Exploring the Behavior of Hospital Composite Performance Measures
An Example From Coronary Artery Bypass Surgery

Sean M. O’Brien, PhD; Elizabeth R. DeLong, PhD; Rachel S. Dokholyan, MPH; Fred H. Edwards, MD; Eric D. Peterson, MD, MPH

Background—Composite scores that combine several performance measures into a single ranking are becoming the accepted metric for assessing hospital performance. In particular, the Centers for Medicare & Medicaid Services Hospital Quality Incentive Demonstration (HQID) project bases financial rewards and penalties on these scores. Although the HQID composite calculation is straightforward and easily understood, its method of combining process and outcome measures has not been validated.

Methods and Results—Using data on 530 hospitals from the Society of Thoracic Surgeons National Cardiac Database, we replicated the HQID methodology with 6 nationally endorsed performance measures (5 process measures plus survival) for coronary artery bypass surgery. Composite scores were essentially determined by process measure performance alone; the survival component explained only 4% of the composite score’s total variance. This result persisted even when the survival component was allowed a 5-fold greater weighting in the composite summary. The popular “all-or-none” measurement approach was also dominated by the process component. Substantial disagreement was found among hospital rankings when several alternative methods were used; up to 60% of hospitals eligible for the top financial reward under HQID would change designation depending on the composite methodology used. The application of a simple statistical adjustment (standardization) to each method would provide more consistent results and a more balanced assessment of performance based on both process and outcomes.

Conclusions—Existing methods used to create composite performance measures have remarkably different weighting of process versus outcomes metrics and lead to highly divergent provider rankings. Simple alternative methods can create more balanced process-outcome performance assessments. (Circulation. 2007;116:2969-2975.)

Key Words: program evaluation coronary artery bypass quality of health care

C omposite measures of healthcare quality are increasingly used to summarize healthcare performance and enhance accountability of providers and hospitals. Often, multiple individual performance measures are used to assess care for the same disease condition. A composite measure combines these multiple related indicators into a single summary that is often simpler to understand and communicate. A recent report from the Institute of Medicine identifies composite measurement as a useful tool in the effort to measure and improve the country’s healthcare quality and calls for research to address the validity and reliability of such indicators.1

Editorial p 2897
Clinical Perspective p 2975

A prominent example of the use of composite measures can be found in the Centers for Medicare & Medicaid Services (CMS) Hospital Quality Incentive Demonstration (HQID) project.2 Hospitals participating in this pay-for-performance program voluntarily submit data on a number of quality indicators for patients with heart attack, heart failure, pneumonia, coronary artery bypass surgery (CABG), and hip and knee replacements. The evaluation of these data involves the calculation of a composite score for each clinical condition. On the basis of these composite scores, CMS pays financial bonuses to the top-ranking 20% of hospitals (2% bonus for best decile; 1% bonus for second-best decile) and penalizes hospitals that fail to meet minimum performance thresholds. Although this particular initiative is voluntary, the growth of such pay-for-performance programs is supported by Congress,3 the Medicare Payment Advisory Council,4 and the Institute of Medicine.5

Although composite measures are increasingly used, there are multiple alternatives for how such measures are formed. For example, the CMS HQID method combines multiple process and outcome measures by assigning a straightforward equal weight to each individual measure.2 An alternative simple method advocated by the Institute of Medicine has been termed “all-or-none measurement,” because providers
are only given credit in the composite if all individual measures are achieved.\textsuperscript{1,6} Although these and other methods exist, there is a paucity of published data investigating how existing popular methodologies behave when applied in practice. As we will demonstrate, seemingly similar composite score methodologies can place remarkably different weights on the relative importance of the same process and outcome metrics. Even methods that appear intuitive, such as the HQID “equal weight for each measure” approach, may in fact weight process and outcome measures in an unexplained manner. These features are not obvious until one explores the behavior of the composite score empirically.

The goal of the present study was to explore the potential implications and consequences of the methodology adopted by CMS for creating composite scores in the HQID pilot pay-for-performance program. Our specific objectives were (1) to elucidate the implications of the HQID method of weighting process and outcome measures, (2) to explore commonly proposed alternatives to the HQID composite methodology, and (3) to determine whether provider rankings differ depending on the composite method used. We selected CABG as our example because this procedure has been at the forefront of provider profiling for more than a decade, has well-defined and accepted measures that pertain to processes of care and risk-adjusted outcomes,\textsuperscript{7} and has known wide variability in provider performance.\textsuperscript{8–10}

**Methods**

**Data Sources**

The Society of Thoracic Surgeons (STS) National Cardiac Database collects demographic, process, and outcomes data for all adult cardiac procedures from >700 surgical practices that perform CABG surgery around the United States. Data from these database participants are submitted semiannually to the Duke Clinical Research Institute, with whom the STS contracts for data warehouse and analysis functions. A series of data quality checks are performed before a site’s data are aggregated into the national sample. Although participation in the STS database is voluntary, the completeness of the data has been confirmed in local and regional audits,\textsuperscript{11,12} as well as an ongoing national audit program.

**Patient Population**

To examine the empirical behavior of the HQID composite method, we analyzed all isolated CABG procedures in the year 2004 for the 530 STS participants that had at least 95% complete data for all of the 5 process measures under study. Patients undergoing CABG combined with valve surgery and other concomitant procedures were excluded, which left 133,319 patients.

**Quality Measures**

In June 2005, the National Quality Forum (NQF) published a set of 21 structure, process, and outcome measures that were endorsed by member organizations for evaluation of the quality of care of adult cardiac surgery providers.\textsuperscript{7} All but 5 of these measures correspond to items collected in the STS database. The present report focuses on the subset of 5 NQF process measures that pertain specifically to isolated CABG surgery, as well as a single risk-adjusted outcome that pertains to CABG: operative mortality. The 5 process measures are preoperative β-blockade, use of internal mammary artery (IMA), discharge antiplatelet medications, discharge β-blockade treatment, and discharge antilipid treatment. Operative mortality was defined as death during the same hospitalization as surgery or after discharge but within 30 days of surgery.

Data for each process measure consist of the number of patients who were eligible to receive the care process during the study period (denominator) and the number of these patients who actually received the care process (numerator). The eligible populations varied from measure to measure, according to NQF measure specifications. In the case of discharge medications, we excluded patients who died before discharge, and for IMA, we excluded patients with a previous CABG surgery. All patients in the study were included in the denominator of the mortality measure, and missing mortality status was imputed to “alive.” Missing data for process measures were imputed by assuming that the patient did not receive the indicated care process. The adoption of this convention has the potential effect of penalizing sites for not completely entering the process data and hence encouraging completeness. All sites in the present study had at least 95% complete data for these measures, and hence, the impact of missing data is minimal.

**Composite Score Methodologies**

**Method A: The HQID Composite Method**

Identical methodology is used in HQID to construct composite scores for heart attack, CABG, and hip and knee replacements. In each case, the HQID composite quality score comprises 2 separate components: a composite process score and a risk-adjusted outcome score.\textsuperscript{2} The process component is combined with the risk-adjusted outcome component to arrive at an overall composite quality score. To create the process component, HQID uses the opportunity-based approach that has been credited to Scinto and coworkers.\textsuperscript{13} An “opportunity” is defined as an instance in which a patient is eligible for a certain process measure; by this definition, a single patient can contribute up to 5 opportunities in the STS database. The total number of eligible opportunities for all individuals at a hospital becomes the denominator for the process component, and the number of successful opportunities becomes the numerator. The final process composite score is defined as:

\[
\text{Process component} = 100 \times \frac{\text{sum of individual process measure numerators}}{\text{sum of individual process measure denominators}}
\]

In HQID, mortality rates are adjusted for case mix by comparing each hospital’s observed mortality rate with its expected mortality rate. The expected mortality rate is calculated from a risk-adjustment model and depends on the hospital’s case mix. To express mortality performance in positive terms, mortality rates are converted to survival rates (ie, observed survival rate=1−observed mortality rate; expected survival rate=1−expected mortality rate.) Then, a “survival index” is created by dividing the hospital’s observed survival rate by its expected survival rate and multiplying by 100:

\[
\text{Survival index} = 100 \times \frac{1−\text{observed mortality rate}}{1−\text{expected mortality rate}}
\]

We defined mortality as “operative mortality” (proportion of isolated CABG patients who die during the same hospitalization as surgery or after discharge but within 30 days of surgery) and calculated expected rates of operative mortality using the previously published STS CABG operative mortality risk model.\textsuperscript{14} The use of the STS CABG risk-adjustment model was consistent with NQF measure specifications.

To form the composite, the HQID method weights each component in proportion to the number of items it comprises, based on the premise of “equal weight for each measure.”\textsuperscript{13} In our example, there are 5 process measures and 1 risk-adjusted outcome measure being evaluated. Thus, the composite process score counts as 5/6 of the overall composite score, and the survival index contributes 1/6 of the HQID composite score. Hence: HQID composite score=(5/6×process score)+(1/6×survival index).
Method B: “All-or-None” Method for Process Component
As a first alternative method, we replaced the HQID “opportunity model” for aggregating process measures with an alternative approach known as the “all-or-none” method.12 The all-or-none process score was defined as the percent of patients who received all of the NQF care processes for which they were eligible. Patients were considered to be eligible for IMA unless they had a previous CABG and were eligible for discharge medications unless they died before discharge. All patients were eligible for preoperative β-blockers. We left the calculation of the outcome component unchanged and used the original HQID weighting of process versus outcome. Thus: Composite score = ((5/6 × all-or-none process score) + (1/6 × survival index)).

Method C: Equal Weight for Process and Survival Components
Although mortality was the only outcome evaluated in the present study, other outcome measures could potentially be included in the composite to boost the weight of the outcomes component. To simulate the effect of increasing the number of outcomes, we multiplied the weight of the mortality component by 5, thus giving equal weight to process and outcomes. This modified composite score was defined as: Composite score = (1/2 × process score) + (1/2 × survival index).

Method D: Mortality Instead of Survival
As a second sensitivity analysis, we considered substituting mortality instead of survival in the HQID composite score described previously. The mortality index was defined as the observed mortality rate divided by the expected mortality rate, multiplied by 100:

\[
\text{Mortality index} = 100 \times \frac{\text{observed mortality rate}}{\text{expected mortality rate}}.
\]

The mortality index and survival index contain virtually identical information about a provider’s risk-adjusted mortality rate, but the mortality index has a larger SD. As shown below, this difference has an impact on the behavior of the composite score. Because larger values of the mortality index imply worse performance, we multiplied the mortality index by minus 1 (−1) before combining it with the process component. We did this so that larger numbers of the outcomes component score would imply better performance (consistent with the process component). Following the principle of “equal weight for each measure,” the final composite score was created by weighting the process and outcome components by 5/6 and 1/6, respectively. Hence: Composite score = [5/6 × process score] + [1/6 × (−1) × mortality index].

Method E: Standardization to Adjust for Unequal Measurement Scales
Textbooks that discuss composite measures emphasize the importance of considering the measurement scale of each variable before assigning them weights for a composite.13 In our example, the survival index is unitless and tends to be clustered in a narrow interval around 100. In contrast, processes adherence is measured on an interquartile range (IQR; 25th to 75th percentile) of 119 to 51%; discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 64.3% to 79.4%) and discharge antiplatelet medication (IQR = 19%). Among outcomes, the mortality index (IQR = 53.3 to 123.4) had a much wider range than the survival index (IQR = 99.4 to 101.3).

Empirical Behavior of Composite Methods
As illustrated in Figure 1A, a provider’s performance on the HQID-based composite score was almost entirely determined by the participant’s performance on the process component. This is reflected in the nearly perfect linear relationship between the provider’s process score and the overall composite (explained variation >99%). In contrast, the hospital’s survival index contributed negligible information to the composite metric (Figure 1B; explained variation = 4%). Results were similar when the all-or-none method was used to create the process component (method B). Process adherence accounted for >99% of the variation in the overall composite score compared with 2% of the variation explained by survival.

As shown in Figure 2, different items contributed different amounts to the variation in the HQID composite score. Among process measures, the amount of variation explained by individual measures ranged from 52% for discharge β-blocker medication to 16% for IMA usage. The amount of variation explained by the survival index (4%) was less than any single process indicator. Results were similar when the all-or-none method was used to create the process component (discharge β-blockers = 42%; discharge antilipid medication = 50%; preoperative β-blockers = 51%; discharge antiplatelet medication = 19%; IMA usage = 13%; and survival = 2%).

Results
Participant sample sizes ranged from 11 to 1513 patients with an interquartile range (IQR; 25th to 75th percentile) of 119 to 323 patients (Table 1). Among the process measures, performance was most variable for preoperative β-blockade (IQR = 64.3% to 79.4%) and discharge antilipid medications (IQR = 67.3% to 88.8%). Among outcomes, the mortality index (IQR = 53.3 to 123.4) had a much wider range than the survival index (IQR = 99.4 to 101.3).

Statistical Analysis
For each composite method, we addressed the following questions: (1) To what extent is each composite score driven by process measures versus outcomes? (2) To what extent does each individual indicator contribute to the overall composite? (3) Are inferences about a provider’s performance sensitive to variations in the method used to create the composite score? As a preliminary step, we summarized the distribution of each individual quality indicator across hospitals in the study population.

For questions 1 and 2, we calculated composite scores for each of the 530 hospitals and then determined the proportion of variation in the overall composite score that was explained by each component (the process component and the survival component) and by each individual process indicator. The proportion of variation explained by a component or indicator is equal to the squared Pearson correlation coefficient (r²) between that component or indicator and the overall composite. The total explained variation does not sum to 100% because the indicators are correlated.

For question 3, we applied each of the alternative composite calculations to the 530 STS participants and compared the results. Agreement between the HQID method and each alternative method was quantified by Spearman’s rank correlation coefficient. To explore agreement in the context of pay-for-performance, we assigned providers to tiers that were loosely based on the reward structure used by HQID. The tiers were best decile, second-best decile, middle 6 deciles, next-to-lowest decile, and lowest decile. We then assessed how often a provider’s rank changed from a bonus grouping (1st or 2nd decile) to middle or low based on alternative methods of calculating the composite score.

The authors had full access to the data and take full responsibility for its integrity. All authors have read and agree to the manuscript as written.
Contrary to our expectations, the HQID-method composite was still largely determined by the process component even when the process and outcome components were weighted equally (method C; variation explained by process component 95%). The amount of variation explained by the survival component was still less than any single process measure (discharge β-blockers 50%; discharge antilipid treatment 48%; preoperative β-blockers 43%; discharge antiplatelet medication 26%; IMA usage 16%; and survival 14%). This finding is explained by the fact that the process component score has a larger SD than the survival index (6.6 versus 1.6). When 2 variables are averaged equally, the contribution of any single variable is proportional to the square of its SD. In our example, the process component has a larger SD, so the process component dominates.

In contrast, when we replaced the survival index with a mortality index, we found that the modified HQID composite score was now weighted heavily toward outcomes (method D; explained variation = 73% for mortality component and 33% for process component). Although the mortality and survival indexes contain virtually identical information, the mortality index has a much larger SD (57.8 versus 1.6). As a result, the mortality index contributes more variation to the overall composite. Compared with the process component score, the mortality index was assigned less weight (1/6 versus 5/6, a 5-fold difference) but had a larger SD (57.8 versus 6.6, a 10-fold difference). The net effect of the unequal weights and SDs was to allow the outcome component to contribute 2-fold variation. The contribution of individual process measures ranged from 18% for discharge β-blockers to 6% for IMA usage. Finally, when we standardized the process and survival components to correct for their unequal SDs (method E), the process and survival components contributed equally to the composite score (explained variation = 57% for each component).

Table 1. Distribution Across 530 Sites of Sample Sizes and Performance on Individual Quality Measures and Quality Domain Components

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Mean</th>
<th>SD</th>
<th>25th Percentile</th>
<th>75th Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of patients</td>
<td>252</td>
<td>199.7</td>
<td>119</td>
<td>323</td>
</tr>
<tr>
<td>No. of care opportunities</td>
<td>1230</td>
<td>975.7</td>
<td>584</td>
<td>1567</td>
</tr>
<tr>
<td>Process measures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preoperative β-blockers, %</td>
<td>71.2</td>
<td>12.0</td>
<td>64.4</td>
<td>79.4</td>
</tr>
<tr>
<td>IMA, %</td>
<td>91.9</td>
<td>6.7</td>
<td>89.7</td>
<td>96.2</td>
</tr>
<tr>
<td>Discharge antiplatelet medications, %</td>
<td>93.2</td>
<td>6.3</td>
<td>91.0</td>
<td>97.4</td>
</tr>
<tr>
<td>Discharge β-blockers, %</td>
<td>82.4</td>
<td>11.2</td>
<td>76.6</td>
<td>90.5</td>
</tr>
<tr>
<td>Discharge antilipid medications, %</td>
<td>76.8</td>
<td>14.9</td>
<td>67.3</td>
<td>88.8</td>
</tr>
<tr>
<td>HQID process component score</td>
<td>83.0</td>
<td>6.6</td>
<td>78.9</td>
<td>87.9</td>
</tr>
<tr>
<td>All-or-none process, %</td>
<td>44.4</td>
<td>15.1</td>
<td>33.2</td>
<td>55.5</td>
</tr>
<tr>
<td>Risk-adjusted outcome</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Survival index*</td>
<td>100.3</td>
<td>1.6</td>
<td>99.4</td>
<td>101.3</td>
</tr>
<tr>
<td>Mortality index†</td>
<td>91.4</td>
<td>57.8</td>
<td>53.3</td>
<td>123.4</td>
</tr>
</tbody>
</table>

*Survival index = 100 × observed survival rate/expected survival rate. †Mortality index = 100 × observed mortality rate/expected mortality rate.

**Figure 1.** Relationship of process and survival components with HQID-based overall composite score. Each circle represents 1 of the 530 participants.

**Figure 2.** Contribution of individual indicators to overall HQID composite score. Total explained variation does not sum to 100% because items are correlated. D/C indicates discharge; Preop, preoperative.
Agreement Between Rankings Based on Alternative Methods

There was close agreement between the HQID method composite and the variation that used the all-or-none method for process measures (rank correlation = 0.98). Among the 53 sites that ranked in the top decile on the basis of the HQID method, 43 of these providers (81%) also ranked in the top decile by the all-or-none composite, and none of them ranked below the second decile. Thus, all of them would qualify for a bonus under the HQID incentive scheme regardless of the choice between the HQID composite versus all-or-none measurement. When the process component was examined separately, the rank correlation coefficient between the original HQID process score and the all-or-none process score was 0.98, which indicates high agreement between the 2 methods with respect to the measurement of process performance.

Reweighting of the HQID composite (method C) to place equal weight on process and survival components did not cause hospital rankings to change substantially (rank correlation = 0.98). No site was ranked in the top 20% by the original HQID method and bottom 20% by the reweighted method.

Agreement was considerably less when the HQID composite was modified by replacing the survival index with a mortality index (method D). Of the 53 participants in the top decile according to the HQID-based composite, only 21 of these participants (39.6%) ranked in the top performing decile according to the modified version. In other words, 60% of the hospitals that would be eligible for the top pay-for-performance bonus under HQID would change designation based on the choice of composite methodology. Two participants that would have received a 2% bonus (top decile) and 5 participants that would have received a 1% bonus (second decile) based on the HQID reward structure would actually have been in the bottom 2 deciles according to the modified HQID composite score. Examples of these 7 participants with almost opposite performance ratings depending on composite methodology used are displayed in Table 2. Although each of these would have received a financial bonus with the HQID method, they also had mortality rates that were nearly twice those expected. For 3 of these providers, the excess mortality was statistically significant (P < 0.05 based on 1-sided exact binomial test), and for 1 provider, the excess mortality was marginally significant (P < 0.10). Hence, there is some risk that the HQID method will reward process while ignoring poor outcomes.

There was also imperfect agreement between the HQID-method composite and the variation in which the process and outcome components were first standardized by dividing by their SDs and then weighted equally (method E). One site that ranked in the second highest decile by the HQID composite ranked in the second lowest decile based on the modified version of the composite.

Reweighting the HQID composite to place equal weight on process and survival did not cause hospital rankings to change substantially (rank correlation = 0.98). Again, no site was ranked in the top 20% by the HQID method and bottom 20% by the reweighted method.

Discussion

Quality is a difficult concept to define, measure, and compare. Yet, in order to reward high quality, definitions must be created, measurements taken, and comparisons made. At this point, there is general agreement on most of the metrics that define quality of care for CABG patients but much less agreement about the best way to form composite scores. The present study demonstrates that existing and seemingly similar methods used to define composite measures can have a significant impact on performance rankings. Even minor variations, such as rescaling measures or considering the opposite of survival (mortality) as the outcome, can profoundly affect which providers are considered “top” versus “bottom” performers.

The method that CMS uses to create composite scores for HQID is simple and seemingly transparent, and as a result, it currently serves as a model for pay-for-performance programs. However, when the HQID methodology was applied to NQF CABG measures, we found that the resulting composite score did not behave as expected: The outcome component contributed virtually no information to the composite score. This did not occur simply because there were 5 process measures and only 1 outcome. In fact, process measures still dominated the composite even when we increased the weight of the outcomes component 5-fold to equal the number of process measures. And survival still
contributed less variation to the composite than any single process item. This occurred because the process component is measured on a scale that has a much larger SD.

Although there is recognized value in adherence to recommended processes, there is strong justification for integrating outcomes into the comprehensive assessment of hospital performance. This point is emphasized by Porter and Teisburg in their assessment of the wrong competitive forces defining the current healthcare shortfalls. They propose the development of information systems that allow for the direct risk adjustment of patient outcomes as a means for meaningful competition.

In addition to the HQID “opportunity model” for combining process measures, we also examined an alternative advocated approach known popularly as “all-or-none measurement.” This method of combining process measures was advocated on the grounds that it promotes an appropriately high standard of excellence. Although the HQID method and all-or-none method are different conceptually, we found that the choice between these 2 methods of combining process measures made relatively little difference for actual hospital CABG rankings (rank correlation = 0.98). This high level of agreement is consistent with previously reported results from the same STS data set.

Although the method of combining process measures made little difference for performance rankings, other small variations had a large impact. Merely reversing the outcome assessment from survival to mortality ratios resulted in some providers switching from best to worst performers. Whereas the HQID method was dominated by process measures, the HQID variation based on mortality was largely driven by outcomes. As mentioned previously, the difference arose because the HQID method does not adjust for the unequal measurement scales of the process and outcome component scores. Simply rescaling the outcome and process component scores to a common measurement scale would prevent this inconsistent behavior.

The results of the present study illustrate important features of composite scores. First, the present data indicate that the choice of methodology can have a potentially large impact on hospital rankings based on composite scores. Although the HQID and all-or-none methods behaved similarly, other related approaches produced highly discrepant hospital rankings. Second, the present data indicate the potential risk of relying on a default objective method (eg, “equal weights”) to determine the appropriate weighting of process versus outcome measures. Although a casual observer might assume that equally weighted items would contribute roughly equal information to the HQID composite score, this proved not to be the case. Process measures dominated because they are measured on a scale with a much larger SD. In general, the choice of equal weights for a composite is only meaningful if the different items have a compatible measurement scale or if the items have been normalized to have a similar SD. Finally, the present study indicates the importance of conducting an empirical validation to assess the properties of a composite score. Although the HQID composite was simple to calculate, its properties were not obvious until we explored the method empirically.

The results of the present study should not be extrapolated beyond what was actually studied. We focused on a single therapeutic area (isolated CABG) and a single measure set (NQF-endorsed CABG measures). The HQID methodology may perform differently when applied in other contexts. However, some important features of our example (several process measures with high performance and a limited set of relatively rare outcomes events) are common to other conditions (eg, acute myocardial infarction, stroke, and heart failure). Because we did not apply the HQID methodology to the measure set that is currently used by CMS to evaluate CABG surgery, we are unable to assess the validity of the actual score that is currently used for the distribution of pay-for-performance incentives. The present study merely indicates the potential risk of relying on “equal weight” to produce a composite score with desirable properties.

Several important issues were not addressed in this report. First, the choice of which individual indicators to include or exclude from a composite potentially has a large impact on hospital ratings, but this issue was beyond the scope of the present study. These issues have been addressed to date by the NQF and other groups. In addition to measuring process and outcomes, the Institute of Medicine has recommended a comprehensive approach to quality measurement that also includes measures of efficiency, equity, and patient-centered care. Second, the method of handling missing data may impact hospital ratings, but this was not investigated because data were quite complete (>95%). We focused specifically on the behavior of alternative weighting methods and did not address other aspects of the validity of composite scores. Third, there are multiple possible alternative composite measurement methodologies, and we chose to consider the most commonly proposed and simple alternatives. Fourth, although the present study indicates the potential behavior of composite methods, the decision of whether to combine process and outcomes into a global composite and/or the “best” weighting of these inputs is subjective.

Neither the HQID method nor the variations we explored involve sophisticated statistical modeling, such as the type recommended by O’Brien et al. In these more complex analyses, hierarchical models are used to produce “shrunk” estimates that more reliably reflect each provider’s long-run performance. These shrunken estimates can be combined with a simple weighting scheme that is “balanced” in its weighting of process and outcome measures. Although such model-based approaches are desirable from a statistical perspective, straightforward methods are likely to continue to appeal to some groups, especially those that do not have the resources to apply more intensive methods. On the basis of the data from the present study, there do appear to be challenges to the use of certain simple methods such as the CMS HQID “equal weights for each measure” method. Yet, as we have shown, there are simple modifications to these approaches that can result in a more balanced assessment of provider performance that reflects process as well as outcomes.

Conclusions

Although composite measurement offers a means of summarizing provider performance, the construction of valid and
reliable composite measures is challenging. Because there is no single “right way” to create such a score, a variety of methods should be compared and the rationale for the chosen method documented. Empirical testing can make the behavior of the selected composite score transparent, so that raters and providers can understand exactly what is being measured and rewarded.

Sources of Funding
This work was supported by Duke Clinical Research Institute and the STS National Cardiac Surgery Database.

Disclosures
None.

References

CLINICAL PERSPECTIVE
Numerous process and outcome performance indicators have been developed to assess the quality of healthcare provided by hospitals. To simplify reporting, individual indicators are often rolled up into a single summary “composite” score, which is then used to rank and compare hospitals. The calculation of a composite score, however, requires the assignment of relative weights to the individual performance indicators that contribute to the composite. We undertook this study to explore the implications of the method of weighting process and outcome measures when creating composite scores for bypass surgery. Using data on 530 hospitals from the Society of Thoracic Surgeons National Cardiac Database, we first replicated the methodology that was used by Medicare in its Hospital Quality Incentive Demonstration (HQID) pay-for-performance project. We replicated the HQID methodology by assigning equal mathematical weighting to 6 nationally endorsed performance measures (5 process measures plus survival). Owing to larger standard deviations in the process measures, this approach led to a composite score that was essentially determined by process performance alone, with risk-adjusted mortality accounting for only 4% of the total variance of the composite score. There was also substantial disagreement between hospital rankings based on other alternative methods such as “all-or-none” weighting methodologies. Up to 60% of hospitals eligible for the top financial reward under HQID would change designation depending on the composite methodology used. These findings highlight the importance of validating composite scores and the need for empirical testing to elucidate the implications of the chosen weighting of individual indicators.
Exploring the Behavior of Hospital Composite Performance Measures: An Example From Coronary Artery Bypass Surgery

_Circulation_. 2007;116:2969-2975; originally published online December 3, 2007; doi: 10.1161/CIRCULATIONAHA.107.703553
_Circulation_ is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2007 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/116/25/2969

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/