; 2001.
8. ACCF/AHA Task Force on Practice Guidelines Methodology Manual and Policies From the ACCF/AHA Task Force on Practice Guidelines. Available at: http://assets.cardiosource.com/Methodology_Manual_for_ACC_AHA_Writing_Committees.pdf. Accessed June 25, 2010.
9. Tricoci P, Allen JM, Kramer JM, et al. Scientific evidence underlying the ACC/AHA clinical practice guidelines. JAMA. 2009;301:831–41.
10. Gold MR, Siegel JE, Russell LB, Weinstein MC, eds. Cost-Effectiveness in Health and Medicine. 1st ed. New York, NY: Oxford University Press; 1996.
11. Diamond GA, Kaul S. Cost, effectiveness, and cost-effectiveness. Circ Cardiovasc Qual Outcomes. 2009;2:49–54.
12. Weintraub W, Cohen D. The limits of cost-effectiveness analysis. Circ Cardiovasc Qual Outcomes. 2009;2:55–8.

Downloaded from http://circ.ahajournals.org/ by guest on March 1, 2017

13. Mark DB, Nelson CL, Anstrom KJ, et al. Cost-effectiveness of defibrillator therapy or amiodarone in chronic stable heart failure: results from the Sudden Cardiac Death in Heart Failure Trial (SCD-HeFT). Circulation. 2006;114:135–42.
14. Sanders GD, Hlatky MA, Owens DK. Cost-effectiveness of implantable cardioverter-defibrillators. N Engl J Med. 2005;353:1471–80.
15. Zwanziger J, Hall WJ, Dick AW, et al. The cost effectiveness of implantable cardioverter-defibrillators: results from the Multicenter Automatic Defibrillator Implantation Trial (MADIT)-II. J Am Coll Cardiol. 2006;47:2310–8.
16. Hayward RA, Kent DM. 6 EZ steps to improving your performance (or how to make P4P pay 4U!). JAMA. 2008;300:255–6.
17. Specifications and categorization of measures exclusions: recommendations to Physician Consortium for Performance Improvement Work Group. Available at: http://www.ama-assn.org/ama1/pub/upload/mm-370/exclusions053008.pdf. Accessed June 25, 2010.
18. Smedley B, Stith A, Nelson A, eds. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. Washington, DC: National Academies Press; 2003.
19. Krumholz HM, Merrill AR, Schone EM, et al. Patterns of hospital performance in acute myocardial infarction and heart failure 30-day mortality and readmission. Circ Cardiovasc Qual Outcomes. 2009;2:407–13.
20. Krumholz HM, Normand SL, Spertus JA, et al. Measuring performance for treating heart attacks and heart failure: the case for outcomes measurement. Health Affairs. 2007;26:75–85.
21. Spertus JA. Evolving applications for patient-centered health status measures. Circulation. 2008;118:2103–10.
22. Krumholz HM, Brindis RG, Brush JE, et al. Standards for statistical models used for public reporting of health outcomes: an American Heart Association scientific statement from the Quality of Care and Outcomes Research Interdisciplinary Writing Group: cosponsored by the Council on Epidemiology and Prevention and the Stroke Council. Circulation. 2006;113:456–62.
23. American College of Cardiology, American Heart Association, Physician Consortium for Performance Improvement. Clinical performance measures: chronic stable coronary artery disease. Available at: http://www.ama-assn.org/ama1/pub/upload/mm/370/cadminisetjune06.pdf. Accessed June 25, 2010.
24. Bonow RO, Bennett S, Casey DE Jr, et al. ACC/AHA clinical performance measures for adults with chronic heart failure: a report of the American College of Cardiology/American Heart Association Task Force on Performance Measures (Writing Committee to Develop Heart Failure Clinical Performance Measures). J Am Coll Cardiol. 2005;46:1144–78.
25. Krumholz HM, Anderson JL, Bachelder BL, et al. ACC/AHA 2008 performance measures for adults with ST-elevation and non–ST-elevation myocardial infarction: a report of the American College of Cardiology/American Heart Association Task Force on Performance Measures (Writing Committee to Develop Performance Measures for ST-Elevation and Non–ST-Elevation Myocardial Infarction), developed in collaboration with the American Academy of Family Physicians and American College of Emergency Physicians. J Am Coll Cardiol. 2008;52:2046–99.
26. Bradley EH, Nallamothu BK, Herrin J, et al. National efforts to improve door-to-balloon time results from the Door-to-Balloon Alliance. J Am Coll Cardiol. 2009;54:2423–9.
27. Krumholz HM, Bradley EH, Nallamothu BK, et al. A campaign to improve the timeliness of primary percutaneous coronary intervention: Door-to-Balloon: An Alliance for Quality. J Am Coll Cardiol Cardiovasc Interv. 2008;1:97–104.
28. Bonow RO, Masoudi FA, Rumsfeld JS, et al. ACC/AHA classification of care metrics: performance measures and quality metrics: a report of the American College of Cardiology/American Heart Association Task Force on Performance Measures. J Am Coll Cardiol. 2008;52:2113–7.
29. Chen ZM, Pan HC, Chen YP, et al. Early intravenous then oral metoprolol in 45,852 patients with acute myocardial infarction: randomised placebo-controlled trial. Lancet. 2005;366:1622–32.
30. Rigotti NA, Munafo MR, Murphy MF, Stead LF. Interventions for smoking cessation in hospitalised patients. Cochrane Database Syst Rev. 2001;(2):CD001837.
31. Dawood N, Vaccarino V, Reid KJ, et al. Predictors of smoking cessation after a myocardial infarction: the role of institutional smoking cessation programs in improving success. Arch Intern Med. 2008;168:1961–7.
32. Reeves GR, Wang TY, Reid KJ, et al. Dissociation between hospital performance of the smoking cessation counseling quality metric and cessation outcomes after myocardial infarction. Arch Intern Med. 2008;168:2111–7.
33. Wachter RM, Flanders SA, Fee C, Pronovost PJ. Public reporting of antibiotic timing in patients with pneumonia: lessons from a flawed performance measure. Ann Intern Med. 2008;149:29–32.
34. Larson DM, Menssen KM, Sharkey SW, et al. "False-positive" cardiac catheterization laboratory activation among patients with suspected ST-segment elevation myocardial infarction. JAMA. 2007;298:2754–60.
35. Masoudi FA. Measuring the quality of primary PCI for ST-segment elevation myocardial infarction: time for balance. JAMA. 2007;298:2790–1.
36. American Heart Association. Heart Disease and Stroke Statistics: Our Guide to Current Statistics and the Supplement to Our Heart and Stroke Facts. 2008 Update At-a-Glance. Available at: http://www.americanheart.org/downloadable/heart/1200078608862HS_Stats%202008.final.pdf. Accessed June 25, 2010.
37. Lee TH. Eulogy for a quality measure. N Engl J Med. 2007;357:1175–7.
38. National Quality Forum. Measure Evaluation Criteria. August 2008. Available at: http://www.qualityforum.org/uploadedFiles/Quality_Forum/Measuring_Performance/tbEvalCriteria2008-08-28Final.pdf. Accessed June 25, 2010.
39. Dove JT, Brindis RG, Fasules JW, et al. President's page: American College of Cardiology's quality leadership influencing health system reform. J Am Coll Cardiol. 2007;50:2016–7.
40. Radford MJ, Heidenreich PA, Bailey SR, et al. ACC/AHA 2007 methodology for the development of clinical data standards: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Data Standards. J Am Coll Cardiol. 2007;49:830–7.
41. Estes NA III, Halperin JL, Calkins H, et al. ACC/AHA/Physician Consortium 2008 clinical performance measures for adults with nonvalvular atrial fibrillation or atrial flutter: a report of the American College of Cardiology/American Heart Association Task Force on Performance Measures and the Physician Consortium for Performance Improvement (Writing Committee to Develop Clinical Performance Measures for Atrial Fibrillation), developed in collaboration with the Heart Rhythm Society. J Am Coll Cardiol. 2008;51:865–84.
42. Jacobson AK. The Home INR Study (THINRS). November 2008. Available at: http://www.theheart.org/article/921079.do. Accessed July 27, 2009.
43. Marciniak TA, Mosedale L, Ellerbeck EF. Quality improvement at the national level: lessons from the Cooperative Cardiovascular Project. Eval Health Prof. 1998;21:525–36.
44. 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.
45. Jacobs R, Goddard M, Smith PC. How robust are hospital ranks based on composite performance measures? Med Care. 2005;43:1177–84.
46. Zaslavsky AM, Shaul JA, Zaborski LB, et al. Combining health plan performance indicators into simpler composite measures. Health Care Financ Rev. 2002;23:101–15.
47. Kaplan S, Normand SL. Conceptual and analytical issues in creating composite measures of ambulatory care performance. 2006. Available at: http://www.qualityforum.org/Projects/c-d/Composite_Evaluation_Framework/5_24_07_Kaplan-Normand_pdf.aspx. Accessed June 25, 2010.
48. Normand SL, Wolf RE, McNeil BJ. Discriminating quality of hospital care in the US. Med Decis Making. 2008;28:308–22.
49. O'Brien S, Shahian DM, DeLong ER, et al. The measurement of adult cardiac surgery quality, 2: statistical considerations in composite measure scoring and provider rating. Ann Thorac Surg. 2007;83(suppl):S13–26.
50. Shahian DM, Edwards FH, Ferraris VA, et al. The measurement of adult cardiac surgery quality, 1: conceptual framework and measurement selection. Ann Thorac Surg. 2007;83(suppl):S3–12.

Downloaded from http://circ.ahajournals.org/ by guest on March 1, 2017

51. Peterson ED, DeLong ER, Masoudi FA, et al. 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.
52. National Quality Forum. Composite measure evaluation framework and national voluntary consensus standards for mortality and safety—composite measures. 2009. Available at: http://qualityforum.org/Publications/Composite_Measures.aspx. Accessed July 10, 2010.
53. National Quality Forum. National voluntary consensus standards for hospital care: outcomes and efficiency phase II. Available at: http://www.qualityforum.org/projects/hospital_outcomes-and-efficiency_II.aspx. Accessed July 10, 2010.
54. Fisher ES, Staiger DO, Bynum JP, Gottlieb DJ. Creating accountable care organizations: the extended hospital medical staff. A new approach to organizing care and ensuring accountability [published online December 5, 2006]. doi:10.1377/hlthaff.26.1.w44. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2131738. Accessed July 10, 2010.
55. O'Kane M, Corrigan J, Foote SM, et al. Crossroads in quality: successful reform must address coverage, costs, and quality at the same time. Health Affairs 2008;27(3):749–758. Available at: http://content.healthaffairs.org/cgi/content/full/27/3/749. Accessed May 25, 2010.
56. Shortell SM, Casalino LP. Health care reform requires accountable care systems. JAMA. 2008;300:95–7.
57. American College of Physicians. Patient-centered medical home. Available at: http://www.acponline.org/running_practice/pcmh/. Accessed July 10, 2010.

KEY WORDS: AHA Scientific Statements ■ health policy and outcome research ■ quality indicators ■ quality measurement

Downloaded from http://circ.ahajournals.org/ by guest on March 1, 2017

Circulation

ACCF/AHA New Insights Into the Methodology of Performance Measurement: A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Performance Measures

WRITING COMMITTEE MEMBERS, John A. Spertus, Robert O. Bonow, Paul Chan, George A. Diamond, Joseph P. Drozda, Jr, Sanjay Kaul, Harlan M. Krumholz, Frederick A. Masoudi, Sharon-Lise T. Normand, Eric D. Peterson, Martha J. Radford and John S. Rumsfeld

Circulation. 2010;122:2091-2106; originally published online November 8, 2010;
doi: 10.1161/CIR.0b013e3181f7d78c
Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2010 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/122/20/2091

Data Supplement (unedited) at:
http://circ.ahajournals.org/content/suppl/2010/11/06/CIR.0b013e3181f7d78c.DC3
http://circ.ahajournals.org/content/suppl/2010/11/06/CIR.0b013e3181f7d78c.DC1
http://circ.ahajournals.org/content/suppl/2010/11/06/CIR.0b013e3181f7d78c.DC2

Downloaded from http://circ.ahajournals.org/ by guest on March 1, 2017

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in *Circulation* can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to *Circulation* is online at:
http://circ.ahajournals.org//subscriptions/

Copyright 2010 by the American College of Cardiology Foundation and
American Heart Association, Inc.

Appendix E: The Challenge of Using Procedures Versus Patient Scenarios in Applying Outcomes as Performance Measures

To amplify the concern of selecting a clinical state as the appropriate denominator of patients in whom to apply an outcome-based performance measure, consider patients with chronic angina and 2-vessel coronary disease at intermediate risk for periprocedural death. Hospital A preferentially referred such patients to coronary artery bypass surgery (CABG), whereas Hospital B referred them for percutaneous coronary intervention (PCI). Such patients would likely lower the average periprocedural mortality rate for both CABG and PCI at Hospital A while increasing them for both CABG and PCI at Hospital B. Given the limitations of risk adjustment and the modest proportion of outcomes explained, it is not possible to rely on risk adjustment alone to completely level the playing field between these 2 hospitals.

This same logic also applies to examination of efficiency by procedures rather than disease states. In using procedures rather than the patient as the denominator for analysis, it turns out that episodes of care, from which costs and efficiency are calculated, are derived from sentinel procedures (eg, CABG) rather than management of patients with a particular condition (eg, coronary artery disease [CAD] requiring coronary revascularization). Consider the same 2 hospitals with different practice styles for coronary revascularization as described above: Hospital A (dotted lines) uses PCI 80% of the time and CABG 20% of the time, whereas Hospital B (solid lines) uses PCI 20% of the time and CABG 80% of the time. Assume that the average cost of PCI is $20,000 at

Copyright 2010 by the American College of Cardiology Foundation and
American Heart Association, Inc.

Hospital A and $15,000 at Hospital B, whereas the cost of CABG is $55,000 at Hospital A and $50,000 at Hospital B.

Using the current efficiency methodology, Hospital B seems far more efficient than Hospital A at managing these 100 patients with acute myocardial infarction because the costs of both procedures are lower. As shown in Figure 1, however, if patients with CAD who require revascularization were used as the denominator (rather than procedures), it would be apparent that Hospital B costs far more than Hospital A. Accordingly, the Task Force on Performance Measures believes that it is a research priority to begin developing the methodology to calculate efficiency at the patient level rather than the procedural level.[1,2]

Copyright 2010 by the American College of Cardiology Foundation and
American Heart Association, Inc.

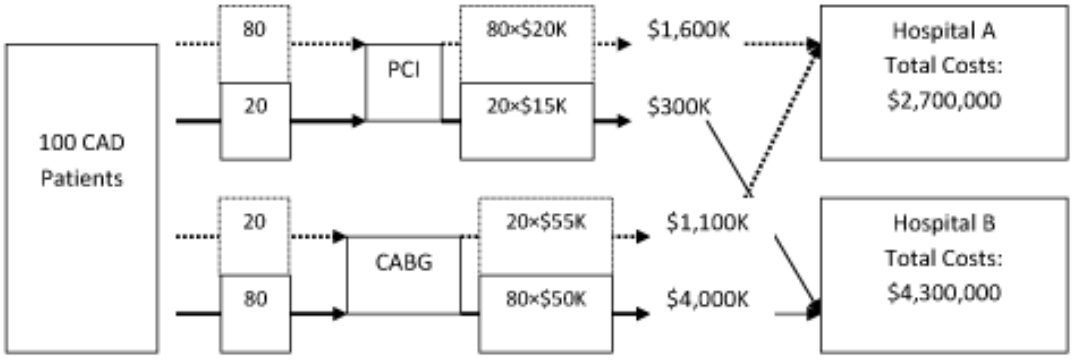

Figure 1. Comparing PCI and CABG Procedures

Copyright 2010 by the American College of Cardiology Foundation and
American Heart Association, Inc.

References

1. Krumholz HM, Keenan PS, Brush JE, Jr., et al. Standards for measures used for public reporting of efficiency in health care: a scientific statement from the American Heart Association Interdisciplinary Council on Quality of Care and Outcomes Research and the American College of Cardiology Foundation. Circulation. 2008;118:1885-93.

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. Circulation. 2005;111:1703-12.

Copyright 2010 by the American College of Cardiology Foundation and
American Heart Association, Inc.

Appendix C: Establishing the Strength of Evidence

Conventional Inference

The evidence used to support a guideline or performance measure is currently exclusively based on "frequentist" approaches to assessing benefit. This approach focuses on *P* values and confidence intervals (CIs).[1-3] A *P* value or observed significance level provides a measure of the inconsistency of the data with respect to a specific hypothesis. For example, in clinical trials investigators prespecify a significance level that represents the maximum probability they will tolerate for rejecting the null hypothesis when it is true. If a maximum risk of 5% of incorrectly rejecting the hypothesis is tolerable, weighed against the consequences of such a wrong decision, then the significance level will be 0.05. However, because *P* values are dependent on sample size, for decision-making purposes, a *P* value of 0.001 does not necessarily provide more support for rejecting the null hypothesis than one of 0.05. CIs, in contrast, provide an assessment of the accuracy of a point estimate of a treatment effect. If investigators were to repeat a study 100 times under exactly the same conditions and compute 95% CIs, 95 of the 100 intervals would contain the true treatment effect.

Despite the historical reliance of the interpretation of clinical trials from a frequentist perspective, complementary approaches to evaluate the strength of evidence are available. Bayesian inference provides one such approach. A formal quantitative algorithm for the construction of guidelines using Bayes' theorem[1,4-8] involves integrating the clinical trial evidence with prior belief. Prior belief can range from a skeptical point

Copyright 2010 by the American College of Cardiology Foundation and
American Heart Association, Inc.

of view embodied in the null hypothesis (the treatment is expected to produce no reduction in risk) to an enthusiastic point of view embodied in the alternative hypothesis (the treatment is expected to produce a specific clinically important reduction in risk).[4,8] In most applications, prior belief is non-informative in the sense that all plausible benefits are permitted. The strength of evidence is often summarized using a Bayes factor which is a measure of how well 2 competing hypotheses predict the data. Suppose in a trial of experimental versus placebo group outcomes, a Bayes factor of 1 is observed. This is interpreted to mean that the evidence supports the null hypothesis as strongly as it does the alternative hypothesis; if the Bayes factor were 1/10 then the evidence supports the null hypothesis (no difference in treatment groups) 1/10 as strongly as it does the alternative. Practical guidelines for interpreting Bayes factors exist.[9] Bayes factors provide a substantially different perspective than *P* values. Specifically, *P* values cannot represent measures of comparative evidence because *P* values, by construction, assume the null hypothesis is true.

Another complementary approach to evaluate the strength of evidence, proposed by Sackett, is a critical evaluation of the CIs.[10] If the summary treatment effect is large, and the CIs are precise, such that even the lower CIs of the effect would be considered clinically important, then the strength of evidence is greater. The decision whether a particular difference is clinically important would ideally depend on the seriousness and the frequency of the outcome of interest.

No particular guidelines exist for deciding that the magnitude of a distinction is “clinically significant” or “practically important.”[2] Consequently, there has been a tendency to equate statistical significance with clinical importance. In some instances,

Copyright 2010 by the American College of Cardiology Foundation and
American Heart Association, Inc.

statistically significant results may not be clinically important (ie, a small magnitude of difference in studies with large sample size) and, conversely, statistically insignificant results do not completely rule out the possibility of clinically important effects (ie, large magnitudes of difference in studies with insufficient power/sample size). Ideally, clinical practice guidelines would explicitly assess both the statistical significance and clinical importance of evidence in order to evaluate the potential of a proposed intervention to improve human health. Those therapies for which there was adequate certainty of benefit as well as good evidence for a clinically important effect would be those most appropriately targeted by performance measurement.

The Task Force on Performance Measures felt that no hard and fast rules of minimal clinically important differences (MCID) could be created outside the context of a particular intervention and outcome. However, converting the benefits of therapy on both survival and health into numbers-needed-to-treat could help writing committees establish a justifiable standard for their use in creating a performance measurement set.

To illustrate the types of data derived from clinical studies, Figure 1 demonstrates the range of effect sizes and CIs that might be obtained relative to a hypothetical MICD. Example A exemplifies a treatment estimate that is “statistically not significant and clinically not important,” because the entire CI lies to the right of MCID and crosses the null line. Example B describes a treatment benefit that is “statistically not significant but may be clinically important” because the CI crosses the MCID as well as the null lines consistent with an indeterminate effect. A “statistically significant but not clinically important” intervention is diagrammed in Example C, where the entire CI lies between the MCID and the null line, whereas a “statistically significant and potentially clinically

Copyright 2010 by the American College of Cardiology Foundation and
American Heart Association, Inc.

important" intervention is shown in Example D, where the CI is centered to the left of MCID and does not cross the null. A "statistically significant and clinically important" treatment is shown in Example E, where the entire CI lies to the left of the MCID line.[2,10] Most clinical trials are only large enough to generate C- or D-type results, that is, better than nothing, while an E-type result would be what is most desired in defining an "established effective therapy". This is, however, most commonly achieved through very large studies, those with very large treatment benefits or through meta-analyses of several comparable trials.[10] Nevertheless, one can reasonably argue that only Class IA recommendations that meet the criterion of "established effective therapy" should qualify as performance measures.

Copyright 2010 by the American College of Cardiology Foundation and
American Heart Association, Inc.

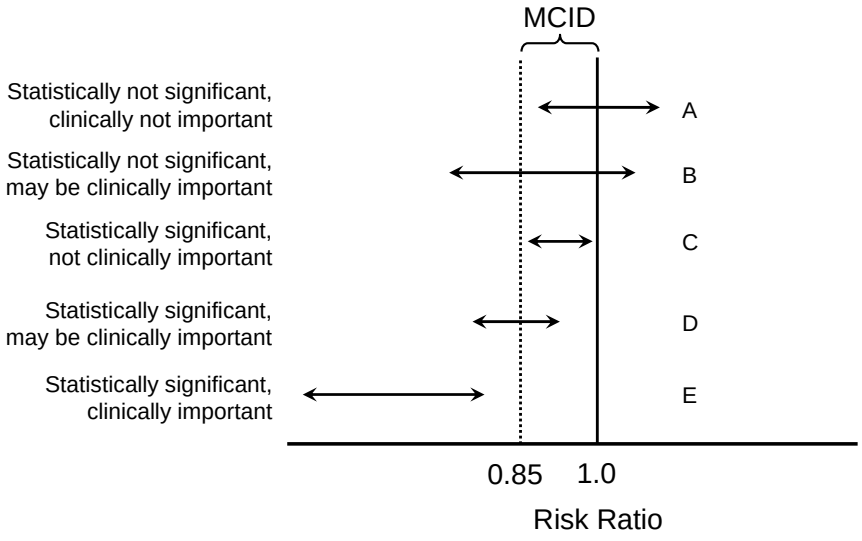

Figure 1. Graphic demonstration of statistically significant and clinically important treatment benefits.[2,10] Five trial results (A to E) and their interpretation with reference to zero effect (a risk ratio of 1.0) and a minimal clinically important difference (MCID) of 15% risk difference (corresponding to a risk ratio of 0.85) are shown. Treatment effects (double arrows) are expressed as 95% confidence intervals.

Copyright 2010 by the American College of Cardiology Foundation and
American Heart Association, Inc.

References

1. Diamond GA, Kaul S. Prior convictions: Bayesian approaches to the analysis and interpretation of clinical megatrials. J Am Coll Cardiol. 2004;43:1929-39.

2. Kaul S, Diamond GA. Trial and error: how to avoid commonly encountered limitations of published clinical trials. J Am Coll Cardiol. 2010; 55:415-27.

3. Sterne JAC. Sifting the evidence-what's wrong with significance tests? Br Med J. 2001;322:226-31.

4. Diamond GA, Kaul S. Bayesian classification of clinical practice guidelines. Arch Intern Med. 2009;169:1431-5.

5. Goodman SN. Toward evidence-based medical statistics, 2: the Bayes factor. Am Intern Med. 1999;130:1005-13.

6. Greenland S. Bayesian perspectives for epidemiological research, I: foundations and basic methods. Int J Epidemiol. 2006;35:765-75.

7. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Bayesian methods in health technology assessment: a review. Health Technol Assess. 2000;4:1-130.

8. Wijeysundera DN, Austin PC, Hux JE, et al. Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials. J Clin Epidemiol. 2009;62:13-21.

9. Kass RE, Raftery AE. Bayes factors. J Am Statistical Assoc. 1995;90:773-95.

10. Hayes R, Sackett D, Guyatt GH, et al. The Principles Behind the Tactics of Performing Therapeutic Trials. Clinical Epidemiology: How to Do Clinical Practice Research. 3rd ed. Philadelphia, Pa: Lippincott Williams & Wilkins; 2006.

Copyright 2010 by the American College of Cardiology Foundation and
American Heart Association, Inc.

Appendix D: The Clinical Rationale for the Use of Exclusions in Performance Measurement

Medical exceptions allow for physician clinical judgment to be factored into performance measurement. No matter how carefully performance measures are constructed, there are always patients for whom the intervention or medication does not apply. For instance, comorbidities such as terminal cancer may make the prescription of certain medications or implementation of some interventions futile. In addition, for a given patient, new evidence available to the physician from clinical trials (or, in the future, pharmacogenomics) may point to a superior alternative to the intervention or drug called for in the measure. No matter how frequently practice guidelines and performance measures are updated, these sorts of situations will always arise.

Clinical judgment is also important when exceptions represent relative contraindications to drugs or procedures. In these instances, patients who qualify for one of these exceptions might still receive the intervention after the physician has carefully considered the entire clinical picture. The patient would then be captured in both the measure's numerator and denominator and would not be excluded from analysis.

Exceptions also create greater comparability among denominator populations. This holds true for medical, patient, and system exceptions. Not only is it desirable to have clinically more homogeneous denominators, but it is also important to "level the playing" field in terms of socioeconomic factors in order not to discriminate against physicians who care for disadvantaged populations and thereby risk aggravating healthcare disparities.[1] Of course, the opposite also holds true when measuring at the

Copyright 2010 by the American College of Cardiology Foundation and
American Heart Association, Inc.

system or societal level in that allowing such exceptions can serve to condone such disparities. In addition, the use of exceptions also reduces the risk of penalizing physicians who care for medically complex patients.

Patient exceptions allow for inclusion of patient preferences in performance measures. Patient autonomy and the principle of respect for persons require that informed patients be allowed to choose whether or not to accept a medical intervention or to take a medication. This, of course, assumes that patients or their surrogates are well informed about their options. If a patient exception is identified, the physician should document in the record that the patient or patient's representative was adequately informed about the recommended test, procedure, or treatment, including its risks and benefits. The physician should also document that alternative approaches were also discussed. However, requiring all of the elements of informed consent as is done in clinical trials would be impractical in the practice setting and would likely add little value.

Finally, exceptions help to avoid the moral hazard of providing incentives for care that would be inappropriate for a given patient. Without the ability to use exceptions in accountability measures, physicians might be tempted to employ diagnostic and therapeutic options that are not in the patient's best interests in order to "meet the performance measure."

Special Considerations in the Use of Exceptions

A number of special considerations about the use of exceptions are worthy of comment. First, not all performance measures require exceptions in all 3 categories, and some do not require exceptions at all. For instance, it would generally not be appropriate

Copyright 2010 by the American College of Cardiology Foundation and
American Heart Association, Inc.

to apply exceptions to structural measures, such as the presence of accreditation of an ambulatory imaging center. In addition, it would not be necessary to use exceptions in measures that call for interventions or treatments to be "considered" rather than actually implemented, because a physician could always consider an intervention or treatment no matter what the clinical, patient, or socioeconomic circumstances. In other words, the exception decision is "built into" the consideration process.

Secondly, whenever exceptions are allowed, physicians should be required to document in the medical record the rationale for their use in specific cases. The documentation need not be extensive but should make it clear that the action called for in the measure was indeed considered but not performed, and it should delineate the alternative course of action to be taken.

The third consideration is the classification of economic exceptions. One could make an argument that such exceptions are not system reasons for excluding patients from a measure as they are classified in the American Medical Association Physician Consortium for Performance Improvement construct[2] but are instead patient reasons. Some patients, for instance, may not have insurance because they chose to spend their money on other discretionary items even though they could have afforded the premiums. Nevertheless, because of the multiple factors that have an impact on insurance coverage, most of which are beyond the control of the individual patient, it would seem appropriate, granted a bit arbitrary, to classify these as system exceptions. Having said that, there will be instances where it is clear that the reason a given patient declines a test or treatment is economic but not related to affordability. ("I prefer to spend my money on cigarettes and not on copays for my cholesterol drug.") In such instances the exception should be classified in the Patient category.

Copyright 2010 by the American College of Cardiology Foundation and
American Heart Association, Inc.

Fourth is the controversial issue of physician "gaming" of the system. Some—generally payers—express angst that allowing physicians to use exceptions would give them the opportunity to use those exceptions inappropriately to make their performance appear better than it actually is. This, of course, is an issue only when measures are used for physician accountability purposes. Although at first glance it might appear that physicians could indeed be tempted to pad their numbers when used for public reporting or reimbursement purposes, a closer examination raises significant questions over whether or not this is likely to happen with any frequency.

Most interventions, medications, and procedures around which performance measures are built are those in which there is a significant gap in physician or healthcare team performance. This sort of gap generally arises because providers either do not know the requirement or forget about it. In either instance, the provider is not likely to list an exception to the measure. Exceptions will be used when a physician recognizes that a patient otherwise qualifies for the performance measure. In other words, the physician is aware of the requirement in the instances where exceptions are invoked. One can only speculate why a physician would inappropriately use an exception in this circumstance rather than provide the service called for in the measure. It does not seem that such inappropriate use of exceptions would be very common.

A final point is that not much is known about the use of exceptions in physician performance measurement because this remains a nascent field and there has been little robust testing of measures. Cardio-HIT is a study funded by the Agency for Healthcare Research and Quality of the use of American College of Cardiology Foundation/American Heart Association/American Medical Association Physician

Copyright 2010 by the American College of Cardiology Foundation and
American Heart Association, Inc.

Consortium for Performance Improvement measures in the electronic health records of several large physician groups. Phase II of this study, which should be published in the near future, deals specifically with exception reporting and should begin to provide the sort of data that measures developers will need to refine the use of exceptions in their future work.

References

1. Werner RM, Asch DA, Polsky D. Racial profiling: the unintended consequences of coronary artery bypass graft report cards. Circulation. 2005;111:1257-63.

2. Specifications and categorization of measures exclusions: recommendations to Physician Consortium for Performance Improvement work group. Available at: http://www.ama-assn.org/ama1/pub/upload/mm370/exclusions053008.pdf. Accessed June 25, 2010.