HSR Health Services Research


DOI: 10.1111/j.1475-6773.2012.01407.x
RESEARCH ARTICLE

# Composite Measures for Rating Hospital Quality with Major Surgery

*Justin B. Dimick, Douglas O. Staiger, Nicholas H. Osborne, Lauren H. Nicholas, and John D. Birkmeyer*

**Objective.** To assess the value of a novel composite measure for identifying the best hospitals for major procedures.

**Data Source.** We used national Medicare data for patients undergoing five high-risk surgical procedures between 2005 and 2008.

**Study Design.** For each procedure, we used empirical Bayes techniques to create a composite measure combining hospital volume, risk-adjusted mortality with the procedure of interest, risk-adjusted mortality with other related procedures, and other variables. Hospitals were ranked based on 2005–2006 data and placed in one of three groups: 1-star (bottom 20 percent), 2-star (middle 60 percent), and 3-star (top 20 percent). We assessed how well these ratings forecasted risk-adjusted mortality rates in the next 2 years (2007–2008), compared to other measures.

**Principal Findings.** For all five procedures, the composite measures based on 2005–2006 data performed well in predicting future hospital performance. Compared to 1-star hospitals, risk-adjusted mortality was much lower at 3-star hospitals for esophagectomy (6.7 versus 14.4 percent), pancreatectomy (4.7 versus 9.2 percent), coronary artery bypass surgery (2.6 versus 5.0 percent), aortic valve replacement (4.5 versus 8.5 percent), and percutaneous coronary interventions (2.4 versus 4.1 percent). Compared to individual surgical quality measures, the composite measures were better at forecasting future risk-adjusted mortality. These measures also outperformed the Center for Medicare and Medicaid Services (CMS) Hospital Compare ratings.

**Conclusion.** Composite measures of surgical quality are very effective at predicting hospital mortality rates with major procedures. Such measures would be more informative than existing quality indicators in helping patients and payers identify high-quality hospitals with specific procedures.

**Key Words.** Administrative data uses, econometrics, modeling, multi-level, risk adjustment for clinical outcomes, quality of care/patient safety (measurement)

With wide recognition that surgical outcomes vary across hospitals, information on surgical quality is in high demand. Patients, families, and referring physicians are looking for information to help them select the best hospital for specific procedures (Kizer 2001). Payers and large health care purchasers need

accurate hospital ratings to inform their selective referral and value-based purchasing programs (Galvin 2001). To meet these demands, several organizations are publicly reporting measures of surgical quality. The Leapfrog Group, a large coalition of public and private purchasers, reports information on mortality and hospital volume with major surgical procedures (Dimick et al. 2009). As its primary measure of surgical quality, the Center for Medicare and Medicaid Services (CMS) publicly reports hospital compliance with several process measures on its Hospital Compare website (Stulberg et al. 2010).

Whether such measures are optimal for helping patients and payers identify the safest hospitals for major procedures is uncertain, however. Measures based on structure, process, and outcomes all have distinct limitations (Birkmeyer, Dimick, and Birkmeyer 2004). Hospital volume, for example, is a useful predictor of outcomes for some operations, but it is a relatively weak proxy of quality for most procedures (Birkmeyer et al. 2002; Halm, Lee, and Chassin 2002). Outcome measures, such as mortality and morbidity, are often too "noisy" to reliably measure hospital quality due to small sample sizes and low event rates (Shahian et al. 2001; Dimick, Welch, and Birkmeyer 2004). Process measures, such as those of the Surgical Care Improvement Program (SCIP) reported by CMS on Hospital Compare, are only weakly related to hospital outcomes with surgery (Stulberg et al. 2010). Providing patients with all available information may be better than any single indicator, but they would still require guidance on how to weight various measures, particularly when they conflict.

In previous work, we have evaluated several different approaches for improving hospital quality measurement for surgery. Using national Medicare data, we recently described the value of a simple composite measure for profiling hospital quality with surgery based on volume and mortality alone (Dimick et al. 2009). This measure has since become the quality standard for the Leapfrog Group and was recently endorsed by the National Quality Forum (NQF) for use with esophagectomy, pancreatectomy, and abdominal aortic aneurysm repair. In another recent publication, we demonstrated the

---

Address correspondence to Justin B. Dimick, M.D., M.P.H., Assistant Professor of Surgery, University of Michigan, 2800 Plymouth Road, Building 520, Office 3144, Ann Arbor, MI 48109, e-mail: jdimick@umich.edu. Douglas O. Staiger, Ph.D., John French Professor, is with the Department of Economics, Dartmouth College, Hanover, NH; Nicholas H. Osborne, M.D., M.S., Surgical House Officer, is with the Department of Surgery at the University of Michigan, Ann Arbor, MI; Lauren H. Nicholas, Ph.D., Faculty Research Fellow, is with the Institute for Social Research, University of Michigan, Ann Arbor, MI; and John D. Birkmeyer, M.D., George Zuidema Professor, is with the Department of Surgery, University of Michigan, Ann Arbor, MI.

value of empirical Bayes techniques for filtering out statistical "noise" in surgical mortality measurement (Dimick, Staiger, and Birkmeyer 2010). Finally, we have demonstrated the feasibility of a more comprehensive composite measure that takes into account a broader array of measures, albeit in a narrow clinical context (Staiger et al. 2009).

In this article, we demonstrate the value of a comprehensive composite measure for a broad range of surgical conditions. We combine the insights from our prior work, and others, to create a composite measure that incorporates all sources of information, including quality indicators for other, related operations, and uses empirical Bayes techniques to filter out statistical "noise." Given the manner in which hospital ratings are used, we examined the ability of these measures to forecast future hospital mortality, compared to several other existing quality measures, including hospital volume, risk-adjusted mortality, simple empirical Bayes methods, and the process measures reported by the Center for Medicare and Medicaid Services (CMS) on its Hospital Compare website.

## METHODS

### *Data Source and Study Population*

We used data from the Medicare Analysis Provider and Review (MEDPAR) files for 2005–2008 to create our main analysis datasets. This dataset contains hospital discharge abstracts for all fee-for-service acute care hospitalizations of U.S. Medicare recipients, which accounts for approximately 70 percent of such admissions in the Medicare patients. The Medicare eligibility file was used to assess patient vital status at 30 days. The study protocol was approved by the Institutional Review Board at the University of Michigan.

Using appropriate procedure codes from the International Classification of Diseases, version 9 (ICD-9), we identified all patients aged 65–99 undergoing five high-risk surgical procedures that are targeted by the Leapfrog Group as part of their evidence-based hospital referral program: coronary artery bypass grafting, aortic valve replacement, percutaneous coronary interventions, and resection of pancreatic and esophageal cancers (Dimick et al. 2009). To enhance the homogeneity of hospital case mix, we excluded small patient subgroups with much higher baseline risks, including those with procedure codes indicating that other operations were simultaneously performed (e.g., coronary artery bypass and valve surgery) or were performed

for emergent indications (e.g., acute myocardial infarction with percutaneous coronary intervention) (Birkmeyer et al. 2002).

### *Hospital Ratings*

Hospitals were rated using information from 2005 to 2006. In creating our composite measures, we considered several individual quality measures, including measures of hospital structure (volume, teaching status, and nurse-to-patient ratios), and outcomes (risk-adjusted mortality and nonfatal complications). For each operation, we considered hospital volume and mortality not only for the index operation but also for other, related procedures (e.g., coronary bypass mortality and volume were tested as inputs to the composite measure for aortic valve replacement). These quality indicators were selected because they can be readily estimated from administrative data and they have been shown in previous studies to correlate with mortality for many surgical procedures (Goodney et al. 2003; Dimick, Staiger, and Birkmeyer 2006). In preliminary analyses, we also considered other quality indicators such as patient length of stay and alternative definitions of nonfatal complications, but these indicators did not correlate with mortality and were therefore not included.

Hospital volume was calculated as the number of Medicare cases performed during a 2-year period (2005–2006). We constructed three separate measures: volume for the index procedure, volume for all related procedures (i.e., procedures in the same clinical specialty), and total hospital volume (i.e., including procedures from other specialties). 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 2005.

Risk-adjusted mortality and nonfatal complication rates were calculated at all hospitals over a 2-year period (2005–2006). 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 (Lawthers et al. 2000; Weingart et al. 2000). Complication rates were calculated as one or more complication for each patient. For mortality and nonfatal complications, risk adjustment was performed using logistic regression of patient covariates, including age, gender, race, urgency of admission,

socioeconomic position, and comorbid diseases. To adjust for socioeconomic position, we used a zip code-level measure derived from the most recent U.S. census data. Comorbid diseases were ascertained from secondary diagnoses using ICD-9 codes using the methods of Elixhauser (Southern, Quan, and Ghali 2004). Each comorbid disease was entered into the risk-adjusted model as an independent variable. We used a logistic regression model to estimate the predicted probability of the outcome for each patient. These probabilities were then summed to generate the number of expected deaths. Risk-adjusted mortality was then calculated by dividing the observed by the expected deaths and multiplying by the average procedure-specific mortality rate.

We developed a composite measure that incorporates information from multiple quality indicators to optimally predict "true" risk-adjusted mortality. Our composite measure is a generalization of the standard shrinkage estimator that places more weight on a hospital's own mortality rate when it is measured reliably but shrinks back toward the average mortality when a hospital's own mortality is measured with error (e.g., for hospitals with small numbers of patients undergoing the procedure) (Staiger et al. 2009). While the simple shrinkage estimator is a weighted average of a single mortality measure of interest and its mean, 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.

Although the statistical methods used to create these measures are described in detail elsewhere (Staiger et al. 2009), we will provide a brief conceptual overview (full mathematical details can be found in the Appendix S1). Perhaps the best way to conceptualize these methods is to imagine a simple regression equation where the outcome (e.g., a hospital's mortality rate in the next year) is not known. The goal of the analyses described below is to estimate the regression coefficients (i.e., the weights on different quality indicators) to best predict next year's mortality. We then evaluate our "model," the composite measure, based on how well it actually predicts mortality in a subsequent time period.

The first step in creating the composite measure was to determine the extent to which each individual quality indicator predicts risk-adjusted mortality for the index operation. To evaluate the importance of each potential input, we first estimated the proportion of systematic variation in risk-adjusted

Table 1: Components of the Composite Measure for All Five Procedures Are Shown, along with the Proportion of Nonrandom Hospital-Level Mortality Explained by Each

| *Procedure* | *Individual Quality Measures* | *Proportion of Hospital-Level Variation Explained (%)* |
|---|---|---|
| Coronary artery by pass grafting | Mortality | 50 |
| | Mortality with aortic valve surgery | 25 |
| | Mortality with mitral valve surgery | 20 |
| | Mortality with percutaneous coronary interventions | 14 |
| | Hospital volume | 11 |
| | Hospital volume with other operations | 11 |
| Aortic valve replacement | Hospital volume | 36 |
| | Mortality | 36 |
| | Mortality with coronary artery bypass surgery | 36 |
| | Mortality with mitral valve surgery | 23 |
| | Hospital volume with other operations | 18 |
| | Mortality with percutaneous coronary interventions | 14 |
| Percutaneous coronary interventions | Mortality | 51 |
| | Hospital volume | 18 |
| | Hospital volume with other operations | 18 |
| | Mortality with coronary artery bypass surgery | 12 |
| Pancreatic cancer resection | Hospital volume | 65 |
| | Hospital volume with other operations | 43 |
| | Hospital teaching status | 39 |
| | Nurse-to-patient ratios | 35 |
| | Mortality | 24 |
| | Mortality with colon surgery | 24 |
| Esophageal cancer resection | Hospital volume | 48 |
| | Hospital volume with other operations | 34 |
| | Hospital teaching status | 24 |
| | Mortality with colon surgery | 21 |
| | Mortality | 19 |
| | Mortality with aortic valve surgery | 15 |
| | Mortality with pancreatic resection | 12 |
| | Nurse-to-patient ratios | 12 |

mortality explained by each individual quality indicator (Table 1), where the systematic variation is defined as the hospital variation (i.e., between hospital variance) derived from hierarchical logistic regression models. We then entered each quality indicator into the model and assessed the degree to which it reduced the hospital-level variance, as described in our prior work (Staiger et al. 2009). Thus, when using the mortality from the index operation itself as

the quality indicator, this estimate reflects the reliability of this outcome measure. We included any quality indicator in the composite measure that explained more than 10 percent of hospital variation in risk-adjusted mortality over a 2-year period (2005–2006).

Next, we calculated weights for each quality indicator. The weight placed on each quality indicator in our composite measure was based on two factors (Staiger et al. 2009). The first is the hospital-level correlation of each quality indicator with the mortality rate for the index operation. The strength of these correlations indicates the extent to which other quality indicators can be used to help predict mortality for the index operation. The second factor affecting the weight placed on each quality indicator is the reliability with which each indicator is measured. Reliability ranges from 0 (no reliability) to 1 (perfect reliability). 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 ("noise"). For example, in smaller hospitals, less weight is placed on mortality and complication rates because they are less reliably estimated. We assume that structural characteristics of each hospital (such as hospital volume) are not estimated with error and, therefore, have reliability equal to 1.

### *Comparing Performance of Hospital Ratings*

We determined the value of our composite measure by determining how well it predicted future risk-adjusted mortality compared to several other approaches. For each operation (data from years 2005–2006), hospitals were ranked based on the composite measure. Each hospital was assigned one of three rankings (1-star, 2-star, and 3-star). The "worst" hospitals (bottom 20 percent) received a 1-star rating, the middle of the distribution (60 percent) received a 2-star rating, and the "best" hospitals (top 20 percent) received a 3-star rating. Many hospital rating systems determine tiers of performance by designating high and low outliers by testing for statistically significant differences from the average. Because we used empirical Bayes methods, which adjust each hospital's composite for imprecision (i.e., hospital rankings are a valid indicator of relative performance), we used percentile cut-offs for this analysis. We then calculated the risk-adjusted mortality rates for 1-star, 2-star, and 3-star hospitals during the subsequent 2 years (data from years 2007–2008).

We next assessed the ability of our composite measure to predict future performance compared to other widely used surgical quality. As with the composite measure, for each operation (data from years 2005–2006), hospitals

were ranked based on hospital volume, risk-adjusted mortality, "reliability adjusted" mortality (i.e., empirical Bayes shrinkage alone), and process measures from Hospital Compare. Hospital Compare reports data on each hospital's compliance with processes of care from the Surgical Care Improvement Program (SCIP). These measures track adherence to processes of care which are proven to prevent common surgical complications, including infection, venous thromboembolism, and cardiac complications (Stulberg et al. 2010). For all of these analyses, we evaluated the discrimination in future risk-adjusted mortality, comparing the 1-star hospitals (bottom 20 percent) to the 3-star hospitals (top 20 percent) for each of the measures.

We also assessed compared the ability of the composite measure and these other quality measures to explain future (2007–2008) hospital-level variation in risk-adjusted mortality. To avoid problems with "noise variation" in the subsequent time period, we determined the proportion of systemic hospital-level variation explained. We generated hierarchical models with mortality as the dependent variable (2007–2008) and used them to estimate the hospital-level variance. We first used an "empty model" that contained only patient variables for risk adjustment. We then entered each historical quality measure (assessed in 2005–2006) into the model. We then calculated the degree to which the historical quality measures reduced the hospital-level variance, an approach described in our prior work (Staiger et al. 2009).All statistical analyses were conducted using STATA 10.0 (College Station, Texas).

## RESULTS

### *Components of the Composite Measure*

For each of the five procedures, several individual measures explained a significant proportion of hospital-level variation in risk-adjusted mortality (Table 1). The importance of each individual measure varied across procedures. For example, hospital volume with the procedure of interest explained 65 percent of the variation in risk-adjusted mortality for pancreatic resection, but only 11 percent for coronary artery bypass surgery (Table 1). Hospital volume with other operations was also important in explaining variation in mortality with all five procedures, explaining from 11 percent (coronary artery bypass surgery) to 43 percent (pancreatic resection). Other structural measures, including hospital teaching status and nurse-to-patient ratios, explained a large proportion of the variation in mortality for pancreatic resection and esophageal resection, but they were not important for other procedures (Table 1).

The amount of hospital-level variation explained by each procedure's own mortality rate varied, ranging from 50 percent with coronary artery bypass surgery to only 19 percent for esophageal resection (Table 1). Mortality with other, related procedures was important in explaining hospital-level variation for all five procedures (Table 1). For example, mortality with aortic valve surgery and mitral valve surgery explains 25 and 20 percent of the hospital-level variation in risk-adjusted mortality with coronary artery bypass surgery, respectively. Rates of nonfatal complications were not important in explaining variation in mortality rates for any of the five procedures.

The average weights on each input measure also varied across procedures (Table 2). For each procedure, most weight was placed on the structural variables such as hospital volume. The weight placed on structural variables such as volume also includes the weight placed on the average mortality (consistent with empirical Bayes "shrinkage") and therefore appears relatively high for all operations. Nonetheless, the amount of weight placed on each

Table 2: Weights Placed on Each Input Measure for the Composite Measure for All Five Procedures Are Shown

| *Procedure* | *Individual Quality Measures* | *Weight in the Composite Score (%)* |
|---|---|---|
| Coronary artery by pass grafting | Mortality expected given hospital volume | 47 |
| | Mortality | 31 |
| | Mortality with aortic valve surgery | 10 |
| | Mortality with percutaneous coronary interventions | 9 |
| | Mortality with mitral valve surgery | 3 |
| Aortic valve replacement | Mortality expected given hospital volume | 49 |
| | Mortality with coronary artery bypass surgery | 18 |
| | Mortality | 16 |
| | Mortality with percutaneous coronary interventions | 12 |
| | Mortality with mitral valve surgery | 4 |
| Percutaneous coronary interventions | Mortality expected given hospital volume | 49 |
| | Mortality | 46 |
| | Mortality with coronary artery bypass surgery | 5 |
| Pancreatic cancer resection | Mortality expected given hospital volume and other structural characteristics | 60 |
| | Mortality | 30 |
| | Mortality with colon surgery | 10 |
| Esophageal cancer resection | Mortality expected given hospital volume and other structural characteristics | 48 |
| | Mortality with colon surgery | 29 |
| | Mortality | 14 |
| | Mortality with aortic valve surgery | 9 |

conditions own mortality rate varied from 46 percent with percutaneous coronary interventions to only 14 percent with esophageal resection (Table 2). For esophagectomy, the least common operation, the weight placed on mortality with another, related operation (colon resection) was higher than the weight placed on its own mortality (29 versus 14 percent).

*Ability of the Composite to Predict Future Performance*

The composite score created by combining these individual measures performed well at predicting future hospital performance (Figure 1). Compared to 1-star hospitals, risk-adjusted mortality was much lower at 3-star hospitals for esophagectomy (6.7 versus 14.4 percent), pancreatectomy (4.7 versus 9.2 percent), coronary artery bypass surgery (2.6 versus 5.0 percent), aortic valve replacement (4.5 versus 8.5 percent), and percutaneous coronary interventions (2.4 versus 4.1 percent). These differences in mortality were not explained by observable differences in patient severity of illness, as the differences in patient characteristics shown in Table 3 were adjusted for in all comparisons.

For all five procedures, the composite measure based on 2005–2006 data was better at discriminating future performance in 2007–2008 when compared

Figure 1: Future Risk-Adjusted Mortality Rates (2007–2008) for 1-Star, 2-Star, and 3-Star Hospitals as Assessed Using the Composite Measure from the Two Previous Years (2005–2006)

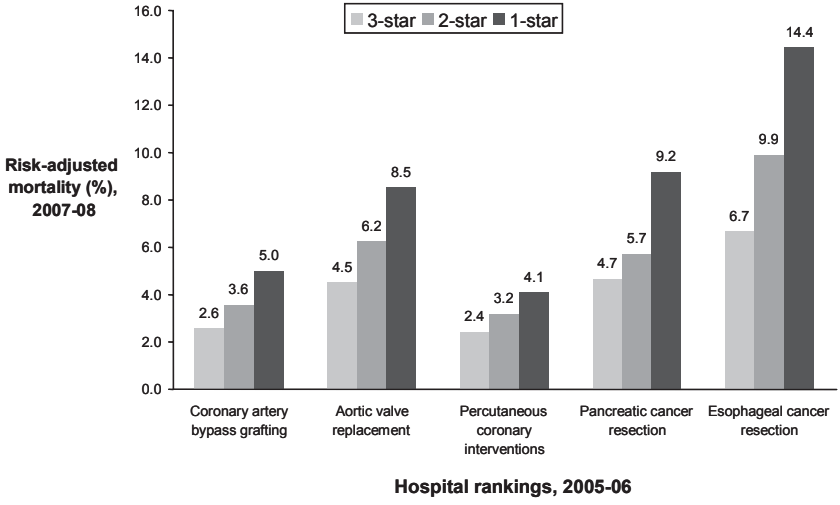

Table 3: Patient Characteristics for 1-Star, 2-Star, and 3-Star Hospitals in 2005–2006

| | *Hospital Rankings (2005–2006)* | | |
|---|---|---|---|
| *Procedure* | *1-Star Bottom 20% of Hospitals* | *2-Star Middle 60% of Hospitals* | *3-Star Top 20% of Hospitals* |
| Coronary artery bypass grafting | | | |
| Age, mean | 74.5 | 74.8 | 74.9 |
| Black race,(%) | 6.3 | 5.3 | 4.3 |
| Male,(%) | 67 | 68 | 68 |
| Urgent admission,(%) | 25 | 28 | 31 |
| Bottom third of socioeconomic position, (%) | 30 | 26 | 19 |
| Two or more comorbid diseases, (%) | 40 | 41 | 40 |
| Aortic valve replacement | | | |
| Age, mean | 76.5 | 76.9 | 77.2 |
| Black race, (%) | 4.3 | 3.8 | 2.4 |
| Male, (%) | 57 | 57 | 57 |
| Urgent admission, (%) | 18 | 21 | 18 |
| Bottom third of socioeconomic position, (%) | 27 | 18 | 8.4 |
| Two or more comorbid diseases, (%) | 39 | 40 | 38 |
| Percutaneous coronary interventions | | | |
| Age, mean | 74.5 | 74.8 | 74.9 |
| Black race, (%) | 6.3 | 5.3 | 4.3 |
| Male, (%) | 67 | 68 | 68 |
| Urgent admission, (%) | 25 | 28 | 31 |
| Bottom third of socioeconomic position, (%) | 27 | 23 | 21 |
| Two or more comorbid diseases, (%) | 40 | 41 | 41 |
| Pancreatic cancer resection | | | |
| Age, mean | 74.7 | 74.9 | 74.7 |
| Black race, (%) | 8.3 | 6.3 | 5.2 |
| Male, (%) | 49 | 48 | 47 |
| Urgent admission, (%) | 14 | 12 | 5.0 |
| Bottom third of socioeconomic position, (%) | 23 | 17 | 14 |
| Two or more comorbid diseases, (%) | 56 | 54 | 50 |
| Esophageal cancer resection | | | |
| Age, mean | 75.1 | 74.5 | 74.2 |
| Black race, (%) | 7.9 | 5.9 | 4.2 |
| Male, (%) | 68 | 71 | 71 |
| Urgent admission, (%) | 11 | 10 | 6.9 |
| Bottom third of socioeconomic position, (%) | 22 | 17 | 15 |
| Two or more comorbid diseases, (%) | 55 | 51 | 49 |

to risk-adjusted mortality (Table 4). For example, with coronary artery bypass surgery, historical mortality predicted a smaller difference between the 1-star (bottom 20 percent) and 3-star (top 20 percent) hospitals (OR, 1.61; 95 percent

Table 4: Relative Ability of Hospital Rankings Based on Individual and Composite Measures from 2005 to 2006 to Forecast Risk-Adjusted Mortality in 2007–2008

| | *Adjusted Odds Ratio for Risk-Adjusted Mortality in 2007-08, 3-Star (top 20%) versus 1-Star (bottom 20%) from 2005 to 2006 Hospital Rankings(95% CI)* | | | | |
|---|---|---|---|---|---|
| *Procedure* | *Hospital Volume* | *Operative Mortality* | *Reliability-Adjusted Mortality* | *Hospital Compare* | *Composite Measure* |
| Coronary artery bypass grafting | 1.46 (1.35–1.59) | 1.61 (1.48–1.75) | 1.88 (1.67–2.11) | 1.25 (1.15–1.36) | 2.10 (1.93–2.28) |
| Aortic valve replacement | 1.65 (1.49–2.34) | 1.52 (1.38–1.69) | 1.76 (1.51–2.06) | 1.11 (1.00–1.23) | 2.10 (1.89–2.34) |
| Percutaneous coronary interventions | 1.43 (1.36–1.51) | 1.45 (1.38–1.52) | 1.68 (1.55–1.81) | 1.15 (1.10–1.21) | 1.81 (1.72–1.91) |
| Pancreatic cancer resection | 2.41 (1.63–3.58) | 1.44 (1.06–1.94) | 2.12 (1.40–3.19) | 1.40 (1.02–1.91) | 3.29 (2.27–4.77) |
| Esophageal cancer resection | 2.58 (1.76–3.79) | 1.25 (0.95–1.65) | 2.28 (1.55–3.35) | 1.15 (.87–1.53) | 3.91 (2.74–5.58) |

CI, 1.48–1.75) when compared to the composite measure (OR, 2.10; 95 percent CI, 1.93–2.28) (Table 4). For all five procedures, the composite measures based on 2005–2006 data were better at discriminating future performance in 2007–2008 when compared to hospital volume (Table 4). For example, with aortic valve replacement, hospital volume predicted a smaller difference between the 1-star (bottom 20 percent) and 3-star (top 20 percent) hospitals (OR, 1.65; 95 percent CI, 1.89–2.34) when compared to the composite measure (OR, 2.10; 95 percent CI, 1.89–2.34) (Table 4). Although reliability adjusted mortality (i.e., empirical Bayes shrinkage) performed better than risk-adjusted mortality assessed using standard techniques, the composite measure was better at discriminating future performance for all five procedures (Table 4).

Composite measures were also much better at discriminating future performance than the measures publicly reported on the Hospital Compare website. When comparing the 1-star (bottom 20 percent) to 3-star (top 20 percent) hospitals, the composite predicted much larger differences than the Hospital Compare ratings for all five procedures (Table 4). For example, with pancreatic cancer resection, the difference in the risk of future mortality between 1-star and 3-star hospitals was much smaller (OR 1.40; 95 percent CI, 1.02–1.91) than the difference between 1-star and 3-star ratings based on the composite measure (OR, 2.28; 95 percent CI, 1.55–3.36).

Table 5: Proportion of Subsequent (2007–2008) Hospital-Level Variation in Risk-Adjusted Mortality Explained by Rankings for Each Quality Measure Assessed in the Prior Two Years (2005–2006)

| | *Proportion of Subsequent (2007-08) Hospital-Level Variation Explained by Measures from 2005 to 2006* | | | | |
|---|---|---|---|---|---|
| *Procedure* | *Hospital Volume (%)* | *Operative Mortality (%)* | *Reliability-Adjusted Mortality (%)* | *Hospital Compare (%)* | *Composite Measure (%)* |
| Coronary artery bypass grafting | 14 | 27 | 38 | 5 | 54 |
| Aortic valve replacement | 18 | 12 | 21 | 3 | 47 |
| Percutaneous coronary interventions | 16 | 15 | 35 | 4 | 45 |
| Pancreatic cancer resection | 59 | 7 | 41 | 3 | 98 |
| Esophageal cancer resection | 60 | 8 | 50 | 2 | 92 |

The composite measures were also much better at explaining systemic hospital-level variation in risk-adjusted mortality in the next 2 years (Table 5). Although the ability to explain future risk-adjusted mortality varied across measures, the composite measure outperformed all individual measures, including reliability adjustment (i.e., simple empirical Bayes shrinkage (Table 5).

## DISCUSSION

In this study, we investigated the value of empirically weighted composite measures for assessing surgical performance. We found that several input measures explained a large proportion of hospital-level variation in risk-adjusted mortality, but the relative importance of each measure varied across procedures. Composite measures combining various types of information about quality were better at forecasting future performance than existing quality indicators, including hospital volume, risk-adjusted mortality, reliability adjusted mortality, and the quality ratings reported by Center for Medicare and Medicaid Services (CMS) on its Hospital Compare website.

There is growing interest in composite measures of performance in health care (Peterson et al. 2011). Recent pay-for-performance programs,

including the Center for Medicare and Medicaid Services (CMS)/Premier pilot program, use composite quality measures for several medical conditions and surgical procedures (O'Brien et al. 2007). The Society of Thoracic Surgeons (STS), which maintains a large national registry in cardiac surgery, recently created a composite measure of hospital performance by combining process and outcome measures (Shahian et al. 2007). Like our composite measures, these approaches are created by combining multiple input measures. However, they were designed with distinctly different goals in mind. The CMS and STS composite scores aim to provide a summary score of multiple domains of quality. In contrast, our measure was designed to optimize the prediction of one particularly central measure of quality—risk-adjusted mortality. As a result, we use a different approach to weighting input measures. Many existing approaches for creating composite measures, including those of the CMS and STS, assign equal weight to all measures (i.e., the all-or-none approach) or weight measures according to expert opinion. Our method differs from existing composites by providing an empirical weighting process that takes into account the importance of each input.

Our findings suggest that weight-applied composite measures need to be tailored to the specific operation. We found that the reliability of mortality as a hospital quality measure varies dramatically across procedures. For very common operations, such as coronary artery bypass, more weight is placed on the mortality rate, largely because it is measured with more precision. At the other end of the spectrum, less common operations like pancreatic and esophageal cancer resection are not performed often enough to measure mortality precisely, and very little weight should be placed on the mortality rate. Another reason that weighting of inputs to a composite measure should be tailored to the procedure is that hospital volume matters more for some operations than others. It is well known that the strength of the volume outcome relationship varies across procedures (e.g., very strong for pancreatic and esophageal resection, much less so for coronary artery bypass) (Halm, Lee, and Chassin 2002). Prior to weighting measures in a composite score, this relationship should be systematically evaluated and used to guide the empirical weighting of input measures.

Given the value of our composite measure in predicting future hospital outcomes, we believe our measures would be particularly valuable for public reporting or value-based purchasing. In such contexts, arguably the most important criterion of their usefulness is the extent to which measures based on historical information can predict outcomes here and now (Birkmeyer, Dimick, and Staiger 2006). In quality improvement contexts, however,

information about past performance is arguably most relevant to help hospitals target quality improvement efforts. While our composite measures perform well in discriminating hospitals on their historical performance, their summary nature makes them more limited (i.e., less actionable) for purposes of quality improvement.

We should acknowledge several limitations to this study. Because we used Medicare claims data, our adjustment for patient case mix is limited. Although we adjusted for several patient variables, including age, gender, race, urgency or admission, socioeconomic status, and secondary diagnoses, problems with risk adjustment using administrative data are well known (Iezzoni 1997). If differences in patient risk varied systematically across hospitals, our analysis would tend to overestimate the ability of hospital ratings to forecast future mortality. However, random, year-to-year differences in patient risk would bias our results toward a null finding and lead to an underestimation of the predictive power of composite measures. While there is little empiric data establishing whether these differences are random or systematic, there is a growing body of evidence suggesting that hospital case-mix may not vary substantially, especially among patients undergoing the same surgical procedure (Dimick and Birkmeyer 2008).

The use of Medicare claims data limits our study in several other ways. First, we used Medicare fee-for-service volume rather than total hospital volume. Although these volume measures are highly correlated, a more complete ascertainment of hospital volume would likely further improve the predictive ability of our measure. Second, administrative data are limited in their ability to accurately ascertain nonfatal complications. We focused on a subset of complications previously shown to have a high sensitivity and specificity on medical chart review (Lawthers et al. 2000; Weingart et al. 2000). Using data from a clinical registry, where complications are determined from the medical chart based on rigorous definitions could potentially enhance the ability of our composite measure to predict future performance. Although there are a growing number of hospitals participating in clinical registries in surgery (i.e., the National Surgical Quality Improvement Program), such detailed data are not currently available for most U.S. hospitals (Khuri et al. 2008).

Our findings also suggest that surgical quality measures publicly reported by CMS on the Hospital Compare website are not ideal for helping patients identify the safest hospitals for surgery. The Hospital Compare measures are processes of care related to preventing surgical complications, as developed for the Surgical Care Improvement Program (SCIP). Although these measures were selected because of clinical trials linking them to better

outcomes, there is growing evidence that these processes do not account for hospital-level variations in important surgical outcomes, such as complications and mortality (Hawn 2010). The composite measures described in this study would be much better at helping patients and payers identify low-mortality hospitals for high-risk surgery.

Our findings may also have implications for the Hospital Compare measures for nonsurgical conditions. CMS publicly reports risk-standardized mortality and readmission rates for several common, inpatient medical conditions, such as acute myocardial infarction, congestive heart failure, and pneumonia. The modeling strategy used by CMS for these measures addresses the problem of small hospital size (i.e., statistical "noise") using Bayesian analyses somewhat analogous to those in this article, with one key difference. The Hospital Compare measures do not include hospital volume or other structural variables in their modeling strategy. Because of this exclusion, they implicitly make the assumption that small hospitals have average performance, which is not supported by the empirical data (Ross et al. 2010). As shown in this article, hospital volume and other structural variables are important for optimally predicting a hospitals future performance. Silber et al. have recently brought attention to this issue and assert that a measure based on mortality and hospital volume combined would more accurately reflect "true" hospital performance (Silber et al. 2010). Thus, the methods presented in this paper may have important implications outside surgery.

Numerous stakeholders would benefit from better measures of surgical quality. Patients would benefit by having access to data that could help them increase their chances of surviving surgery by choosing the right hospital. Payers and health care purchasers would benefit by having reliable measures—created using readily available data—that would help them identify high-quality hospitals for their selective referral and value-based purchasing programs. Although these composite measures could no doubt be improved with better inputs, they represent a significant advance over current surgical quality indicators.

## ACKNOWLEDGMENTS

*Joint Acknowledgment/Disclosure Statement:* The authors acknowledge Wenying Zhang at the Center for Healthcare Outcomes and Policy for assistance with data management and analysis.

*Funding:* This study was supported by a career development award to Dr. Dimick from the Agency for Healthcare Research and Quality (K08 HS017765) and a grant to Dr. Birkmeyer and Dr. Staiger from the National Institute of Aging (P01AG019783). The views expressed herein do not necessarily represent the views of the Center for Medicare and Medicaid Services or the U.S. Government.

*Prior presentation:* This work was presented at the American College of Surgeons Clinical Congress Papers Session in Washington, DC, October, 2010.

*Disclosures:* Drs. Dimick, Staiger, and Birkmeyer are paid consultants and have an equity interest in ArborMetrix, Inc, a company that provides software and analytics for assessing hospital quality and efficiency. The company had no role in the conduct of the study herein.

## REFERENCES

Birkmeyer, J. D., J. B. Dimick, and N. J. Birkmeyer. 2004. "Measuring the Quality of Surgical Care: Structure, Process, or Outcomes?" *Journal of the American College of Surgeons* 198 (4): 626–32.

Birkmeyer, J. D., J. B. Dimick, and D. O. Staiger. 2006. "Operative Mortality and Procedure Volume as Predictors of Subsequent Hospital Performance." *Annals of Surgery* 243 (3): 411–7.

Birkmeyer, J. D., A. E., Siewers, E. V. Finlayson, T. A. Stukel, F. L. Lucas, I. Batista, H. G. Welch, and D. E. Wennberg. 2002. "Hospital Volume and Surgical Mortality in the United States." *New England Journal of Medicine* 346 (15): 1128–37.

Dimick, J. B., and J. D. Birkmeyer. 2008. "Ranking Hospitals on Surgical Quality: Does Risk-Adjustment Always Matter?" *Journal of the American College of Surgeons* 207 (3): 347–51.

Dimick, J. B., D. O. Staiger, and J. D. Birkmeyer. 2006. "Are Mortality Rates for Different Operations Related? Implications for Measuring the Quality of Noncardiac Surgery." *Medical Care* 44 (8): 774–8.

———. 2010. "Ranking Hospitals on Surgical Mortality: The Importance of Reliability Adjustment." *Health Services Research* 45 (6 Pt 1): 1614–29.

Dimick, J. B., H. G. Welch, and J. D. Birkmeyer. 2004. "Surgical Mortality as an Indicator of Hospital Quality: The Problem with Small Sample Size." *Journal of the American Medical Association* 292 (7): 847–51.

Dimick, J. B., D. O., Staiger, O. Baser, and J. D. Birkmeyer. 2009. "Composite Measures for Predicting Surgical Mortality in the Hospital." *Health Affairs (Millwood)* 28 (4): 1189–98.

Galvin, R. S. 2001. "The Business Case for Quality." *Health Affairs (Millwood)* 20 (6): 57–8.

Goodney, P. P., O'Connor, G. T., D. E. Wennberg, and J. D. Birkmeyer. 2003. "Do Hospitals with Low Mortality Rates in Coronary Artery Bypass Also Perform

Well in Valve Replacement?" *Annals of Thoracic Surgery* 76 (4): 1131–6; discussion 1136-7.

Halm, E. A., C. Lee, and M. R. Chassin. 2002. "Is Volume Related to Outcome in Health Care? A Systematic Review and Methodologic Critique of the Literature." *Annals of Internal Medicine* 137 (6): 511–20.

Hawn, M. T. 2010. "Surgical Care Improvement: Should Performance Measures Have Performance Measures." *Journal of the American Medical Association* 303 (24): 2527–8.

Iezzoni, L. I.. 1997. "Assessing Quality Using Administrative Data." *Annals of Internal Medicine* 127 (8 Pt 2): 666–74.

Khuri, S. F., W. G., Henderson, J. Daley, O. Jonasson, R. S. Jones, D. A. Campbell Jr, A. S. Fink, R. M. Mentzer Jr, L. Neumayer, K. Hammermeister, C. Mosca, and N. Healey. 2008. "Successful Implementation of the Department of Veterans Affairs' National Surgical Quality Improvement Program in the Private Sector: The Patient Safety in Surgery Study." *Annals of Surgery* 248 (2): 329–36.

Kizer, K. W. 2001. "Establishing Health Care Performance Standards in an Era of Consumerism." *Journal of the American Medical Association* 286 (10): 1213–7.

Lawthers, A. G., E. P., McCarthy, R. B. Davis, L. E. Peterson, R. H. Palmer, and L. I. Iezzoni. 2000. "Identification of In-Hospital Complications from Claims Data: Is It Valid?" *Medical Care* 38 (8): 785–95.

O'Brien, S. M., E. R., DeLong, R. S. Dokholyan, F. H. Edwards, and E. D. Peterson. 2007. "Exploring the Behavior of Hospital Composite Performance Measures: An Example from Coronary Artery Bypass Surgery." *Circulation* 116 (25): 2969–75.

Peterson, E. D., E. R., Delong, F. A. Masoudi, S. M. O'Brien, P. N. Peterson, J. S. Rumsfeld, D. M. Shahian, R. E. Shaw, D. C. Goff Jr, K. Grady, L. A. Green, K. J. Jenkins, A. Loth, and M. J. Radford. 2011. "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)." *Circulation* 121 (15): 1780–91.

Ross, J. S., S. L., Normand, Y. Wang, D. T. Ko, J. Chen, E. E. Drye, P. S. Keenan, J. H. Lichtman, H. Bueno, G. C. Schreiner, and H. M. Krumholz. 2010. "Hospital Volume and 30-Day Mortality for Three Common Medical Conditions." *New England Journal of Medicine* 362 (12): 1110–8.

Shahian, D. M., F. H., Edwards, V. A. Ferraris, C. K. Haan, J. B. Rich, S. L. Normand, E. R. DeLong, S. M. O'Brien, C. M. Shewan, R. S. Dokholyan, and E. D. Peterson. 2001. "Cardiac Surgery Report Cards: Comprehensive Review and Statistical Critique." *Annals of Thoracic Surgery* 72 (6): 2155–68.

———. 2007. "Quality Measurement in Adult Cardiac Surgery: Part 1–Conceptual Framework and Measure Selection." *Annals of Thoracic Surgery* 83 (4 suppl): S3–12.

Silber, J. H., P. R., Rosenbaum, T. J. Brachet, R. N. Ross, L. J. Bressler, O. Even-Shoshan, S. A. Lorch, and K. G. Volpp. 2010. "The Hospital Compare Mortality

Model and the Volume-Outcome Relationship." *Health Services Research* 45 (5 Pt 1): 1148–67.

Southern, D. A., H. Quan, and W. A. Ghali. 2004. "Comparison of the Elixhauser and Charlson/Deyo Methods of Comorbidity Measurement in Administrative Data." *Medical Care* 42 (4): 355–60.

Staiger, D. O., J. B., Dimick, O. Baser, Z. Fan, and J. D. Birkmeyer. 2009. "Empirically Derived Composite Measures of Surgical Performance." *Medical Care* 47 (2): 226–33.

Stulberg, J. J., P., Delaney, Neuhauser, D. C. Aron, P. Fu, and S. M. Koroukian. 2010. "Adherence to Surgical Care Improvement Project Measures and the Association with Postoperative Infections." *Journal of the American Medical Association* 303 (24): 2479–85.

Weingart, S. N., L. I., Iezzoni, R. B. Davis, R. H. Palmer, M. Cahalane, M. B. Hamel, K. Mukamal, R. S. Phillips, D. T. Davies Jr, and N. J. Banks. 2000. "Use of Administrative Data to Find Substandard Care: Validation of the Complications Screening Program." *Medical Care* 38 (8): 796–806.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.
Appendix S1: Technical Appendix.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.