Empirically Derived Composite Measures of Surgical Performance

Douglas O. Staiger, PhD, Justin B. Dimick, MD, MPH,† Onur Baser, PhD,† Zhaohui Fan, MS,† and John D. Birkmeyer, MD†*

Background: Individual quality measures have significant limitations for assessing surgical performance. Despite growing interest in composite measures, empirically-based methods for combining multiple domains of surgical quality are not well established.

Objective: To develop and validate a composite measure of surgical performance that best describes variation in hospital mortality rates and forecasts future performance.

Research Design: Using the national Medicare claims database, we identified all patients undergoing aortic valve replacement in 2000 to 2001 (n = 53,120). To serve as input variables, we identified hospital-level predictors of mortality with aortic valve replacement, including hospital volume, complication rates, and mortality with other procedures. Hospital-specific predicted mortality rates were then determined using Bayesian-derived modeling techniques and assessed against subsequent hospital mortality (2002–2003).

Results: Our composite measure explained 78% of the variation in aortic valve replacement mortality rates (2000–2001). The most important input variables were hospital volume, mortality with aortic valve replacement, and mortality for other high-risk cardiac procedures. The composite measure forecasted 70% of future hospital-level variation in mortality rates (2002–2003), and was substantially better in this regard than individual measures. Hospitals scoring in the bottom quintile on the composite measure in 2000 to 2001 had 2-fold higher mortality rates in 2002 to 2003 than hospitals in the top quintile (adjusted odds ratio, 1.97; 95% CI, 1.73–2.23).

Conclusions: Compared with individual surgical quality indicators, empirically derived composite measures are superior in explaining variation in hospital mortality rates and in forecasting future performance. Such measures could be useful for public reporting, value-based purchasing, or benchmarking for quality improvement purposes.

From the *Department of Economics and the Dartmouth Institute for Health Policy and Clinical Practice, Dartmouth College, Hanover, New Hampshire; and †Michigan Surgical Collaborative for Outcomes Research and Evaluation, Department of Surgery, University of Michigan, Ann Arbor, Michigan.

Supported by the National Institute on Aging (1R21AG027819-01). The views expressed herein do not necessarily represent the views of Center for Medicare and Medicaid Services or the US Government.

Reprints: Justin B. Dimick, MD, MPH, Department of Surgery, University of Michigan, M-SCORE offices, 211 N. Fourth Avenue, Suite 201 and 202, Ann Arbor, MI 48104. E-mail: jdimick@umich.edu.

ISSN: 0025-7079/09/4702-0226

Key Words: quality, surgery, composite, mortality

(*Med Care* 2009;47: 226–233)

There is growing demand for better hospital quality indicators for surgery. Patients, families, and referring physicians are looking for information about best hospitals for selected procedures.[1,2] Measures that best identify top performers would be useful for clinical leaders and hospital administrators looking for benchmarks to guide their quality improvement efforts. Finally, payers and policy makers need good measures for their public reporting and value-based purchasing initiatives in surgery.[3]

Unfortunately, currently available quality indicators in surgery, whether based on structure, process, or direct outcome measures, all have significant limitations.[1,4] For example, hospital volume, the primary focus of many selective referral initiatives, seems to be important for a relatively small number of high-risk procedures and does not reliably predict performance for individual hospitals. Process measures, the primary basis of new pay-for-performance plans launched by the Center for Medicare and Medicaid Services, often relate to secondary outcomes (eg, prophylactic antibiotics for surgical site infection) and often fail to explain observed variations in hospital mortality rates.[5,6] Finally, because of sample size limitations, direct outcome measures, like risk-adjusted morbidity and mortality, are often imprecisely estimated and do not yield a reliable estimate of hospital performance for most procedures.[7]

Composite measures, which combine information from multiple quality domains into a single summary measure, have the potential to obviate many of these limitations associated with individual quality indicators. Along with their growing use in ambulatory and preventative care, composite measures are becoming increasingly popular for assessing surgical performance. For example, CMS' pay-for-performance plan for coronary artery bypass graft surgery is based on a composite of 7 specific processes of care and outcome measures.[8] However, the validity of these new composite measures has not been well established. Among other issues, quality indicators are often combined without regard to their relative impact on patient outcomes, or else weighted according to expert opinion and consensus. Their superiority over individual quality indicators in discriminating among hospi-

tals or forecasting hospital performance has not been carefully assessed.

In this study, we describe the development of an empirically-based composite measure for predicting operative mortality for 1 procedure—aortic valve replacement. We chose this procedure because it is relatively common, high-risk, and currently targeted by at least 1 major value-based purchasing initiative (the Leapfrog Group).[9] We focus on predicting operative mortality because it is the most commonly reported measure of hospital quality for this operation (including publicly reported data for New York, Pennsylvania, California, and New Jersey). Using national Medicare data, we apply an empirical Bayes methodology for optimally combining information from multiple quality domains, including procedure volume and other structural variables, procedure-specific morbidity and mortality, and hospital mortality with other surgical procedures. The resulting composite measure is the best prediction of the "true" risk-adjusted mortality rate in each hospital. We then evaluate the utility of this composite measure in explaining variation in hospital mortality rates and in forecasting subsequent hospital performance.

METHODS

Data Source and Study Population

We used 100% national analytic files from the Center for Medicare and Medicaid Services for years 2000 through 2003. MEDPAR files, which contain hospital discharge abstracts for all fee-for-service acute care hospitalizations of all US Medicare recipients, were used to create our main analysis datasets. The Medicare eligibility file was used to assess patient vital status at 30 days. The 2000 Census files and 2000 American Hospital Association files were used for supplemental information on patient and hospital characteristics, respectively, as described below.

Using appropriate procedure codes from the International Classification of Diseases, version 9 (ICD-9 codes), we identified all patients aged 65 to 99 undergoing aortic valve replacement. To minimize the potential for case-mix differences between hospitals, we excluded small patient subgroups with much higher baseline risks, including those with procedure codes indicating that other operations were simultaneously performed. We also identified (with similar exclusions) all patients undergoing 7 other relatively common elective procedures: percutaneous coronary interventions, coronary artery bypass surgery, mitral valve surgery, aortic aneurysm repair, carotid endarterectomy, esophagectomy, and pancreatic resection. These procedures were selected because they represent a large proportion of hospital deaths with elective surgery (either because they are very common or associated with high mortality risks).

Individual Surgical Quality Indicators

As potential inputs for our composite measure, we constructed hospital-level indicators from several different domains of surgical quality including measures of hospital structural characteristics (volume, teaching status, and nurse staffing levels), and measures of mortality and nonfatal complications for aortic valve replacement and each of the 7 other procedures. These quality indicators were selected because they can be assessed by available administrative data and because they have been shown in previous studies to correlate with operative mortality for many surgical procedures. In preliminary analyses, we also considered other quality indicators such as patient length of stay and more narrow definitions of nonfatal complications, but these indicators did not correlate with operative mortality and the results are not reported here.

Hospital procedure volume was assessed as the total number of procedures performed by each hospital in Medicare beneficiaries during 2000 to 2001. We constructed 3 separate measures: volume for aortic valve replacement, volume for all cardiac procedures (aortic and mitral valve surgery, percutaneous coronary interventions, and coronary artery bypass), and total volume for all 8 procedures. After testing several transformations, we used the natural log of the continuous volume variable for each operation in our analyses. Hospital teaching status (membership in the Council of Teaching Hospitals) and nurse ratios (Registered Nurse hours per patient day) were assessed using data from the American Hospital Association survey data from 2000.

Operative mortality was defined as death occurring before discharge or within 30 days of surgery. Because of the well-known limitations of ICD-9 coding for complications, we focus on a subset of complications from the Complications Screening Project that have been demonstrated to have good sensitivity and specificity for use with surgical patients.[10,11] The complications include pulmonary failure, pneumonia, myocardial infarction, acute renal failure, venous thrombosis, and postoperative hemorrhage.

We calculated risk-adjusted hospital mortality and nonfatal complication rates for each procedure using standard methods. For each operation, we determined the ratio of actual deaths or complications to the number of expected deaths (the O/E ratio). The number of expected deaths was the sum over all patients of the predicted probability of death or complications derived from a logistic regression model estimated on all patients undergoing a given procedure. The dependent variable in the logistic model was death or complications and the independent variables were patient covariates, which have been used to adjust for risk in previous work.[12] The patient characteristics included age, gender, race, admission acuity, and coexisting diseases assessed using the Elixhauser method.[13] A zip code-level measure of socioeconomic status was derived from data from the 2000 census.

Construction of the Composite Quality Measure

Our composite measure is a generalization of the standard shrinkage estimator that places more weight on a hospital's own standardized mortality ratio when it is measured reliably, but shrinks back toward the average mortality when a hospital's own mortality is measured with error (eg, for hospitals with small numbers of patients undergoing the procedure). Although the simple shrinkage estimator is a weighted average of a single mortality measure of interest and its mean (in our case, aortic valve replacement), our composite measure is a weighted average of all available

quality indicators—-the mortality and complication rates for all procedures along with all of the observable hospital structural characteristics (hospital volume, nurse-staffing ratios, and teaching status) that are thought to be related to patient outcomes.

The weight on each quality indicator is determined for each hospital to minimize the expected mean squared prediction error, using an empirical Bayes methodology outlined in Morris[14] (see the Technical Appendix for details, http://links.lww.com/A659). Empirical Bayes methods and more fully Bayesian hierarchical methods have been used previously to form hospital quality measures.[15,16] We use the empirical Bayes method because it is computationally simpler than fully Bayesian approaches, which allows the efficient use of large numbers of hospitals and quality indicators.[17,18]

The weight placed on each quality indicator in our composite measure depends on 2 factors. The first is the hospital-level correlation of each quality indicator with the mortality rate for aortic valve replacement. The strength of these correlations indicates the extent to which other quality indicators can be used to help predict mortality for aortic valve replacement, which is the ultimate goal. To estimate these hospital-level correlations, we first calculate the variance and covariance of the quality indicators, and then adjust for sampling variability by subtracting the average variance and covariance of the sampling error (averaged across hospitals). This approach is motivated by a hierarchical model in which the adjusted variance and covariance in the quality indicators captures the underlying hospital-level variation in quality of care across procedures net of any estimation error, and has been used successfully to estimate the correlation across quality measures in previous applications.[19]

The second factor affecting the weight placed on each quality indicator is the reliability with which each indicator is measured. The reliability of each quality indicator refers to the proportion of the overall variance that is attributable to true hospital-level variation in performance, as opposed to estimation error in the indicator. As in the usual shrinkage estimator, less weight is placed on quality indicators that are less reliable. For example, in smaller hospitals less weight is placed on the mortality and complication rates because they are less reliably estimated (because they are estimated from smaller samples of patients resulting in greater estimation error). We estimate the reliability of the indicators for each procedure as the ratio of the hospital-level variance to total variance in each indicator. The hospital-level variance is adjusted for sampling variability as discussed above. The total variance (and therefore reliability) varies by hospital, and is the sum of hospital-level variance and estimation error (which is larger in hospitals with fewer patients). We assume that structural characteristics of each hospital (such as volume) are not estimated with error and, therefore, have reliability equal to 1.

The resulting composite measure of surgical mortality has a number of attractive properties. First, it incorporates information in a systematic way from many quality indicators into the predictions of any 1 outcome. Moreover, if all of the estimated parameters (the hospital-level correlations and reliability of individual indicators) were known with certainty, the composite measure represents the optimal linear predictor of each hospital's true mortality rate, based on a mean squared error criterion. Because these parameters are more precisely estimated as the number of hospitals increases, the composite estimates are asymptotically (in the number of hospitals) the optimal linear predictor. Finally, these estimates maintain many of the attractive aspects of fully Bayesian approaches, while dramatically simplifying the complexity of the estimation.

To determine the relative importance for our composite measure, we calculated the correlation of each individual quality indicator with the mortality rate for aortic valve replacement, and calculated the average reliability of the standardized mortality and complication ratios for each procedure. Although one could in principal construct composite quality measures from every available quality indicator in our analysis, limiting the analysis to those quality indicators that are most likely to be strong predictors may improve the out-of-sample forecast performance of the composite measure by limiting the number of parameters that must be estimated. Finally, to summarize the importance of each quality indicator, we constructed a simple empirical Bayes prediction of mortality for aortic valve replacement that only depended on the single indicator, and then calculated the amount of variation in this prediction as a percentage of all hospital-level variation (adjusted for sampling variation as discussed above). This is a simple statistic, analogous to the R-squared from a regression, which summarizes the ability of a given prediction to explain the hospital-level variation in mortality for aortic valve replacement. We also used this statistic to compare the predictive performance of composite measures that incorporated various subsets of the quality indicators.

Validation of the Composite Measure

To validate the composite measures, and to compare the relative performance of various composite measures, we assessed how well composite measures from 2000 to 2001 predicted subsequent hospital-level variation in mortality rates for patients undergoing aortic valve replacement in 2002 to 2003. Hospitals were ranked according to the composite measure from 2000 to 2001 and grouped into quintiles. We then estimated random effect logistic models of mortality at the patient level using data from the subsequent 2 years (2002–2003), controlling for the same patient covariates as before and including either the composite measure directly (logged to better match the logistic specification) or including indicators for the quintile in which the hospital was ranked. The random effect logistic model allows for an unobserved hospital-level component (the random effect), which captures any hospital-level factors that were omitted from the model and systematically increases or decreases mortality for all patients in that hospital. Inclusion of the random effect corrects the standard errors for the resulting within-hospital correlation (clustering) in patient outcomes and provides an estimate of the variance of these unobserved differences across hospitals. We constructed an R-squared statistic for the 2002 to 2003 forecast (analogous to the statistic we used to evaluate the explanatory power of the composites in the 2000–2001 estimation sample) equal to the amount of variation being predicted by the composite measureas

a percentage of all hospital-level variation (predicted variance plus variance of the hospital random effect). Predicted variance is the variance of β*ln(Composite), where β it the estimated coefficient on the log of the composite from the random effect logit model. We then evaluate the forecast performance using the proportion of hospital-level variation in mortality in 2002 to 2003 that was explained by each composite measure, and using the difference in mortality in 2002 to 2003 across quintiles of each composite measure. All analyses were performed using stataMP 10.0 (Stata-Corp, College Station, TX).

RESULTS

We first determined the relative importance of structural variables in explaining variation in the standardized mortality ratio for aortic valve surgery (Table 1). In this and all subsequent tables, we report estimates that were weighted by the number of patients at each hospital. Hospital volume with aortic valve surgery, hospital volume with all cardiac procedures, and hospital volume with all major procedures were all inversely related to operative mortality (Table 1). Hospital volume with aortic valve replacement explained 19% of the hospital-level variation in mortality rates. Hospital volume with all cardiac procedures (10%) and all other major procedures combined (6%) explained a smaller proportion. Nurse staffing ratios and hospital teaching status were not strongly correlated with mortality and did not explain a significant proportion of observed mortality rates.

TABLE 1. Relationships Between Various Structural Measures and Operative Mortality With AVR, Based on National Medicare Data (2000–2001)

Structural Characteristic	Hospital-Level Correlation With AVR Mortality (O/E Ratio)	Predicted Variation as a % of All Hospital-Level Variation in AVR Mortality
Aortic valve replacement hospital volume	−0.17	19%
All cardiac procedure volume	−0.13	10%
All major procedure volume	−0.10	6%
Member in Council of Teaching Hospitals	−0.02	0%
Registered Nurse hours per patient day	0.01	0%

We next determined the importance of mortality and complications with other, related procedures in explaining variation in mortality with aortic valve surgery (Table 2). In the first column of Table 2 we report the average reliability of each measure across all hospitals doing aortic valve surgery. Recall that hospitals with fewer patients will have less reliable measures and these measures will be less useful as predictors (in the extreme, a hospital with no patients for a given surgery has a useless measure with reliability of zero). Several of the procedure-specific mortality rates demonstrated high enough reliability to be useful as inputs to the composite measures (ranged from 0.58 for percutaneous coronary interventions to 0.09 for pancreatectomy).

The second column of Table 2 reports the hospital-level correlation of each quality indicator with the mortality rate for aortic valve replacement (adjusted for sampling variation as discussed in the methods section). We found the strongest correlation between the mortality rate for aortic valve surgery and the mortality rate for other cardiac procedures, including both coronary artery bypass (adjusted correlation coefficient, 0.83) and mitral valve surgery (adjusted correlation coefficient, 1.0). In other words, hospitals with low mortality rates with aortic valve surgery also had low mortality rates for mitral valve surgery and coronary bypass surgery. The correlation with other, noncardiac operations was still significant but weaker (Table 2).

The final column of Table 2 reports the amount of variation in aortic valve replacement mortality that was predicted by each measure (using a simple empirical Bayes prediction), as a percentage of all hospital-level variation. The amount of variation explained by each measure is approximately equal to the product of each measure's reliability and the square of its correlation with aortic valve replacement mortality. Thus, measures with low reliability or correlation explain little variation. The proportion of hospital-level variation

TABLE 2. Relationships Between Various Outcome Measures and Operative Mortality With AVR, Based on National Medicare Data (2000–2001)

Outcome Measure (O/E Ratio)	Average Reliability of Measure	Hospital-Level Correlation With AVR Mortality (O/E Ratio)	Predicted Variation as a % of All Hospital-Level Variation in AVR Mortality
Mortality rates			
Aortic valve replacement	0.34	1.00	35%
Coronary artery bypass grafting	0.55	0.83	39%
Mitral valve replacement	0.20	1.00	21%
Percutaneous coronary interventions	0.58	0.49	15%
Carotid endarterectomy	0.10	0.67	4%
Elective aortic aneurysm repair	0.21	0.33	2%
Esophagectomy	0.15	0.23	1%
Pancreatectomy	0.09	0.72	7%
Non-fatal complications following AVR*	0.62	0.32	7%

*Complication rates with other procedures accounted for less than 2% of nonrandom variation in AVR mortality and are not listed here.

TABLE 3. Ability of Various Composite Measures in Describing Variation in Hospital Mortality With AVR and in Forecasting Subsequent Mortality

	Components Included in Composite Measure						In-Sample Prediction (2000–2001)	Out-of-Sample Prediction (2002–2003 Mortality)	
	Procedure Volume (2000–2001)			Mortality (2000–2001)			Predicted Variation as a % of All Hospital-Level Variation in AVR Mortality	Predicted Variation as a % of All Hospital-Level Variation in AVR Mortality	Adjusted Odds Ratio, Best Vs. Worst Hospital Quintile (95% CI)
	AVR	Other Cardiac Procedures	Other Major Procedures	AVR	Other Cardiac Procedures	Other Major Procedures			
Model 1	X						19%	9%	1.32 (1.14–1.53)
Model 2				X			35%	32%	1.66 (1.45–1.91)
Model 3	X			X			44%	34%	1.66 (1.45–1.91)
Model 4	X	X		X	X		78%	70%	1.97 (1.73–2.23)
Model 5	X	X	X	X	X	X	72%	66%	1.86 (1.64–2.12)
Model 6*	X	X	X	X	X	X	73%	66%	1.98 (1.74–2.25)
Model 7†	X	X	X	X	X	X	75%	62%	1.87 (1.65–2.13)

*Nursing staff, teaching status, and nonfatal complications were also included.
†Model 6 plus complications for all of the other 7 procedures.

in aortic valve mortality explained was highest for coronary artery bypass mortality (39%), aortic valve mortality (35%), mitral valve mortality (21%), and percutaneous coronary interventions (15%). For these procedures, mortality was either strongly correlated with aortic valve mortality and/or was measured reliably. The remaining procedures were either unreliable or were weakly correlated with aortic valve mortality and explained no more than 7% of the hospital-level variation. Finally, nonfatal complication rates for aortic valve replacement and for all other surgeries (not reported in the table) were weak predictors, explaining no more than 7% of the hospital variation in aortic valve mortality.

Table 3 shows a comparison of several combinations of these inputs to create different composite scores. We compared the ability of each composite to predict hospital-level variation in aortic valve mortality both during the period when the rankings were created (in-sample) and during the subsequent 2 years (out-of-sample forecast). In model 1, hospital volume with aortic valve surgery (alone) explained 19% of the variation in 2000 to 2001 but forecasted only 9% of the variation in 2002 to 2003. In model 2, the risk-adjusted mortality rate with aortic surgery explained 35% of the variation in 2000 to 2001 and forecasted 32% of the variation in 2002 to 2003. In model 3, the combination of aortic valve mortality and volume performed better than either measure alone, explaining 44% (2000–2001) and 34% (2002–2003) of the variation. Model 4, the combination of aortic valve volume, aortic valve mortality, and mortality and volume with other cardiac procedures, seemed to be the most parsimonious model with the best ability to predict future performance, explaining 78% (2000–2001) and 70% (2002–2003) of the variation. Although models 5–7 added additional quality indicators to the composite measure and explained similar amounts of variation in 2000 to 2001, their ability to predict subsequent performance was inferior to the more parsimonious model 4.

The composite measure based on model 4 was also superior to all other models in terms of forecasting future mortality differences across the quintiles of previous performance (Table 3, last column). For example, hospitals scoring in the bottom quintile on the model 4 composite measure in 2000 to 2001 had 2-fold higher mortality rates in 2002 to 2003 than hospitals in the top quintile (adjusted odds ratio, 1.97; 95% CI, 1.73–2.23). Figure 1 illustrates the superior ability of the composite measure from model 4 to identify subsequent differences in aortic valve mortality in 2002 to 2003, compared with aortic valve replacement volume or mortality alone.

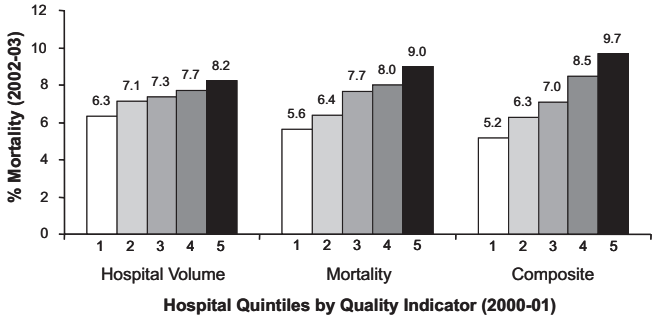

FIGURE 1. Ability of various historical (2000–2001) quality indicators to forecast subsequent (2002–2003) risk-adjusted mortality with aortic valve replacement (AVR) Hospitals sorted into quintiles according to hospital volume alone, risk-adjusted mortality alone, and composite measure based on AVR mortality, AVR hospital volume, and mortality and volume with other cardiac procedures.

The composite measure based on model 4 obtained improved predictive power by placing substantial weight on risk-adjusted mortality rates from other cardiac procedures. The composite measure for each hospital's aortic valve replacement mortality was the sum of 5 terms. The first term was the predicted mortality based on patient volume. The remaining 4 terms were risk-adjusted mortality rates (after removing the volume effect) for each cardiac related surgery, with less weight placed on these measures in a hospital with a small sample of patients. For an average hospital in our

sample, model 4 composites placed the most weight on mortality in coronary artery bypass graft surgeries (0.29), followed by aortic valve replacement (0.12), percutaneous coronary interventions (0.10), and mitral valve replacement (0.08). Without information on mortality in the other surgeries, model 3 composites placed more weight on mortality in aortic valve replacement (0.27) but this resulted in less accurate predictions as reported in Table 3.

DISCUSSION

Despite a growing need for reliable indicators of surgical performance, little consensus exists on the optimal measures for this purpose. The present study demonstrates that an empirically derived composite measure offers several advantages over existing approaches, and may be the best approach for measuring surgical performance. The method set forth in this manuscript uses empiric weighting techniques to combine multiple domains of quality, and optimally uses all available information for a given operation. For aortic valve replacement, the composite measure (based on a combination of operative mortality, hospital volume, and mortality and volume with other related operations) explained 78% of hospital-level variation in aortic valve mortality rates. The composite measure was also better at forecasting future performance, with a nearly 2-fold difference between the best and worst hospital quintiles in the subsequent 2 years.

The idea of combining multiple measures to create a composite is not novel. Many existing pay-for-performance efforts already employ composite measures of performance for medical and surgical diagnoses.[8,20,21] For example, the Center for Medicare and Medicaid Services (CMS)/Premier Hospital Quality Incentive Demonstration Project uses a composite score for coronary artery bypass surgery. This composite is a combination of 5 process variables and 3 outcome variables, which are simply weighted according to the number of measures (5/8 weight to the process measures and 3/8 to the outcome measures). Although the CMS/Premier approach deals with the problem of multiple measures, this simple weighting strategy ignores the simple fact that some measures may be more important than others. For example, risk-adjusted mortality rates may be more important than the appropriate timing of antibiotic prophylaxis. In contrast to this approach, the empirically-based methods used in our present paper weight the inputs according to how reliable they are and how strongly they are related to the outcome of interest.

The Society of Thoracic Surgeons (STS) has recently developed a method similar to ours for measuring the quality of coronary artery bypass surgery.[20,21] The STS convened a task force to select candidate measures and test several approaches for creating a composite score of quality. The STS method for creating a composite involves estimating scores for each of 4 domains: mortality, morbidity, internal mammary artery usage, and perioperative medications. Each of these domain scores is then combined using "all or none" weighting into a composite measure. At first glance, this method appears quite different from ours. However, each of the domain-specific scores is generated using a method analogous to that presented in the present study. The STS method integrates information across multiple measures to optimally predict the "true" rate for each domain, similar to our approach. However, there are 2 key differences. First, we use an empirical Bayes approach while the STS methods are fully Bayesian. Second, we do not perform the second step—adding together multiple domains to create a "global" measure of quality. In our method, we assume that mortality is the gold standard and empirically weight each input to optimally predict this 1 domain. Despite the similarities in analytic methods, this second step is the key distinction between our method and those of the STS. The STS composite score is designed to represent global quality (a latent construct) and our measure is designed to optimally predict future risk-adjusted mortality.

Although the application of empirically derived composite measures to surgery is new, previous work has documented their use in medical populations. McClellan and Staiger combined information on multiple outcomes (mortality rates at 30 days, 90 days, and 1 year; and readmission rates) to measure quality for more than 200,000 Medicare patients hospitalized for acute myocardial infarction.[17] Using composite measures, they were able to dramatically improve the consistency of quality rankings from year to year and were able to accurately forecast large differences in subsequent mortality.[17] However, in this previous work, McClellan et al focused only on mortality outcomes assessed at different time points. In the present article, we also combined information from other domains (eg, hospital volume) to improve the predictive accuracy of our composite measure.

In addition to combining conventional measures already in use, such as mortality rates and hospital volume, we demonstrate the value of adding information that is often overlooked—outcomes for other, related procedures. Previous studies have shown that mortality rates are correlated at the hospital level, especially for cardiac surgery.[22] In creating the composite measure for aortic valve replacement, we combined information on mortality for coronary artery bypass surgery, mitral valve surgery, and percutaneous coronary interventions. We have shown that adding this previously overlooked source of information improves predictive ability and contributes to explaining the observed hospital-level variation in mortality rates.

In contrast, we found that information on postoperative complications did not improve predictive ability. Others have also found a low correlation between morbidity and mortality rankings.[22,23] This may be because of poor accuracy of administrative data in ascertaining which patients had complications, despite our focus on complications with demonstrated accuracy for use with surgical patients.[10,11] The lack of value of adding complications may also suggest, as some previous studies have found, that failure to rescue from complications rather than complications themselves are more strongly correlated with mortality.[23]

Rather than choosing input measures on previous evidence or expert opinion, we treated all inputs (including mortality with other seemingly unrelated operations) as potentially valid indicators of the construct being measured.

This is a novel aspect of our article—we incorporate previously overlooked sources of quality information. In some ways, mortality with other operations could be a more valid indicator for aortic valve surgery than certain processes of care that turn out to be unrelated to outcomes. For example, mortality with other operations may pertain to the general strength of the surgical department or the quality of postoperative nursing care. A valuable aspect of our composite method is the explicit consideration of how each input measure relates to the "gold standard" quality measure—risk-adjusted mortality—-rather than assuming these relationships a priori.

However, there are potential limitations of this approach to creating composite measures. The weights placed on each component reflect its importance in predicting an operative mortality. We focus on predicting operative mortality because it is the most commonly reported measure of hospital quality for this operation (including publicly reported data for New York, Pennsylvania, California, and New Jersey). More generally, it may be appropriate to construct a composite measure that predicts other domains of quality with this operation, such as long-term survival and quality of life. Unfortunately, the inputs for such a measure are not widely available. However, the methods set forth in this article would apply to any gold standard measure if the appropriate input variables were available.

We acknowledge several additional limitations to the current study, most of which relate to our use of claims data. It is well known that claims data do not have the detailed clinical data necessary for optimal case-mix adjustment. This problem is not unique to our composite measure and is a problem with any attempt to use claims data to profile hospitals. If differences in case-mix across hospitals are systematic (eg, certain hospitals consistently treat sicker patients), our ability to forecast future performance would have been overestimated. Although it is impossible to know whether differences in unobserved case-mix are random or systematic, we tried to minimize this bias by limiting the cohort to a homogenous group of patients undergoing the same operation. In terms of the variables available in claims data, patient severity did not vary systematically across hospitals in our cohort. For example, expected mortality was 7.5% in the highest mortality hospitals and 7.4% in the lowest mortality hospitals. Correspondingly, when we estimated our composite measure with and without controlling for the Elixhauser comorbidity variables, the resulting composite measures were almost identical (correlation = 0.96). Thus, although the comorbidity variables are highly significant at the patient level, they are unrelated to hospital performance in our cohort and controlling for them does not substantively alter our estimates of hospital performance. It is not known whether a more detailed assessment of patient characteristics would alter these findings.

The use of administrative data also limits the available inputs to our composite measure. If a richer source of data were used, other domains of quality, including clinical process measures, could be considered. For example, preoperative β-blocker use, and the appropriate selection and timing of antibiotics are all used as quality measures for aortic valve replacement.[24] More detailed data sources would also provide other outcomes measures, including more accurate nonfatal clinical outcomes, which could be used as inputs to a composite measure. It is possible that these process measures and more detailed nonfatal outcomes would account for some of the unexplained systematic variation in mortality and improve the predictive ability of our measure.

Some may also view our focus on a single operation as a limitation. In some ways, aortic valve replacement is the ideal operation for illustrating the advantages of composite measures. Largely, this is because there are no good standalone quality measures for this operation. Caseloads at individual hospitals can be low with this procedure, making mortality rates too "noisy" to be useful. Hospital volume is related to mortality but the relationship is not as strong as with other procedures. In contrast, for other procedures, the composite approach may have smaller incremental benefits over individual quality indicators. For coronary artery bypass, the caseloads at individual hospitals are large enough to use mortality as a reliable measure. For esophageal resection, hospital volume is strongly related to mortality and can be used as a good proxy measure of quality. Although the relative advantages are not as great for these operations, the composite approach will be at least as good, and most likely better, than the best available individual quality indicator.

The composite measures outlined in this article will be useful for many purposes. Because they are more reliable at discriminating hospital performance than existing measures, they will be useful for benchmarking performance for quality improvement. For example, these composite measures can be used to identify high and low performing hospitals. A standardized instrument could be developed and used to review the medical charts at these hospitals to identify whether certain evidence-based processes of care were used. By comparing the relative use of these processes of care at high and low performing hospitals, we will be able to understand the mechanisms that explain variation in performance. After these processes are identified, this knowledge can be disseminated with the aim of incrementally improving care at all hospitals.

Perhaps the most obvious use for composite measures is for steering patients towards the best hospitals, through either public reporting or payer-led selective contracting. We validated our measure by showing that it predicted subsequent outcomes, which is exactly what consumers and payers want to know. With the several year lag in most hospital report cards, any useful measure must be a valid predictor of future performance. With the ongoing proliferation of pay-for-performance initiatives, the demand for reliable summary measures of performance continues to grow. The empirically derived composite measures outlined in the paper could help meet this demand.

REFERENCES

1. Pronovost PJ, Miller M, Wachter RM. The GAAP in quality measurement and reporting. *JAMA.* 2007;298:1800–1802.
2. Kizer KW. Establishing health care performance standards in an era of consumerism. *JAMA*. 2001;286:1213–1217.

3. Rosenthal MB, Dudley RA. Pay-for-performance: will the latest payment trend improve care? *JAMA*. 2007;297:740–744.
4. Birkmeyer JD, Dimick JB, Birkmeyer NJ. Measuring the quality of surgical care: structure, process, or outcomes? *J Am Coll Surg*. 2004; 198:626–632.
5. 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;296:72–78.
6. Centers for Medicare and Medicaid Services (CMS), HHS. Medicare program: reporting hospital quality data for FY 2008 inpatient prospective payment system annual payment update program–HCAHPS survey, SCIP, and mortality. *Fed Regist*. 2006;71:67959–68401.
7. Dimick JB, Welch HG, Birkmeyer JD. Surgical mortality as an indicator of hospital quality: the problem with small sample size. *JAMA*. 2004; 292:847–851.
8. http://www.cms.hhs.gov/HospitalQualityInits/35_HospitalPremier.asp. Accessed November 6, 2007.
9. The Leapfrog Group. Evidence to Based Hospital Referral Fact Sheet. Available at: http://www.leapfroggroup.org/. Accessed November 6, 2007.
10. Weingart SN, Iezzoni LI, Davis RB, et al. Use of administrative data to find substandard care: validation of the complications screening program. *Med Care*. 2000;38:796–806.
11. Lawthers AG, McCarthy EP, Davis RB, et al. Identification of in-hospital complications from claims data. Is it valid? *Med Care*. 2000; 38:785–795.
12. Birkmeyer JD, Dimick JB, Staiger DO. Operative mortality and procedure volume as predictors of subsequent hospital performance. *Ann Surg*. 2006;243:411–417.
13. Southern DA, Quan H, Ghali WA. Comparison of the Elixhauser and Charlson/Deyo methods of comorbidity measurement in administrative data. *Med Care*. 2004;42:355–360.
14. Morris CN. Parametric Empirical Bayes Inference: theory and applications. *J Am Stat Assoc*. 1988;78:47–55.
15. Gatsonis CA, Epstein AM, Newhouse JP, et al. Variations in the utilization of coronary angiography for elderly patients with acute myocardial infarction: an analysis using hierarchical logistic regression. *Med Care*. 1995;33:625–642.
16. Normand ST, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. *J Am Stat Assoc*. 1997;92:803–814.
17. McClellan MB, Staiger DO. Comparing the quality of health care providers. In: Alan Garber, ed. *Frontiers in Health Policy Research*. Vol 3. Cambridge MA: The MIT Press; 2000:113–136.
18. Rogowski JA, Staiger DO, Horbar JD. Variations in the quality of care for very-low-birthweight infants: implications for policy. *Health Aff (Millwood)*. 2004;23:88–97.
19. Zaslavsky AM, Cleary PD. Dimensions of plan performance for sick and healthy members on the Consumer Assessments of Health Plans Study 2.0 survey. *Med Care*. 2002;40:951–964.
20. Shahian DM, Edwards FH, Ferraris VA, et al. Society of thoracic surgeons quality measurement task force. Quality measurement in adult cardiac surgery: part 1–conceptual framework and measure selection. *Ann Thorac Surg.* 200;83(4 suppl):S3–S12.
21. O'Brien SM, Shahian DM, DeLong ER, et al. Quality measurement in adult cardiac surgery: part 2–statistical considerations in composite measure scoring and provider rating. *Ann Thorac Surg*. 2007;83(4 suppl):S13–S26.
22. Goodney PP, O'Connor GT, Wennberg DE, et al. Do hospitals with low mortality rates in coronary artery bypass also perform well in valve replacement? *Ann Thorac Surg*. 2003;76:1131–1136.
23. Silber JH, Rosenbaum PR, Williams SV, et al. The relationship between choice of outcome measure and hospital rank in general surgical procedures: implications for quality assessment. *Int J Qual Health Care*. 1997;9:193–200.
24. National Quality Forum-endorsed standards for acute care hospital performance. Available at: http://www.qualityforum.org/pdf/lsEndorsedStandardsALL08-14-07corrected.pdf. Accessed November 6, 2007.

