Composite Measures For Predicting Surgical Mortality In The Hospital

A simple measure—based on hospital case counts and number of deaths—can explain variations in hospital mortality rates.

by Justin B. Dimick, Douglas O. Staiger, Onur Baser, and John D. Birkmeyer

ABSTRACT: Although payers increasingly report information on hospital volume and mortality from surgery, the value of these data is uncertain. Using national Medicare data for six surgical operations (covering the years 2003–2006), we created a composite measure based on these two quality indicators. We found that this simple measure was a strong predictor of future performance for all six operations. In this regard, it was more effective than the individual measures. Such measures would be useful for helping patients and payers identify low-mortality hospitals for major surgery. [*Health Affairs* 28, no. 4 (2009): 1189–1198; 10.1377/hlthaff.28.4.1189]

Aiming to foster accountability and encourage quality improvement, payers are increasingly collecting and reporting information on the quality of surgery in the hospital.[1] 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 numbers of deaths from seven different procedures. This information is used to categorize hospitals for purposes of public reporting or selective contracting.[2]

However, it remains unclear whether this information is useful for identifying the best hospitals for surgery. Mortality rates are often too "noisy" to accurately reflect hospital quality with surgery.[3] In addition, although very important for some operations, hospital volume is a relatively weak proxy for mortality with most procedures.[4] Furthermore, when multiple measures are considered, it is not clear how to best weight or interpret them when they conflict.[5] For example, some hospitals will have low volumes but low mortality; others will have high volumes

Justin Dimick (jdimick@umich.edu) is an assistant professor of surgery at the University of Michigan in Ann Arbor. Douglas Staiger is a professor of economics at Dartmouth College in Hanover, New Hampshire. Onur Baser is an assistant professor and John Birkmeyer, a professor of surgery, at the University of Michigan.

DOI 10.1377/hlthaff.28.4.1189

but high mortality. It is not clear which group of hospitals should be preferred by patients or payers.

In this paper we describe a simple composite measure of 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 measure for predicting future hospital performance. Because self-reported mortality rates are generally not risk-adjusted, we also evaluated the extent to which risk adjustment is important for predicting hospital outcomes. Because this composite measure has been adopted by the Leapfrog Group for its 2008 Safety Standards, this study evaluates the extent to which patients and payers can use these measures to identify hospitals with low mortality rates.

Study Data And Methods

■ **Data source and study population.** We used data from the Medicare Provider Analysis and Review (MedPAR) files for 2003–2006. These files, which contain hospital discharge abstracts for all fee-for-service acute care hospitalizations of all U.S. Medicare recipients, were used to create our main analysis data sets. The Medicare denominator file was used to assess patient vital status at thirty 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*, Ninth Revision (ICD-9), we identified all patients ages 65–99 undergoing six operations included in the Leapfrog Group's evidence-based hospitals referral initiative: coronary artery bypass grafting (CABG), aortic valve replacement, abdominal aortic aneurysm (AAA) repair, percutaneous coronary interventions (PCI), and resection of pancreatic and esophageal cancer.[6] We did not include the seventh operation, bariatric surgery, because most patients undergoing this procedure are not covered by Medicare. In keeping with Leapfrog data specifications, we excluded small patient subgroups with much higher baseline risks than those of the patients we included in the study, including those with procedure codes indicating that other operations were simultaneously performed (for example, CABG and valve surgery) or were performed for emergent indications (for example, ruptured aortic aneurysms).

■ **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 (for example, mortality rate observed at a hospital) is adjusted for reliability by shrinking it toward the overall mean (for example, the average mortality rate).[7] 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." With this approach, the observed mortality rate is weighted according to how reliably it is estimated, with the remaining weight placed on hospital volume. Because this method includes observed data to the extent that they are reliable and uses the proxy measure only to the extent necessary, it ensures an optimal combination of these two quality domains.

The two inputs to the composite measure are observed mortality rates and hospital volume for each of the six operations. In creating our composite measure, we assumed that we had only two pieces of information from each hospital: the number of cases and the number of deaths. Procedure-specific mortality rates were then calculated for all hospitals by dividing the number of deaths by the number of cases at each hospital over a two-year period (2003–04). Hospital volume was calculated as the number of Medicare-paid procedures performed during the same two years. For each operation, the relationship between hospital volume and risk-adjusted mortality was modeled using linear regression. After we tested the fit of several transformations, hospital volume was modeled as the natural log of the continuous volume variable—the same approach used in our previous work.[8] Using this regression model, we estimated the volume-predicted mortality for each hospital. We refer to the mortality residual from this regression as the between-hospital variation, which represents the variation across hospitals that is not explained by hospital volume.

We then used an empirical Bayes approach in combining these two inputs. Our composite measure theoretically provides the best estimate of a hospital's true mortality rate, taking into account both inputs.[9] The combined measure was calculated as the weight multiplied by the observed mortality plus 1-weight multiplied by the volume-predicted mortality. The weight placed on mortality is equal to its reliability, or the ratio of between-hospital variation to total variation. Total variation was estimated as the between-hospital variation plus the within-hospital variation. Between-hospital variation was estimated from regression, as described above. Within-hospital variation was calculated as the standard error (noise variance) of the mortality rate at each hospital, which is largely a function of sample size. We estimated these parameters using standard methods.[10]

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. For example, consider a hospital that performs eighty AAA repairs over a two-year period, with eight deaths. The observed mortality rate would be 10 percent, and, based on the estimates created for this paper, the weight placed on the mortality component would be 0.15. The volume-predicted mortality for this hospital would be approximately 4.0 percent, and the weight placed on this input is 0.85. The composite mortality prediction is then calculated as a weighted combination of these two inputs, as follows: (10%)(0.15) + (4.0%)(0.85) = 4.9%. This predicted mortality (4.9 percent) is likely a better estimate of the hospital's "true" mortality than either input alone.

■ **Validating the composite measure.** We determined the value of our composite measure by (1) establishing the extent to which it explained hospital-level variation in contemporaneous risk-adjusted mortality rates, and (2) determining how well it predicted future hospital performance. We first estimated the proportion of variation in hospital-level risk-adjusted mortality (2003–04) 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 effects, after adding each measure to the model.[11] We next compared the composite measure to each individual measure, mortality rates and hospital volume. To account for differences in disease severity, patient characteristics were also included in all models. We included patients' age, sex, race, urgency of operation, median income for the patient's ZIP code, and coexisting diseases, as described in our previous work.[12] Coexisting diseases were determined from secondary diagnostic codes using the methods of Anne Elixhauser and colleagues.[13]

We next determined the extent to which the composite measure from 2003–04 predicts future risk-adjusted mortality (2005–06). Risk adjustment was performed using logistic regression to estimate expected mortality rates for each hospital based on the patient characteristics described above. The observed mortality rate at each hospital was then divided by the expected mortality rate to yield the ratio of observed to expected deaths (O/E ratio). The O/E ratio was next multiplied by the average mortality rate for each operation to yield a risk-adjusted mortality rate.

Hospitals were then ranked based on each measure from the earlier time period (data from 2003–04) and divided into four quartiles (each with an equal number of patients). We chose quartiles for clarity of presentation; this composite measure could be used as a continuous variable or with any other number of hospital groups. The subsequent risk-adjusted mortality rates for each quartile of performance were then calculated (data from 2005 and 2006). All statistical analyses were conducted using STATA 10.0.

■ **Sensitivity analysis.** We performed a sensitivity analysis to evaluate the importance of risk-adjusting the "input" mortality rate. Specifically, we compared the ability of risk-adjusted and unadjusted composite measures to predict subsequent risk-adjusted mortality.

Study Results

■ **Variation in weights applied.** The average weight applied to each input to the composite measure (mortality and hospital volume) varied by procedure and hospital (Exhibit 1). For CABG and PCI, the two procedures with the highest hospital caseloads, more weight, on average, was placed on the hospital's observed mortality (51 percent and 55 percent, respectively). Conversely, for esophageal resection, the operation with by far the lowest caseload, much less weight, on average, was placed

EXHIBIT 1
Hospital Caseload And The Weight Applied To Each Of The Two Inputs In The Composite Measure

Procedure	Hospital caseloads, mean (SD)	Weight applied to observed mortality, mean (SD)	Weight applied to hospital volume, mean (SD)
Coronary artery bypass grafting	207 (195)	0.51 (0.18)	0.49 (0.18)
Aortic valve replacement	55 (69)	0.33 (0.18)	0.67 (0.18)
Percutaneous coronary interventions	409 (473)	0.55 (0.18)	0.45 (0.18)
Elective abdominal aortic aneurysm repair	26 (38)	0.15 (0.11)	0.85 (0.11)
Pancreatic cancer resection	6 (13)	0.22 (0.20)	0.78 (0.20)
Esophageal cancer resection	4 (8)	0.16 (0.16)	0.84 (0.16)

SOURCE: Medicare Provider Analysis and Review (MedPAR) files (2003–04).
NOTES: The weight applied to the mortality rate is the reliability and the weight applied to hospital volume is 1-reliability. SD is standard deviation.

on the hospital's observed mortality (16 percent).

■ **Amount of variation explained by the composite measure.** The composite measure explained a large proportion of nonrandom hospital-level variation in risk-adjusted mortality rates (Exhibit 2). The amount of variation explained by the composite measure varied from 45 percent for aortic valve replacement to 71 percent for PCI. Although the composite measure explained more variation than either measure alone for all six operations, the ability of individual measures to explain variation was different for each procedure (Exhibit 2). For the more common operations, such as CABG, mortality rates explained a large proportion of the variation (44 percent),

EXHIBIT 2
Relative Ability Of Each Measure To Explain Hospital-Level Differences In Risk-Adjusted Mortality Rates, 2003–2004

	Proportion of hospital-level variation in mortality rates explained by each performance measure (%)		
	Hospital volume	Observed mortality	Simple composite measure
Coronary artery bypass grafting	7	44	55
Aortic valve replacement	22	30	45
Percutaneous coronary interventions	37	53	71
Elective abdominal aortic aneurysm repair	58	14	64
Pancreatic cancer resection	69	19	76
Esophageal cancer resection	65	14	70

SOURCE: Medicare Provider Analysis and Review (MedPAR) files (2003–04).

and hospital volume explained a small proportion (7 percent). For less common operations, such as pancreatic resection, hospital volume explained a much larger proportion of the variation (69 percent) than mortality rates (19 percent).

■ **Predictive value of the measure.** The composite measure predicted large differences in future risk-adjusted mortality across hospitals (Exhibit 3). The best prediction was achieved with pancreatic resection, with greater than fourfold differences between the "worst" and "best" quartiles (odds ratio [OR], 4.23; 95 percent confidence interval [CI], 3.07–5.84). The composite measure was least predictive for CABG, but future mortality rates were still 1.7 times higher in "worst" compared to the "best" quartile (OR, 1.72; 95 percent CI, 1.57–1.88). When compared to the individual measures by themselves, the composite was better at predicting differences between the "best" and "worst" quartiles for five of the six procedures (Exhibit 3). For esophagectomy, hospital volume and the composite performed the same. In addition to providing good discrimination between the extremes of performance, the simple composite measure predicts differences in mortality for the intermediate strata for all six procedures, including esophagectomy (Exhibit 4 and online Appendix 2).[14] The composite measure is much better at sorting hospitals into quartiles with stepwise, monotonic differences in mortality. In other words, there is a consistent change in the risk of mortality associated with each quartile. In sensitivity analysis, composite measures based on an unadjusted mortality input and a risk-adjusted mortality input were nearly identical and equally good at predicting future performance.[15]

Discussion

Although information on hospital quality is increasingly collected and reported, the usefulness of much of these data is uncertain. In this paper we have as-

EXHIBIT 3
Relative Ability Of Historical Measures (2003–04) To Predict Subsequent Risk-Adjusted Mortality (2005–06)

	Adjusted odds ratio for risk-adjusted mortality, "worst" versus "best" quartile (95% CI)		
Procedure	**Hospital volume alone (2003–04)**	**Observed mortality alone (2003–04)**	**Simple composite measure (2003–04)**
Coronary artery bypass grafting	1.15 (1.04–1.27)	1.71 (1.57–1.87)	1.72 (1.57–1.88)
Aortic valve replacement	1.48 (1.29–1.67)	1.73 (1.54–1.94)	1.97 (1.75–2.23)
Percutaneous coronary interventions	1.49 (1.35–1.64)	1.84 (1.69–2.01)	1.95 (1.79–2.13)
Abdominal aortic aneurysm repair	1.48 (1.26–1.73)	1.20 (1.05–1.38)	1.69 (1.44–1.98)
Pancreatic cancer resection	3.93 (2.85–5.42)	1.36 (1.05–1.76)	4.23 (3.07–5.84)
Esophageal cancer resection	2.89 (2.15–3.89)	.90 (.72–1.12)	2.71 (2.03–3.60)

SOURCE: Medicare Provider Analysis and Review (MedPAR) files (2003–06).
NOTE: CI is confidence interval.

EXHIBIT 4
Future Risk-Adjusted Mortality Rates (2005–06) For Quartiles Of Hospital Rankings Based On Historical (2003–04) Hospital Volume, Risk-Adjusted Mortality Rates, And Composite Measures

	Risk-adjusted mortality in 2005–06 for each quartile of historical hospital rankings (2003–04)			
Procedure/measure	**1**	**2**	**3**	**4**
Coronary artery bypass grafting				
Hospital volume	3.3	3.5	3.6	3.8
Observed mortality	2.7	3.2	3.8	4.6
Composite	2.6	3.2	3.9	4.5
Aortic valve replacement				
Hospital volume	4.9	5.5	6.3	8.2
Observed mortality	5.1	6.1	6.3	7.4
Composite	4.3	5.7	6.6	8.2
Percutaneous coronary interventions				
Hospital volume	1.7	2.2	2.5	3.1
Observed mortality	2.0	2.3	2.3	2.9
Composite	1.6	2.2	2.5	3.1
Abdominal aortic aneurysm repair				
Hospital volume	2.7	3.2	3.4	4.0
Observed mortality	3.4	2.9	3.0	4.1
Composite	2.6	3.1	3.3	4.4
Pancreatic cancer resection				
Hospital volume	3.0	4.4	7.9	11.0
Observed mortality	7.1	2.2	4.5	9.5
Composite	2.7	4.2	8.4	10.5
Esophageal cancer resection				
Hospital volume	4.9	10.3	10.0	13.1
Observed mortality	11.5	5.2	8.3	10.5
Composite	5.3	8.6	11.1	13.2

SOURCE: Medicare Provider Analysis and Review (MedPAR) files (2003–06).

sessed the value of a simple composite measure—based only on hospital case counts and the number of deaths—for predicting future hospital performance. We found that this composite measure explained a large proportion of hospital-level variation in mortality rates and was good at predicting future hospital performance. In this regard, this simple composite measure performed better than individual measures for all six operations. We validated our composite by studying how well it predicts future performance—arguably the most important criterion for assessing the value of measures used for public reporting or selective referral. Patients and payers are interested in how a hospital performs now or in the future, not how the hospital performed over the past several years.

■ **Popularity of composite measures.** Composite measures of performance are gaining increasing popularity in surgery. For example, the Premier/Centers for

Medicare and Medicaid Services (CMS) Hospital Quality Incentive Demonstration uses a composite of process and outcome to measure quality for CABG.[16] The Society of Thoracic Surgeons (STS) Task Force on Quality Measurement advocates a composite score based on a set of outcome and process measures endorsed by the National Quality Forum.[17] Like our simple composite, these measures are created by combining multiple input measures. However, they are designed with a distinctly different goal in mind. The CMS and STS composite scores are designed to provide a summary score representing multiple domains of quality. In contrast, our measure was designed to use readily available information to optimally predict one domain of quality that is crucially important for many procedures: operative mortality.

■ **Difference in our composite design.** Because our composite was designed for a different purpose, we used 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 or weight measures according to expert opinion. We chose to empirically weight the "noisy" mortality measure based on how reliably it was measured. The remainder of the weight was then placed on the proxy measure, hospital volume. We believe that this approach ensures an optimal combination of the two measures because the direct measure (mortality) is weighted to the extent it is reliable and the proxy measure (hospital volume) is only weighted to the extent necessary.

■ **Insights into weighting of inputs.** Our study yields several important insights into the weighting of inputs of a composite measure. These findings suggest that the weight placed on each measure should be tailored to the procedure. For very common operations, such as CABG, more weight is placed on the mortality rate, largely because it is measured with more precision. At the other end of the spectrum, surgeries such as pancreatic and esophageal cancer resections are not performed often enough to measure mortality precisely, and very little weight should be placed on the mortality rate. The weights applied to input measures should also vary across hospitals performing the same procedure. If a hospital performs a high number of a certain operation, its mortality rate will be measured more precisely, and should receive more weight, than will be the case for a hospital that performs fewer cases of that operation.

■ **Caveats.** *Limited value of risk adjustment.* We found that risk adjustment of the mortality rate did not improve the ability of the composite to predict future risk-adjusted mortality. This could be accounted for by our reliance on Medicare claims data. It is widely known that claims data are limited in their ability to adjust for patient risk.[18] A more complete assessment of patient risk factors would potentially yield different results. But this finding may also be due to a lack of systematic variation in patient severity of illness across hospitals. Patients undergoing the same surgical procedure tend to be a homogenous group. A recently published article using rich clinical data showed that adjusted and unadjusted mortality rates for cardiac surgery were highly correlated ($r > 0.9$) and equally good at predicting future per-

"Our simple composite measure will be better than existing alternatives at helping patients identify safer hospitals for surgery."

formance.[19] Determining which of these possible explanations accounts for our findings would require further analysis with detailed clinical data.

Using Medicare claims data. Another limitation of this study is directly related to using Medicare claims data. Only half of the included surgical procedures are performed on Medicare patients, which reduced our effective sample size at each hospital. Because sample size is the primary determinant of reliability (and the weight placed on mortality), we have likely underestimated the importance of this input. If an all-payer data set was used, and all patients at a hospital were included, there would be a larger sample size, and more weight would be placed on mortality. For this reason, our analyses likely underestimate the extent to which simple composite measures are able to predict future mortality.

Limited predictive ability for individual hospitals. Although our simple composite measure performed well at grouping hospitals according to their future mortality rates, it would no doubt perform less well at predicting mortality rates for individual hospitals. Whether this distinction is important depends on both perspective and purpose. Hospitals want a measure that reliably distinguishes their performance from others and also provides information useful for quality improvement. Patients and payers would value a measure that provided a ranking of one to five stars, which would greatly increase their odds of selecting a hospital with low mortality. Although not ideal from the hospital perspective, our simple composite measure would no doubt be valued by patients and payers.

Although we have demonstrated the value of a simple method for combining two measures, the predictive accuracy of composite measures could surely be improved. Using a broader array of input measures, including outcomes with other, related operations, would improve the ability to predict future mortality.[20] For some procedures, clinical process measures may prove useful as inputs to a composite measure. The widespread use of such composite measures will await broadly available, clinically detailed outcome registries. In the meantime, our simple composite measure will be better than existing alternatives at helping patients and payers identify safer hospitals for surgery.

This paper was presented at the AcademyHealth Annual Research Meeting, 5 June 2008, in Washington, D.C. This study was supported by a grant from the National Institute on Aging (Grant no. R21 AG027819). The views expressed herein do not necessarily represent the views of the Centers for Medicare and Medicaid Services or the U.S. government.

NOTES

1. R.S. Galvin, "The Business Case for Quality," *Health Affairs* 20, no. 6 (2001): 57–58; and M.B. Rosenthal and R.A. Dudley, "Pay-for-Performance: Will the Latest Payment Trend Improve Care?" *Journal of the American Medical Association* 297, no. 7 (2007): 740–744.
2. Leapfrog Group, "Evidence-Based Hospital Referral," Fact Sheet, http://www.leapfroggroup.org/media/file/FactSheet_EBHR.pdf (accessed 15 April 2009).
3. J.B. Dimick, H.G. Welch, and J.D. Birkmeyer, "Surgical Mortality as an Indicator of Hospital Quality: The Problem with Small Sample Size," *Journal of the American Medical Association* 292, no. 7 (2004): 847–851.
4. E.L. Hannan et al., "Do Hospitals and Surgeons with Higher Coronary Artery Bypass Graft Surgery Volumes Still Have Lower Risk-Adjusted Mortality Rates?" *Circulation* 108, no. 7 (2003): 795–801; and E.A. Halm, C. Lee, and M.R. Chassin, "Is Volume Related to Outcome in Health Care? A Systematic Review and Methodologic Critique of the Literature," *Annals of Internal Medicine* 137, no. 6 (2002): 511–520.
5. S.M. O'Brien et al., "Quality Measurement in Adult Cardiac Surgery: Part 2—Statistical Considerations in Composite Measure Scoring and Provider Rating," *Annals of Thoracic Surgery* 83, no. 4 Supp. (2007): S13–S26; and B.E. Landon et al., "Physician Clinical Performance Assessment: Prospects and Barriers," *Journal of the American Medical Association* 290, no. 9 (2003): 1183–1189.
6. Leapfrog Group, "Evidence-Based Hospital Referral."
7. C.N. Morris, "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association* 78, no. 381 (1983): 47–55; and M.B. McClellan and D.O. Staiger, "Comparing the Quality of Health Care Providers," in *Frontiers in Health Policy Research*, vol. 3, ed. A. Garber (Cambridge, Mass.: MIT Press, 2000), 113–136.
8. J.D. Birkmeyer et al., "Surgeon Volume and Operative Mortality in the United States," *New England Journal of Medicine* 349, no. 22 (2003): 2117–2127.
9. Morris "Parametric Empirical Bayes Inference."
10. These standard methods are described in ibid; D.O. Staiger et al., "Empirically Derived Composite Measures of Surgical Performance," *Medical Care* 47, no. 2 (2009): 226–233; and Appendix 1, online at http://content.healthaffairs.org/cgi/content/full/28/4/1189/DC1.
11. J.D Birkmeyer, J.B. Dimick, and D.O. Staiger, "Operative Mortality and Procedure Volume as Predictors of Subsequent Hospital Performance," *Annals of Surgery* 243, no. 3 (2006): 411–417.
12. Birkmeyer et al., "Surgeon Volume and Operative Mortality."
13. A. Elixhauser et al., "Comorbidity Measures for Use with Administrative Data," *Medical Care* 36, no. 1 (1998): 8–27.
14. Appendix 2 is online, as in Note 10.
15. See Appendix 3, ibid.
16. See Centers for Medicare and Medicaid Services, "Premier Hospital Quality Incentive Demonstration," 19 February 2009, http://www.cms.hhs.gov/HospitalQualityInits/35_HospitalPremier.asp (accessed 13 April 2009).
17. D.M. Shahian et al., "Quality Measurement in Adult Cardiac Surgery: Part 1—Conceptual Framework and Measure Selection," *Annals of Thoracic Surgery* 83, no. 4 Supp. (2007): S3–S12.
18. L.I. Iezzoni, "Assessing Quality Using Administrative Data," *Annals of Internal Medicine* 127, no. 8, Part 2 (1997): 666–674.
19. J.B. Dimick and J.D. Birkmeyer, "Ranking Hospitals on Surgical Quality: Does Risk-Adjustment Always Matter?" *Journal of the American College of Surgeons* 207, no. 3 (2008): 347–351.
20. Staiger et al., "Empirically Derived Composite Measures."