Composite Measures for Predicting Hospital Mortality with Surgery

Project Manager:

Justin B. Dimick, MD, MPH

Assistant Professor of Surgery

University of Michigan

Research Director:

John D. Birkmeyer, MD

Professor of Surgery

University of Michigan

February, 2008

Executive Summary:

Individual quality measures have significant limitations for assessing surgical performance. Despite growing interest in composite measures, methods for empirically combining multiple domains of surgical quality are not well established. This paper outlines the methods for creating a simple composite score based on a combination of surgical mortality and hospital volume. The methods used are based on empirical Bayes techniques, where the mortality rate is weighted to the extent it is reliable and the remaining weight is placed on hospital volume. We validated this composite measure by assessing how well it predicts subsequent performance. This perspective is particularly relevant for selective-referral or public reporting contexts. Our results demonstrate that a simple composite of mortality and volume is a better predictor of subsequent performance than either measure alone for most operations. These measures will help patients and payers identify hospitals likely to have superior outcomes.

Research Team:

Justin B. Dimick, MD MPH
Assistant Professor of Surgery, Department of Surgery, University of Michigan
M-SCORE offices, 211 N. Fourth Avenue, Suite 301, Ann Arbor, MI   48104
Phone: (734) 998-7470
Fax: (734) 998-7473
Email: jdimick@umich.edu

Douglas O. Staiger, PhD
Professor of Economics, Department of Economics and the Dartmouth Institute for
Health Policy and Clinical Practice
6106 Rockefeller Hall, Dartmouth College, Hanover, NH 03755
Phone: (603) 646-2979
Fax: (603) 646-2122
Email: Doug.Staiger@Dartmouth.edu

Onur Baser, PhD
Assistant Professor of Surgery, Department of Surgery, University of Michigan
M-SCORE offices, 211 N. Fourth Avenue, Suite 201&202, Ann Arbor, MI   48104
Phone: (734) 998-7470
Fax: (734) 998-7473
Email: onur@umich.edu

Zhaohui Fan, MS
Statistical Analyst, Michigan Surgical Collaborative for Outcomes Research and
Evaluation (M-SCORE), University of Michigan
M-SCORE offices, 211 N. Fourth Avenue, Suite 201&202, Ann Arbor, MI   48104
Phone: (734) 998-7470
Fax: (734) 998-7473
Email: fanluck@umich.edu

John D. Birkmeyer, MD
Professor of Surgery and Director, Michigan Surgical Collaborative for Outcomes
Research and Evaluation (M-SCORE), Department of Surgery, University of Michigan
M-SCORE offices, 211 N. Fourth Avenue, Suite 201&202, Ann Arbor, MI   48104
Phone: (734) 998-7470
Fax: (734) 998-7473
Email: jbirkmey@umich.edu

INTRODUCTION:

Aiming to foster accountability and encourage quality improvement, payers are increasingly collecting and reporting information on hospital quality with surgery (1,2). Because reliable, all-payer patient databases are not widely available, these efforts often rely on self-reported information from hospitals. For example, the Leapfrog Group, a large coalition of private payers, asks hospitals for their caseloads and number of deaths with 7 different procedures. This information is used to categorize hospitals for purposes of public reporting or selective contracting (3).

However, it remains unclear whether this information is useful for identifying the best hospitals with surgery. There are at least two reasons to question the value of these data. First, the individual measures most often used to evaluate surgical performance are flawed. Mortality rates are often too "noisy" to reflect hospital quality with surgery (4,5). In addition, although very important for some operations, hospital volume is a weak proxy for performance with most procedures (6). Second, when multiple measures are considered, it is not clear how to best weight them or interpret them when they conflict. For example, some hospitals will have high volumes but high mortality; others will have low volume but low mortality. It is not clear which group of hospitals is likely to have better outcomes.

In this paper, we describe a simple composite score created using hospital volume and observed mortality, as might be self-reported by hospitals. Given how publicly reported information on quality is likely to be used by patients and payers, we assessed the value of the composite at predicting future hospital performance. Because self-

reported mortality rates are not risk-adjusted, we also evaluated the extent to which risk-adjustment is important for predicting hospital outcomes.

METHODS:

Data source and study population. We used data from the Medicare Analysis Provider and Review (MEDPAR) files, which contains 100% of Medicare hospitalizations. 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 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 to 99 undergoing the following six operations: coronary artery bypass grafting, aortic valve replacement, abdominal aortic aneurysm repair; percutaneous coronary interventions, and resection of pancreatic and esophageal cancer. We chose these operations given their relevance to ongoing value-based purchasing initiatives, including the Leapfrog Group's evidence-based hospitals referral initiative. In keeping with Leapfrog data specifications, 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., ruptured aortic aneurysms) (3).

Development of the composite measure. We used an empirical Bayes approach to combine mortality rates with information on hospital volume at each hospital. In traditional empirical Bayes methods, a point estimate (e.g., mortality rate observed at a hospital) is adjusted for reliability by shrinking it towards the overall mean (e.g., overall mortality rate in the population) (7,8). We modified this traditional approach by shrinking the observed mortality rate back toward the mortality rate expected given the volume at that hospital—we refer to this as the "volume-predicted mortality" (See the TECHNICAL APPENDIX for the mathematical details of this method). With this approach, the observed mortality rate is weighted according to how reliably it is estimated, with the remaining weight placed on the information regarding hospital volume. Because this method includes observed data to the extent that it is useful, and only relies on the proxy measure to the extent necessary, it ensures an optimal combination of these two quality domains.

The two inputs to the composite measure are mortality rates and procedure volume for each of the six included operations. Procedure-specific mortality rates were calculated for all hospitals over a 2-year period (2000-01) and this was used as the first input. Hospital volume was calculated as the number of Medicare cases performed during the same time period. For each operation, the relationship between hospital volume and risk-adjusted mortality was modeled using linear regression. (Details of the risk-adjustment strategy will be discussed below.) After testing the fit of several transformations, hospital volume was modeled as the natural log of the continuous volume variable, which is the same approach used in our previous work (9). Using this

regression model, we estimated the volume-predicted mortality, the second input to the composite measure.

We then used the empirical Bayes approach to create an optimal combination of these two inputs. This composite measure theoretically provides the best estimate of a hospitals true mortality rate, taking into account the both available inputs (7,8). The combined measure was calculated as follows: composite mortality prediction = (weight)*(observed mortality) + (1-weight)*(volume-predicted mortality). The weight placed on the point estimate of mortality is the reliability, or ratio of signal to signal plus noise, calculated as follows: weight = variation among hospitals/(variation among hospitals + variation within hospitals). The variation among hospitals was calculated as the variance in observed mortality rates for the hospitals included in the sample. The variation within hospitals was calculated as the standard error of the mortality rate at each hospital. With this method, more weight is placed on the observed mortality rate when a hospital has a high number of cases because it is estimated with more reliability; less weight is placed on the observed mortality rate when a hospital performs a low number of cases because of its lower reliability.

Sensitivity analysis. We performed a sensitivity analysis to determine whether risk-adjustment of the mortality input was important in improving the predictive ability of the composite measure. Risk-adjustment was performed using logistic regression to estimate expected mortality rates for each hospital based on patient age, gender, race, urgency of operation, median income, and coexisting diseases. Coexisting diseases were determined from secondary diagnostic codes using the methods of Elixhauser (10). The observed mortality rate at each hospital was then divided by the expected mortality rate to

yield the ratio of observed/expected deaths (O/E ratio). The O/E ratio was multiplied by the average mortality rate for each operation to yield a risk-adjusted mortality rate. To determine the value of risk-adjustment in the context of selective referral, we compared the ability of risk-adjusted and unadjusted composite measures to predict subsequent performance.

Validating the composite measure. We determined the value of our composite measure by establishing whether it explained hospital-level variation in risk-adjusted mortality rates and by assessing to what degree it was able to predict future hospital performance. We first estimated the proportion of variation in hospital-level mortality (2000-01) explained by the composite measure using random effects logistic regression models. For these analyses, we estimated the proportional change in the hospital-level variance in mortality rates, which was determined from the standard deviation of the random effect, after adding each measure to the model (8,11). We next compared the ability of the composite measure to the individual measures, mortality rates and hospital volume. We should note that these analyses focus on explaining systematic, or non-random, variation, since measurement error (random error) is accounted for and subtracted from the total variation in all analyses (8,12).

We next determined the extent to which the composite measure predicts future risk-adjusted mortality. For this analysis, hospitals were ranked based on each measure from the earlier time period (data from years 2000-01) and divided into four equal size groups (quartiles at the patient level). The subsequent risk-adjusted mortality rates for each quartile of performance were then calculated (data from years 2002-03). We present the subsequent mortality rates across quartiles of the composite measure to graphically

demonstrate its usefulness in discriminating among hospitals for the entire spectrum of

performance. To compare the predictive ability of the composite measures and individual

measures, we also present the subsequent mortality rates in the "worst" compared to the

"best" quartile. All statistical analyses were conducted using STATA 10.0 (College

Station, Texas).

RESULTS:

Hospital caseloads and the weights applied to each input to the composite measure varied for each procedure (**Table 1**). For coronary artery bypass and percutaneous coronary interventions, the two procedures with the highest hospital caseloads, more weight was placed on the mortality input (46% and 48%, respectively). At the other end of the spectrum, for esophageal resection, the operation with by far the lowest caseload, much less weight was placed on the mortality input (14%).

Weights placed on the mortality input also varied across hospitals for the same procedure. For hospitals with higher caseloads of a particular operation, more weight was placed on the mortality input compared to hospitals with lower caseloads. For example, consider a hospital that performs 40 abdominal aortic aneurysm repairs over a two year period, with 4 deaths. The observed mortality rate would be 10% and the weight placed on the mortality component would be 30%. Since this is a high volume hospital, its volume-predicted mortality is estimated at 2.5%. The composite mortality measure is then calculated as follows: (10%)(0.30) + (2.5%)(0.70) = 4.75%. In contrast, consider a hospital performing only 10 abdominal aortic aneurysm repairs without any deaths. This hospital has a mortality rate of 0%, a weight applied to the mortality input of 15%, and a volume-predicted mortality of 5.0%; and the composite mortality prediction is as follows: (0%)(0.15) + (5.0%)(0.85) = 4.25%.

The composite measure explained a large proportion of non-random, hospital-level variation in risk-adjusted mortality rates (**Table 2**). The amount of variation explained by the composite measure varied from 41% for abdominal aortic aneurysm repair to 66% for percutaneous coronary interventions. While the composite measure

explained more variation than either measure alone for all 6 operations, the relative importance of the individual measures varied across procedures (**Table 2**). For the more common operations, such as coronary artery bypass, mortality rates explained a large proportion of the variation (46%), and hospital volume explained a small proportion (9%). For less common operations, such as pancreatic resection, hospital volume explained a much larger proportion of the variation (57%) than mortality rates (23%).

The composite measure predicted large differences in future risk-adjusted mortality between the "best" and "worst" hospital quartiles (**Table 3**). The best prediction was achieved with pancreatic resection, with greater than 4-fold differences between the "worst" and "best" quartiles (Odds Ratio [OR], 4.45; 95% CI, 3.12 to 6.67). The composite measure was least predictive for coronary artery bypass, but future mortality rates were still 1.7 times greater in "worst" compared to the "best" quartile (OR, 1.70; 95% CI, 1.57-1.84). When compared to the individual measures by themselves, the composite was better at predicting differences between the "best" and "worst" quartiles for all 6 operations (**Table 3**). In addition to better discriminating between the extremes of performance, the composite measure was also better at predicting mortality in the intermediate strata of performance (**Figure**). In sensitivity analysis, composite measures based on an unadjusted mortality input and a risk-adjusted mortality input were equally good a predicting future performance (**Figure**).

DISCUSSION:

Although information on hospital quality is increasingly collected and reported, the usefulness of much of this data is uncertain. In the present paper, we assessed the value of a composite measure—based only on the hospital case count and the number of deaths—for predicting future hospital performance. We found that this simple composite measure based on widely available data explained a great deal of hospital-level variation in mortality and predicted nearly 2-fold differences in future hospital performance between the best and worst hospital quartiles. In this regard, this simple composite measure performed better than individual measures for all operations.

Composite measures of performance are gaining increasing popularity in surgery. Most existing pay-for-performance programs use this approach to summarize hospital performance. For example, the Premier/Center for Medicare and Medicaid Services Hospital Quality Incentive Demonstration uses a composite of process and outcome to measure quality for coronary artery bypass surgery (13). The Society of Thoracic Surgeons' Task Force on Quality Measurement advocates a composite score based on a set of outcome and process measures endorsed by the National Quality Forum (14,15).

Despite the increasing interest in composite measures, there are major problems with existing approaches for creating them. Perhaps the biggest limitation relates to the weighting of the input measures. Existing approaches rely on overly simplistic approaches. Among these, assigning equal weight to all measures (i.e., the all or none approach) and relying on expert opinion are the most common. There are several reasons why these approaches may not make optimal use of available information. First, some measures are more reliable or more important than others. Second, as the present study

shows, the relative importance of individual measures varies by procedure. In addition,

the usefulness of these existing approaches to predict subsequent performance, arguably

the most important criteria of validity for public reporting or selective referral, has not

been established. In contrast, the methods set forth in this paper combine individual

measures using an empiric weighting process. The validity of this approach is proved by

the ability of the composite measure to reliably predict future performance. Our simple

composite measure also has the advantage of being relatively transparent and applied

using readily available data, including information that is self-reported or obtained from

administrative datasets.

Our findings also suggest that the weight placed on individual measures used in a

composite score should be tailored to the procedure. For very common operations, such

as coronary artery bypass surgery, much more weight should be 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 this measure. The weights applied to individual measures should also vary

across hospitals performing the same procedure. If a hospital performs a high number of

cases for a specific operation, their mortality will be measured more precisely than a

hospital that performs fewer cases of that operation.

Although perhaps important for face validity, we found that risk-adjustment was

not important in categorizing hospitals into performance groups. Composite measures

based on unadjusted and risk-adjusted mortality rates were equally good at predicting

future risk-adjusted mortality rates. Among potential reasons for this finding, it is

possible that illness severity for patients undergoing the same surgical procedure may not vary across systematically hospitals, especially when compared to patients in other clinical settings in which quality is measured (e.g., trauma or acute myocardial infarction). However, our conclusions about the importance of risk-adjustment must be tempered by our reliance on Medicare claims data for risk-adjustment, the limitations of which are well-described (16). Because of this limitation, we cannot exclude the possibility that composite measures created using mortality rates with more detailed risk-adjustment could result in even better predictive ability.

Relying on Medicare data also cuts our effective sample size by one half at each hospital, another limitation of using this data source. By limiting the sample sizes in this way, we have likely underestimated the reliability of the mortality rate input used in our composite measure. If all patients at a hospital were included, there would be a larger sample size and the mortality input would be more reliable. Composite measures based on more reliable mortality inputs would likely be even better at predicting future performance.

Although our study demonstrates the value of composite measures for categorizing hospitals, these measures may be an imperfect proxy for individual hospital performance. As with many other quality indicators, this limits the usefulness of this measure from the perspective of providers engaged in local quality improvement efforts. However, it is no worse in this regard, and may be better, than what we are currently using (e.g., hospital volume and mortality rates alone). The primary application of these measures would be public reporting and payer-led selective contracting. Patients are interested in a measure that will help them choose a hospital that will increase their

chances of surviving an operation. Payers are interested in selectively contracting with providers likely to have the best outcomes. The composite approach to measurement outlined in this paper is ideally suited for these purposes.

In summary, we have developed a simple composite measure that optimizes the use of readily available information and is good at predicting future hospital performance. Refinements of this approach will include incorporating more detailed inputs, such as nonfatal outcomes, clinical process measures, and outcomes with other, related procedures. Adding these other measures would likely improve the predictive ability of the composite measures. In the meantime, our simple composite measure will be better than existing alternatives at helping patients and payers to identify the safest hospitals for surgery.

REFERENCES:

1. Galvin RS. The business case for quality.Health Aff (Millwood). 2001;20:57-58.

2. Rosenthal MB, Dudley RA. Pay-for-performance: will the latest payment trend improve care? JAMA 2007;297:740-744.

3. The Leapfrog Group. Evidence-Based Hospital Referral Fact Sheet. http://www.leapfroggroup.org/, accessed November 6th, 2007.

4. 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.

5. Dimick JB, Welch HG. The zero mortality paradox in surgery. J Am Coll Surg (In press)

6. Halm EA, Lee C, Chassin MR. Is volume related to outcome in health care? A systematic review and methodologic critique of the literature. Ann Intern Med 2002;137:511-520

7. Morris CN. Parametric Empirical Bayes Inference: Theory and Applications. J Am Stat Assoc 1988;78:47-55.

8. McClellan MB, Staiger DO. Comparing the Quality of Health Care Providers. Alan Garber (ed.) Frontiers in Health Policy Research. Volume 3. 2000 The MIT Press: Cambridge MA, pp. 113-136.

9. Birkmeyer JD, Stukel TA, Siewers AE, et al. Surgeon volume and operative mortality in the United States. N Engl J Med. 2003;349:2117-2127.

10. 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.

11. Birkmeyer JD, Dimick JB, Staiger DO. Operative mortality and procedure volume as predictors of subsequent hospital performance. Ann Surg 2006;243:411-417.

12. 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.

13. http://www.cms.hhs.gov/HospitalQualityInits/35_HospitalPremier.asp, accessed November 6th, 2007.

14. 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 2007;83(4 Suppl):S3-12.

15. 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):S13-26.

16. Iezzoni LI. Assessing quality using administrative data. Ann Intern Med 1997;127:666-674.

Table 1. Hospital caseload and the weight applied to each of the two inputs in the composite measure. The weight applied to the mortality rate is the reliability and the weight applied to hospital volume is 1-reliability.

	Hospital caseloads mean (SD)	Weight applied to observed mortality, mean (SD)	Weight applied to hospital volume, mean (SD)
Coronary artery bypass grafting	321 (319)	.46 (.21)	.54 (.21)
Aortic valve replacement	59 (72)	.26 (.17)	.73 (.17)
Percutaneous coronary interventions	442 (499)	.48 (.27)	.52 (.27)
Elective abdominal aortic aneurysm repair	31 (46)	.21 (.18)	.79 (.18)
Pancreatic cancer resection	4 (10)	.23 (.17)	.77 (.17)
Esophageal cancer resection	4 (8)	.1 4(.13)	.86 (.13)

Table 2. Relative ability of each measure to explain hospital-level differences in risk-adjusted mortality rates (2000-01).

	Proportion of hospital-level variation in mortality rates explained by each performance measure (%)		
	Hospital volume	**Observed mortality**	**Simple composite measure**
Coronary artery bypass grafting	9	46	61
Aortic valve replacement	18	26	47
Percutaneous coronary interventions	12	48	66
Elective abdominal aortic aneurysm repair	28	21	41
Pancreatic cancer resection	57	23	59
Esophageal cancer resection	33	14	44

Table 3. Relative ability of historical measures (2000-01) to predict subsequent risk-adjusted mortality (2002-2003).

	Adjusted odds ratio for risk-adjusted mortality (2002-03), best vs. worst quartile (95% CI)		
	Hospital volume alone (2000-01)	**Observed mortality alone (2000-01)**	**Simple composite measure (2000-01)**
Coronary artery bypass grafting	1.14 (1.03-1.26)	1.56 (1.47-1.70)	1.70 (1.57-1.84)
Aortic valve replacement	1.19 (0.99-1.40)	1.73 (1.41-2.07)	1.79 (1.33-1.85)
Percutaneous coronary interventions	1.42 (1.29-1.50)	1.70 (1.58-1.81)	1.82 (1.69-1.92)
Abdominal aortic aneurysm repair	1.77 (1.52-2.06)	1.35 (1.19-1.53)	2.04 (1.78-2.35)
Pancreatic cancer resection	4.0 (2.53-6.34)	1.56 (1.2-2.04)	4.45 (3.12-6.67)
Esophageal cancer resection	2.06 (1.46-2.91)	1.14 (0.91-1.43)	2.99 (2.45-3.75)

Figure. Future risk-adjusted mortality rates (2002-02) for quartiles of hospital rankings based on historical (2000-01) hospital volume, risk-adjusted mortality rates, and composite measures. Future risk-adjusted mortality is shown for the composite measure created using both risk-adjusted and unadjusted mortality rates.
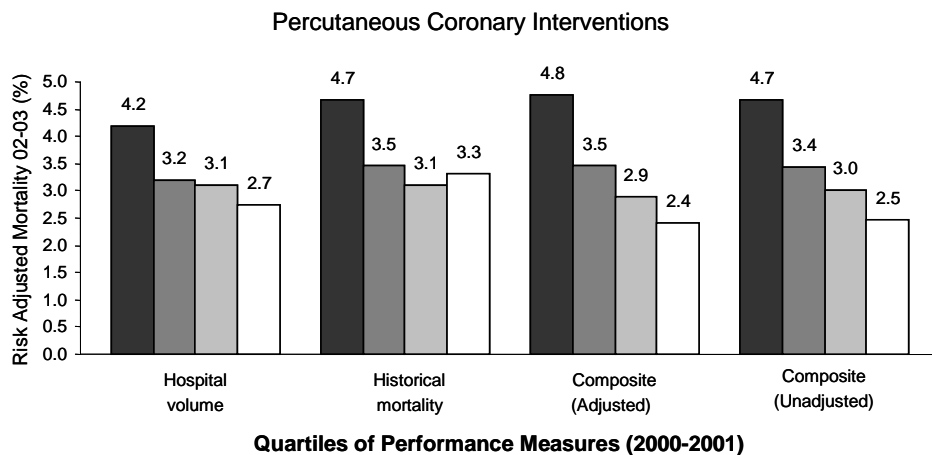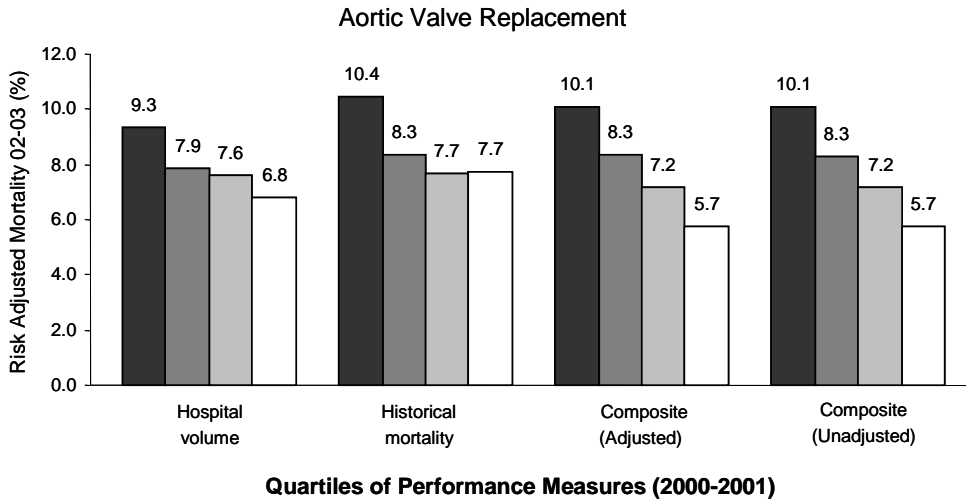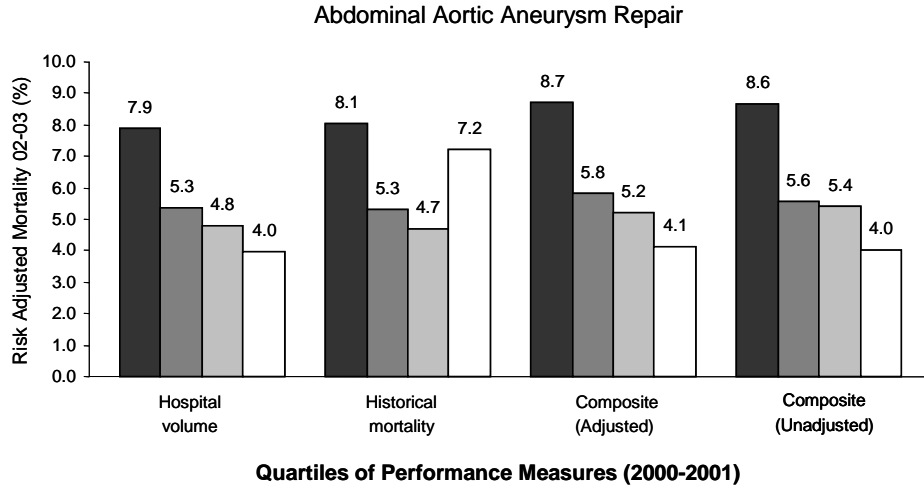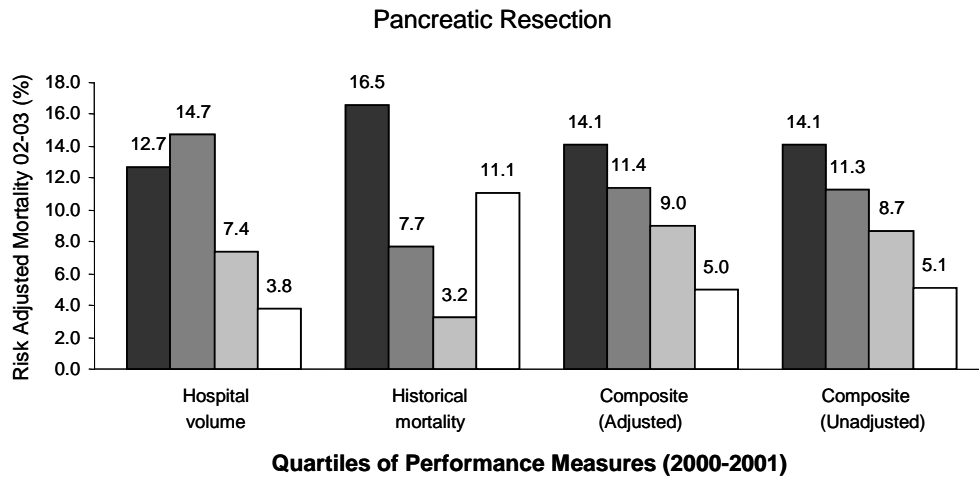
Abdominal Aortic Aneurysm Repair

Aortic Valve Replacement

Percutaneous Coronary Interventions

Figure. (continued)

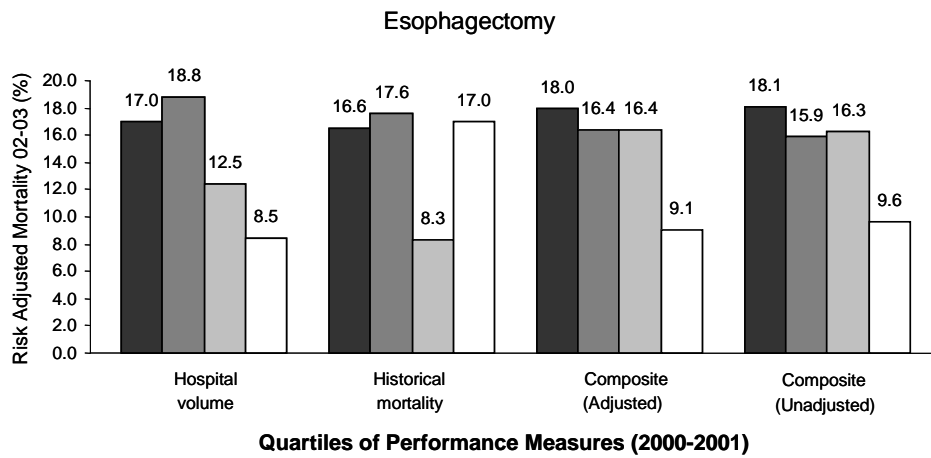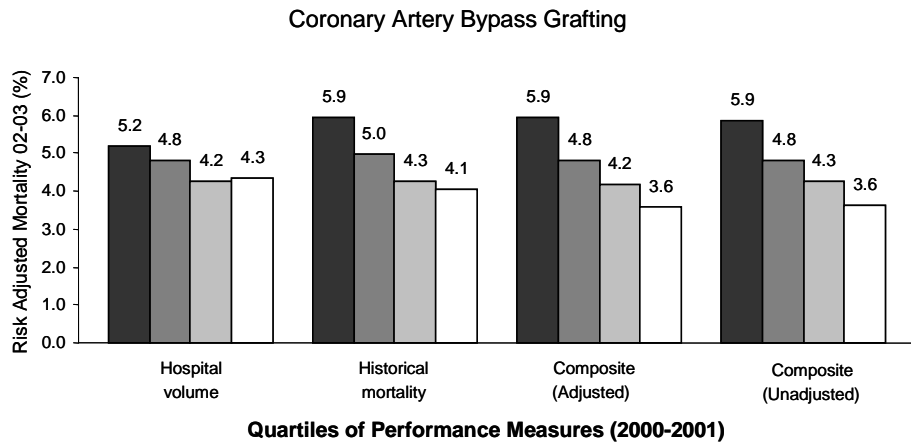### Coronary Artery Bypass Grafting

Esophagectomy

Pancreatic Resection

TECHNICAL APPENDIX:

Herein we describe the methods we used to combine two quality measures for each hospital—the mortality rate and hospital volume. We constructed the composite measure using unadjusted mortality rates and then performed a sensitivity analysis by creating a composite measure with risk-adjusted mortality rates. Because it is methodologically more complex, the method below details the creation of the second composite based on a risk-adjusted mortality rate. To create the composite based on the unadjusted mortality rate, we simply used the observed mortality rate rather than the ratio of observed to expected mortality.

Details of the estimation process. We construct risk-adjusted mortality rates for each hospital and each procedure using standard methods to compare outcomes across providers. For patients receiving a given procedure, let y_{ij} be a dichotomous indicator that is equal to 1 if patient j admitted to hospital i experienced the outcome (mortality). The risk-adjusted outcome (Y_i) for hospital i is the ratio of the observed (O_i) outcome rate to the expected (E_i) outcome rate, so that:

(1) $Y_i = O_i/E_i$, $O_i = \dfrac{1}{n_i}\sum_j y_{ij}$ and $E_i = \dfrac{1}{n_i}\sum_j \hat{p}_{ij}$

where n_i is the number of patients receiving the procedure at hospital i, and $\hat{p}_{ij} = pr(y_{ij}=1|X_{ij})$ is the predicted probability that the outcome occurred for each patient conditional on patient characteristics X. Estimates of the sampling variance for these measures are derived using standard methods. We derive the predicted probability that the outcome occurs for each patient ($\hat{p}_{ij}$) from a logistic regression model estimated on all patients. The dependent variable in the logistic model is operative mortality (y_{ij}) and the independent variables (X_{ij}) are the patient covariates available in the dataset.

The subsequent hospital-level analysis is based on a hierarchical model, in which data at the first (patient) level provides noisy estimates of structural parameters at the second (hospital) level. At the first level, the distribution of the mortality estimate conditional on the structural parameter is:

(1) $E(Y_i \mid \mu_i) = \mu_i$, and $Var(Y_i \mid \mu_i) = V_i$,

where Y_i is the risk-adjusted mortality rate for hospital i, μ_i is the corresponding underlying structural quality parameter that represents the average mortality rate that a typical patient could expect at this hospital, and V_i is the sampling variance for the estimates in Y_i. Note that the hierarchical nature of the data allow us to estimate V_i in a straightforward manner for each hospital, since this is simply the sampling variance of an estimate derived from a sample of patients at hospital i.

At the second level, the distribution of the structural quality parameter conditional on observed hospital volume is:

(2) $E(\mu_i) = \beta_0 + \beta_1 \ln(Volume_i)$ and $Var(\mu_i) = \sigma^2$,

where $\ln(Volume_i)$ is the log of surgical volume at hospital i, β_1 is the coefficient capturing the effect of surgical volume on patient mortality (expected to be negative), and

σ^2 is the variance of the structural quality parameter summarizing the remaining variation in mortality across hospitals after accounting for differences in surgical volume. Preliminary analysis suggested that the log specification adequately summarized the relationship between patient mortality and volume across all of the surgical procedures.

Calculation of hospital-specific measures. Estimation in our hospital-level analysis proceeds in two stages. First, we construct estimates of the higher-level parameters in equation 2 ($\beta_0, \beta_1, \sigma^2$). Second, we combine information on hospital mortality and volume to construct estimates of the underlying structural quality parameters (μ_i) for each hospital. These estimates are derived from the data (Y_i, Volume$_i$, V_i) observed for a sample of N hospitals, where in our application N is large.

The coefficients ($\hat{\beta}_o, \hat{\beta}_1$) determining the relationship between mortality and hospital volume are estimated using a least squares regression of Y on ln(Volume). To estimate the variance of the structural quality parameters (σ^2), we calculate the variance of the risk-adjusted mortality rates (Y_i), and adjust for sampling variability by subtracting the mean sampling-error variance (V_i). The equation is:

$$(3) \quad \hat{\sigma}^2 = \tfrac{1}{N} \sum_{i=1}^{N} \left\{ \left[Y_i - \hat{\beta}_0 - \hat{\beta}_1 \ln(Volume_i) \right]^2 - V_i \right\} = Var(Y_i) - Mean(V_i)$$

To estimate the underlying structural quality parameter (μ_i) for each hospital, we again use the empirical Bayes approach. The empirical Bayes estimate is a weighted average of the noisy data (Y_i) and the regression predictions ($\hat{\beta}_0 + \hat{\beta}_1 \ln(Volume_i)$), where the weights depend on both the signal and noise variance ($\hat{\sigma}^2$ and V_i). The equation is:

$$(4) \quad \hat{\mu}_i = Y_i W_i + (\hat{\beta}_0 + \hat{\beta}_1 \ln(Volume_i))(1 - W_i),$$

where the weight (W_i) is estimated by

$$(5) \quad W_i = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + V_i}.$$

The weight is simply the ratio of signal variance to total variance (the "reliability") in the mortality rate for hospital i. Equation (4) is a standard empirical Bayes (or shrinkage) estimator that places more weight on a hospital's own mortality rate (Y_i) when the signal ratio is high, but shrinks back toward a (conditional) mean when the signal ratio is low.