HSR Health Services Research

© Health Research and Educational Trust
DOI: 10.1111/j.1475-6773.2010.01158.x
RESEARCH ARTICLE

Ranking Hospitals on Surgical Mortality: The Importance of Reliability Adjustment

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

Objective. We examined the implications of reliability adjustment on hospital mortality with surgery.

Data Source. We used national Medicare data (2003–2006) for three surgical procedures: coronary artery bypass grafting (CABG), abdominal aortic aneurysm (AAA) repair, and pancreatic resection.

Study Design. We conducted an observational study to evaluate the impact of reliability adjustment on hospital mortality rankings. Using hierarchical modeling, we adjusted hospital mortality for reliability using empirical Bayes techniques. We assessed the implication of this adjustment on the apparent variation across hospitals and the ability of historical hospital mortality rates (2003–2004) to forecast future mortality (2005–2006).

Principal Findings. The net effect of reliability adjustment was to greatly diminish apparent variation for all three operations. Reliability adjustment was also particularly important for identifying hospitals with the lowest future mortality. Without reliability adjustment, hospitals in the "best" quintile (2003–2004) with pancreatic resection had a mortality of 7.6 percent in 2005–2006; with reliability adjustment, the "best" hospital quintile had a mortality of 2.7 percent in 2005–2006. For AAA repair, reliability adjustment also improved the ability to identify hospitals with lower future mortality. For CABG, the benefits of reliability adjustment were limited to the lowest volume hospitals.

Conclusion. Reliability adjustment results in more stable estimates of mortality that better forecast future performance. This statistical technique is crucial for helping patients select the best hospitals for specific procedures, particularly uncommon ones, and should be used for public reporting of hospital mortality.

Key Words. Quality, surgery, hospital, mortality, hierarchical

Surgical mortality rates are used widely to measure quality with high-risk surgery. New York, Pennsylvania, California, and a growing number of other states publicly report hospital mortality rates for cardiac surgical procedures (Steinbrook 2006). The Agency for Healthcare Research and Quality (AHRQ) advocates the use of mortality rates as surgical quality indicators for a broad range of noncardiac operations, including abdominal aortic aneurysm (AAA) repair, pancreatectomy, and esophagectomy (Dimick, Welch, and Birkmeyer 2004). To date, much of the concern regarding the accuracy of these report cards has focused on the data source (administrative versus clinical) and the adequacy of risk adjustment (Parker et al. 2006).

However, the issue of reliability adjustment has received much less attention. The need to adjust for statistical reliability is particularly important when sample sizes are small. With small numbers of cases, it is hard to know whether extremely high or low mortality rates are due to chance (i.e., bad or good luck) or true differences in quality (Hofer et al. 1999; Dimick, Welch, and Birkmeyer 2004). Reliability adjustment, an application of hierarchical modeling, allows for quantification and removal of this statistical noise using empirical Bayes techniques (Morris 1983; Normand, Glickman, and Gatsonis 1997). This approach has been increasingly applied to health care, including in the public reporting of hospital outcomes for medical conditions on the Center for Medicare and Medicaid Services (CMS) HospitalCompare website (CMS 2010).

We assessed whether reliability adjustment might be similarly important in assessing surgical mortality. Using national Medicare data, we studied three operations commonly targeted by public reporting initiatives and/or pay for performance programs. We first evaluated the degree to which reliability adjustment reduces apparent variation in hospital mortality rates. Given how publicly reported information on quality is likely to be used by patients and payers, we then determined whether reliability adjustment improved the ability of historical hospital mortality rates to forecast future performance.

METHODS

Study Population

We used 100 percent national analytic files from the CMS for the calendar years 2003 through 2006. Medicare Provider Analysis and Review (MEDPAR) files,

Address correspondence to Justin B. Dimick, M.D., M.P.H., Assistant Professor of Surgery, Department of Surgery, University of Michigan, M-SCORE offices, 211 N. Fourth Avenue, Suite 301, Ann Arbor, MI 48104; e-mail: jdimick@umich.edu. Douglas O. Staiger, Ph.D., Professor of Economics, is with the Department of Economics and the Dartmouth Institute for Health Policy and Clinical Practice, Dartmouth College, Hanover, NH. John D. Birkmeyer, M.D., Professor of Surgery and Director of Michigan Surgical Collaborative for Outcomes Research and Evaluation (M-SCORE), is with the Department of Surgery, University of Michigan, Ann Arbor, MI 48104.

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 analytical datasets. The Medicare denominator file was used to assess patient vital status at 30 days.

Using appropriate procedure codes from the International Classification of Diseases, version 9 (ICD-9 codes), we identified all patients aged 65–99 undergoing coronary artery bypass grafting (CABG), elective AAA repair, and pancreatectomy. We chose these operations because they are often targeted by public reporting initiatives and/or pay for performance programs. To minimize the potential for case mix differences between hospitals, we excluded small patient subgroups with much higher baseline risks, including those with procedure codes indicating that other operations were simultaneously performed (e.g., coronary artery bypass and valve surgery) or were performed for emergent indications (e.g., ruptured aortic aneurysms).

Hospital Mortality Rates

We first estimated risk-adjusted hospital mortality rates with all three procedures during 2003–2004. We defined mortality as death within 30 days of operation or before hospital discharge. We use this definition because the 30-day cut-off is somewhat arbitrary, and when a death occurs in the hospital after major elective surgery, it is almost certainly attributable to the operation itself or complications from the surgery. We adjusted for patient age, gender, race, urgency of operation, median ZIP-code income, and coexisting medical conditions. Coexisting medical conditions were obtained from secondary diagnoses in the claims data using the methods of Elixhauser (Southern, Quan, and Ghali 2004). Using logistic regression, we estimated the expected number of deaths in each hospital and then divided the observed deaths by this expected number of deaths to obtain the ratio of observed to expected mortality (O/E ratio). We then multiplied the O/E ratio by the average mortality rate to obtain a risk-adjusted mortality rate for each hospital.

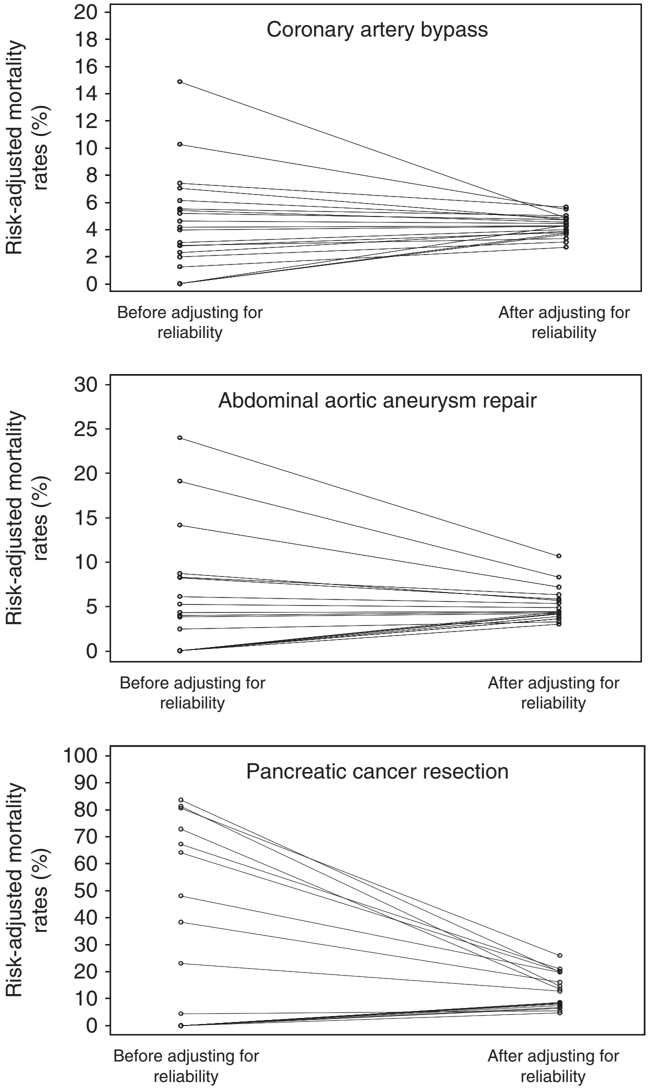
We next used hierarchical modeling techniques to adjust these mortality estimates for reliability (see Appendix SA2 for details). Using random effects logistic regression models, we generated empirical Bayes predictions of mortality for each hospital (Morris 1983; Normand, Glickman, and Gatsonis 1997). This technique shrinks the point estimate of mortality back toward the average mortality rate, with the amount of shrinkage proportional to the reliability at each hospital. Reliability is a measure of precision and is a function of both hospital sample size (which determines "noise" variation) and the amount of true variation across hospitals ("signal"). For example, for hospitals with low caseloads of a particular procedure, mortality rates have lower reliability and are shrunk more toward the average mortality. For hospitals with high caseloads, mortality rates are more reliable and shrunk less toward the average mortality. The resulting reliability-adjusted mortality is considered the best estimate of a hospital's "true" mortality rate with each operation (Morris 1983).

An underlying assumption of reliability adjustment is that hospitals provide average performance until the data are sufficiently robust to prove otherwise. For example, consider a hospital performing 10 pancreatic resections in a year with two deaths (observed mortality rate of 20 percent). Because of the small number of cases, there is considerable likelihood that this estimate of 20 percent is the result of chance and not truly an indication of bad performance. From the empirical Bayes perspective, the true mortality rate lies somewhere between this observed rate of 20 percent and the population-based rate of 5 percent (the average mortality rate across all hospitals). Using reliability adjustment, the observed rate of 20 percent is "shrunk" back toward the average rate of 5 percent. The degree of shrinkage is proportional to the reliability with which the mortality rate is measured. The more reliable the observed mortality rate, the more weight it is afforded. Reliability is assessed on a scale of 0–1, with 1 representing perfect reliability. In this case, suppose the reliability based on 20 cases is 0.15, and the remaining weight (0.85) is placed on the average mortality. Thus, the reliability-adjusted mortality for this hospital is $(0.20)(0.15)+(0.05)(1-0.15) = 7.2$ percent. To further illustrate the impact of reliability adjustment, Figure 1 shows mortality rates before and after reliability adjustment for 20 randomly selected hospitals for each of the three procedures in this study. After reliability adjustment, there is a much less variation across hospitals, because the most extreme observations are shrunk back toward the average mortality rate.

Analysis

We assessed the importance of reliability adjustment from two perspectives. We first examined the extent to which reliability adjustment diminishes apparent variation in mortality across hospitals. We ranked hospitals on risk-adjusted mortality, before and after reliability adjustment, and assessed variation in mortality across hospital quintiles. We performed this analysis using quintiles to make the visual impact of reliability adjustment easier to visualize. We next evaluated the impact of reliability adjustment on the ability of mortality

Figure 1: Variation in Hospital Mortality Rates before and after Adjusting for Reliability. (Twenty Randomly Sampled Hospitals Are Shown for Each Procedure)



rates to forecast future hospital performance. Specifically, we evaluated whether reliability adjustment improves the ability of risk-adjusted mortality rates from an earlier time period (2003–2004) to predict risk-adjusted mortality in a subsequent time period (2005–2006). For this analysis, we calculated risk-adjusted mortality during 2003–2004 and then adjusted each hospital's mortality for reliability, as described above. After merging this data on 2003–2004 hospital performance to the patient level file from 2005–2006, we created quintiles based on mortality rankings, with and without reliability adjustment. We then used logistic regression with death in 2005–2006 as the dependent variable and the quintiles as independent variables. To adjust for patient factors in this analysis, we included all patient covariates, as described above in the section on risk adjustment.

RESULTS

The reliability of mortality rates was highest (mean, 0.44) for CABG, the operation with the highest hospital caseloads. The reliability was lowest (mean, 0.25) for pancreatic resection, the operation with the lowest hospital caseloads (Table 1). The reliability also varied across hospitals for each operation, because hospitals with higher caseloads have much higher reliability (Table 2). For example, with coronary artery bypass surgery, a hospital with only 19 cases had a reliability of 0.08 (5th percentile) compared with a reliability of 0.75 at a hospital with 781 cases (95th percentile) (Table 2).

Hospital mortality rates were widely variable for all three operations. Without reliability adjustment, hospital mortality rates varied from 1.2 percent in the "best" quintile to 9.6 percent in the "worst" quintile for CABG, from 0 percent to 26 percent ("best" to "worst") for AAA repair, and from 0 percent to 57 percent ("best" to "worst") for pancreatic cancer resection (Figure 2). After reliability adjustment, however, variation across hospitals was greatly diminished. For example, with CABG, mortality only ranged from 3.2 percent

Table 1: Total Number of Patients, Average Hospital Caseloads, and the Reliability of Hospital Mortality Rates for Each of Three Operations

Operation	*Medicare Cases*	*Hospital Caseloads, Mean (Standard Deviation)*	*Reliability of Mortality Rates, Mean (Standard Deviation)*
Coronary artery bypass	303,132	267 (256)	0.44 (0.20)
Abdominal aortic aneurysm repair	70,863	33 (49)	0.30 (0.23)
Pancreatic resection	6,192	6 (13)	0.25 (0.20)

Note. National Medicare population, 2003–2004.

Table 2: Reliability and Hospital Caseloads for Three Operations

	Reliability (Hospital Caseloads)		
	Coronary Artery Bypass	*Abdominal Aortic Aneurysm Repair*	*Pancreatic Resection*
Minimum	0.00 (1)	0.00 (1)	0.00 (1)
5th	0.08 (19)	0.00 (1)	0.00 (1)
10th	0.16 (49)	0.04 (2)	0.00 (1)
25th	0.29 (102)	0.08 (5)	0.09 (2)
50th	0.45 (198)	0.25 (15)	0.18 (3)
75th	0.58 (338)	0.48 (41)	0.34 (5)
90th	0.70 (574	0.65 (83)	0.54 (12)
95th	0.75 (781)	0.73 (123)	0.70 (23)
Maximum	0.88 (2053)	0.93 (639)	0.96 (256)

Note. National Medicare population, 2003–2004.

to 6.0 percent from the "best" to "worst" hospital quintile. For this operation, the variation across quintiles decreased from an eightfold difference (1.2–9.6 percent) before reliability adjustment to a twofold difference after reliability adjustment. For pancreatic resection and AAA repair, the differences across quintiles of mortality before after reliability adjustment were even greater (Figure 2).

In assessing the ability of hospital mortality rankings to predict future performance, reliability adjustment was particularly important for the two uncommon operations (Figure 3). For these two operations, pancreatic resection and AAA repair, hospital rankings based on reliability-adjusted mortality were superior at identifying hospitals likely to have the lowest future mortality. Without reliability adjustment, hospitals in the "best" quintile (2003–2004) with pancreatic resection had a mortality of 7.6 percent in 2005–2006; with reliability adjustment, the "best" hospital quintile had a mortality of 2.7 percent in 2005–2006. Similarly, without reliability adjustment, hospitals in the "best" quintile (2003–2004) with AAA repair had a mortality of 4.0 percent in 2005–2006; with reliability adjustment, the "best" hospital quintile had a mortality of 3.2 percent in 2005–2006. In contrast, with CABG, reliability adjustment did not result in a significant improvement in the ability to identify hospitals with the lowest future mortality (Figure 3). However, reliability adjustment did have benefits in predicting future mortality for CABG in the lowest volume hospitals (data not shown). Reliability adjustment did not enhance the ability to discriminate among hospitals at the other end of the spectrum, those with the "worst" performance, as demonstrated by equivalent

Figure 2: Variation in Mortality Rates across Hospital Quintiles, before and after Adjusting for Reliability

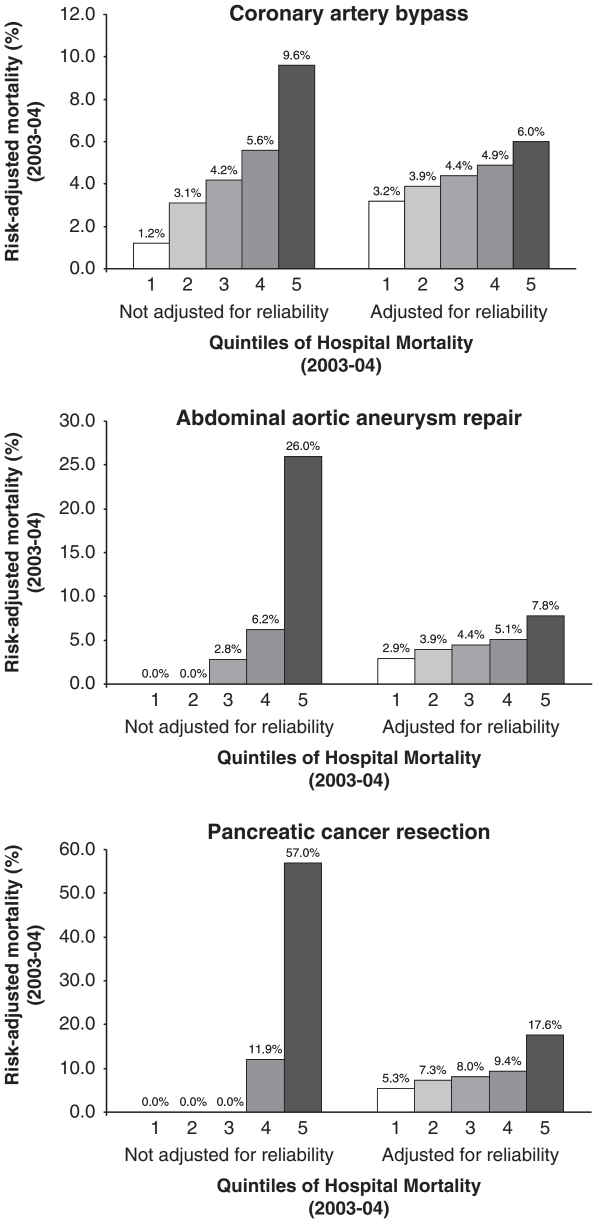


Figure 3: Ability of Mortality Rankings from an Earlier Time Period (2003–2004) to Forecast Future Risk-Adjusted Mortality Rates (2005–2006)

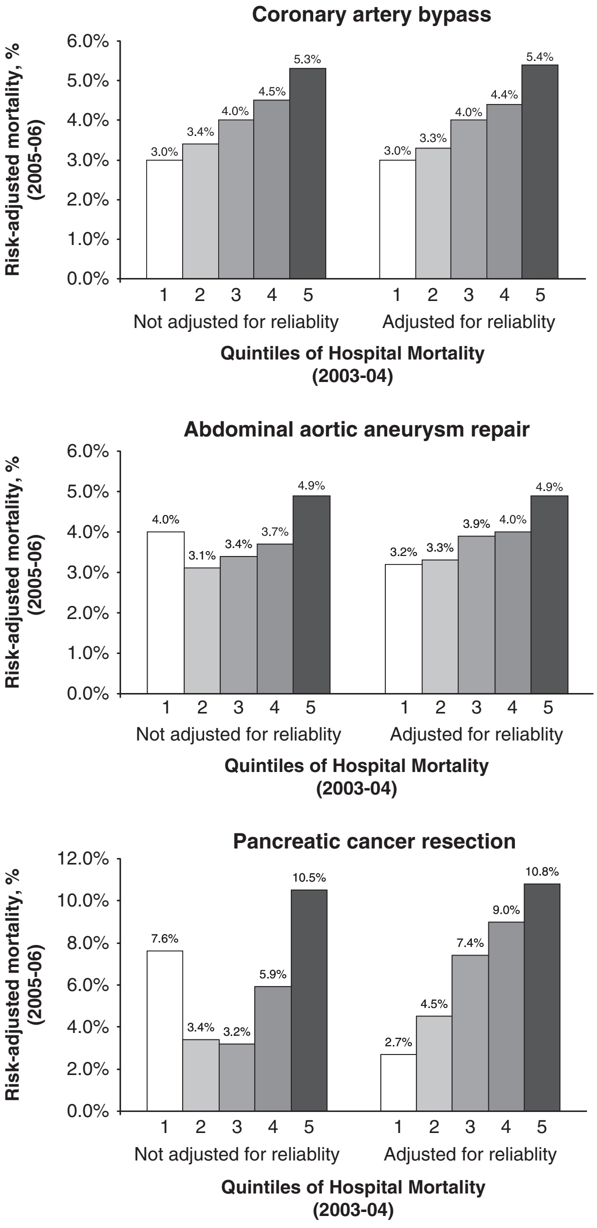


Table 3: Forecasting of Future Mortality (2005–2006) Based on Historical Mortality Rates of (2003–2004), before and after Adjusting for Reliability, Expressed as the Odds Ratio of Mortality in the "Worst" versus the "Best" Hospital Quintile

	Odds Ratio for Mortality in 2005–2006, "Worst" versus "Best" Hospital Quintile (95% CI)	
Operation	*Hospitals Ranked on 2003–2004 Mortality, before Reliability Adjustment*	*Hospitals Ranked on 2003–2004 Mortality, after Reliability Adjustment*
Coronary artery bypass	1.83 (1.71–1.98)	1.87 (1.74–2.02)
Abdominal aortic aneurysm repair	1.27 (1.11–1.45)	1.58 (1.36–1.82)
Pancreatic resection	1.42 (1.11–1.82)	4.08 (2.73–6.09)

future mortality in this quintile (Figure 3). The difference between the "best" and "worst" hospitals can be explained by examining the movement of hospitals into different quintiles after reliability adjustment. For all three operations, the hospitals with no deaths were originally in the "best" quintile. However, many of these were very small hospitals, and they moved to the middle quintiles after reliability adjustment. For example, with pancreatic resection, there were 428 hospitals in the "best" quintile before reliability adjustment and 403 (93 percent) of these had no deaths. After reliability adjustment, all but 16 of these "zero mortality" hospitals moved to the middle quintiles. However, in the "worst" quintile, there was much less movement. Of the 208 hospitals in the "worst" quintile, only six (3 percent) moved toward the middle.

Besides better prediction of future mortality in the "best" hospitals, reliability adjustment improves the ability of hospital mortality rankings to discriminate across the entire spectrum of performance (Figure 3). For pancreatic resection, the odds ratio from the "worst" to "best" quintile went from 1.42 (95 percent CI: 1.11–1.82) before to 4.08 (95 percent CI: 2.73–6.09) after reliability adjustment (Table 3). For AAA repair, the odds ratio of "worst" to "best" quintile went from 1.27 (95 percent CI: 1.11–1.45) before to 1.58 (95 percent CI: 1.36–1.82) after reliability adjustment (Table 3). Although reliability adjustment was good at predicting the extreme quartiles ("best" versus "worst"), it also improved the ability of historical mortality to predict intermediate levels of performance, as represented by the three hospital quintiles in the middle. Reliability adjustment groups hospitals into discrete categories that provide a consistent, monotonic relationship between historical and future mortality (Figure 3). Reliability adjustment had very little impact on the predictive ability of CABG mortality when including all hospitals (Figure 3), because the benefits were limited to the lowest volume hospitals (data not shown).

DISCUSSION

In the present study, we demonstrated the benefits of adjusting surgical mortality rates for statistical reliability. For all three procedures, reliability adjustment greatly reduced apparent variation by eliminating statistical noise. However, when assessing the ability to forecast future performance, the impact of reliability adjustment varied across procedures, with the greatest benefit for the two operations with lower hospital caseloads, pancreatic resection and AAA repair. For these two procedures, mortality rankings based on reliability-adjusted mortality were superior at identifying the "best" hospitals (i.e., those likely to have the lowest mortality in future years). Because most surgical procedures are similar in frequency to pancreatic resection and AAA repair, reliability adjustment would likely improve the accuracy of hospital mortality reporting for most operations.

The importance of reliability adjustment is becoming increasingly recognized. Hofer et al. (1999) popularized reliability adjustment by elucidating its benefits in profiling physician quality in diabetes care. In another application, Zaslavsky et al. (2000) described the benefits of applying reliability adjustment to patient satisfaction reports in the Consumer Assessment of Health Plans Survey (CAHPS). Glance et al. (2006) demonstrated the impact of hierarchical modeling on hospital and surgeon mortality rankings for cardiac surgery in New York State. Consistent with this prior work, and other empirical studies, our study demonstrates the theoretical value of reliability adjustment. However, our present study also adds to this body of work by demonstrating that reliability-adjusted hospital rankings are better at forecasting future performance.

As a result of these and other studies, several organizations are advocating the use of these techniques for quality measurement. The National Quality Forum (2005–2006) expressed a preference for "hierarchical" modeling in their consensus standards for quality monitoring. Although they do not specifically state a preference for the use of empirical Bayes analysis, this is implied in their statement that this technique is especially good at evaluating quality in small hospitals. The AHRQ (2008) also encourages the use of similar methods for use with their *Inpatient Quality Indicators.* The AHRQ website includes downloadable software for creating so-called smoothed estimates of outcomes, which are created using methods similar to those presented in this study.

Reliability adjustment is already seeing real-world application in several quality reporting programs. Perhaps the most visible use of reliability adjustment is the CMS HospitalCompare website (2010), which reports reliability- and risk-adjusted mortality and readmission rates for medical conditions, such as acute myocardial infarction, heart failure, and pneumonia. Massachusetts, which publishes each hospital's mortality rate as part of an annual cardiac surgery report card, recently began emphasizing reliability adjustment (Shahian et al. 2005). The Society of Thoracic Surgeons (STS), which maintains the largest clinical registry for cardiac surgery, adjusts process and outcome measures for reliability before combining them into a composite quality measure (O'Brien et al. 2007). However, as we demonstrate in the present study, this technique is even more valuable for less common operations, and it would likely have even greater value if applied outside cardiac surgery.

In the context of public reporting in surgery, reliability adjustment offers clear advantages over traditional approaches. First, as demonstrated in this paper, the excess variation from statistical noise is greatly diminished. Traditionally, imprecision in measuring mortality is addressed by using confidence intervals, or by comparison to a benchmark, followed by testing for statistically significant differences. Unfortunately, confidence intervals are often misinterpreted and *p*-values are usually relegated to a footnote. In contrast, reliability adjustment directly addresses this problem by estimating the hospital's "true" mortality. Second, reliability adjustment improves the ability of historical mortality to forecast future performance. The ability to forecast future performance is particularly important for public reporting and value-based purchasing, because decisions made by patients and payers about where to have surgery are based on data from several years ago (Birkmeyer, Dimick, and Staiger 2006).

Reliability adjustment also has potential disadvantages. In empirical Bayes methods, the mortality is "shrunk" back toward the average mortality, with the degree of shrinkage related to the reliability, or precision, with which mortality is measured (Morris 1983). Hospitals with low caseloads and unreliable mortality rates are shrunk more toward the average. This technique has very different implications for hospitals that start on each side of the average. For hospitals below average, their mortality may be higher after reliability adjustment. Even hospitals with no deaths will have reliability-adjusted mortality rates greater than zero. Although this seems unfair at first, there is empirical data suggesting that low-volume hospitals with no deaths are actually no better and perhaps worse than average, the so-called zero mortality paradox (Dimick and Welch 2008).

For hospitals above average, reliability adjustment tends to reduce apparent mortality. Because reliability is a function of sample size, small hospitals will be "shrunk" more toward the average than larger hospitals. Thus, this technique gives smaller hospitals the benefit of the doubt. Many critics of reliability adjustment correctly point out that this introduces bias, because small hospitals may not truly have average performance. This criticism is backed by a large body of evidence showing a relationship between higher volume and worse outcome in health care, especially for high-risk surgical procedures (Birkmeyer et al. 2002). However, there is an alternative approach for reliability adjustment that overcomes this potential bias. This alternative approach uses empirical Bayes techniques, but rather than shrinking back toward the average, the hospital's mortality is shrunk back toward the mortality expected given the hospital's volume (Dimick et al. 2009). By taking into account the well-known relationship between lower volume and higher mortality, this approach avoids the bias introduced by assuming small hospitals have average performance. The Leapfrog Group, a large group of health care purchasers, has embraced this approach and will feature these measures in the next iteration of their evidence-based hospital referral initiative for high-risk surgery (The Leapfrog Group 2010).

In this study, we found that reliability adjustment appeared to improve the ability to identify the "best" but not the "worst" hospitals. We also found much greater movement of hospitals out of the "best" quintile (and toward the middle) with reliability adjustment compared with the "worst" quintile. This interesting finding is explained by the fact that the mortality rates for these surgical procedures are closer to "0 percent" than "100 percent." Thus, there is clustering of a large number of small hospitals with no deaths in the "best" quintile (zero mortality hospitals). However, there are only a few "100 percent" mortality hospitals in the "worst" quintile. We would only expect there to be equal movement toward the middle from both sides ("best" and "worst" quintiles) if the average mortality was 50 percent. Nonetheless, we believe that reliability adjustment is still important for both tails of the distribution. The reliability-adjusted mortality for these hospitals is much lower, despite these hospitals still being ranked in the "worst" quintile. Thus, reliability adjustment is important so hospitals do not overestimate their "true" mortality. Further, reliability adjustment results in a net movement of a few small hospitals that should not be included in the "worst" quintile.

This study has several limitations. Because the Medicare population does not account for all patients undergoing these surgical procedures, we likely underestimate the sample size at each hospital. Thus, our analysis may represent an overestimate of the importance of reliability adjustment. However, the impact of this limitation is likely small, because Medicare represents a large proportion of people undergoing each of these three procedures. Further, many existing quality reporting programs create reports based only on Medicare data. Another limitation of this study is the exclusive focus on mortality. Although this outcome is perhaps the most common quality metric in hospital report cards for high-risk procedures, other quality measures such as morbidity, process of care, and patient satisfaction may be important with other procedures. However, the problem of statistical noise is not unique to mortality and is shared by all quality measures. Thus, the accuracy of other measures could be similarly improved using reliability adjustment.

Numerous stakeholders would benefit from better surgical quality measures. Publicly reported surgical outcomes should be adjusted for reliability to help patients choose the best hospitals, thereby improving their odds of surviving surgery. Quality measures used for value-based purchasing should be reliability adjusted to ensure payers and purchasers are steering patients toward hospitals that truly have superior performance. Finally, outcomes data fed back to hospitals should be adjusted for reliability to optimize the impact of provider-led quality improvement registries. Without reliability adjustment, hospitals may waste resources by responding to spuriously high mortality rates, or be lulled into a false sense of security by spuriously low mortality rates. Reliability adjustment is ready for immediate application and should become standard for reporting mortality and other outcomes.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This study was supported by a career development award to Dr. Dimick from the Agency for Healthcare Research and Quality (K08 HS017765) and a research grant to Dr. Dimick from the National Institute of Diabetes and Digestive and Kidney Diseases (R21DK084397). The views expressed herein do not necessarily represent the views of Center for Medicare and Medicaid Services or the United States Government.

Prior dissemination: This work was presented in part at the Academy Health Annual Research Meeting in Chicago, IL, June 29, 2009.

Disclosures: None.
Disclaimers: None.

References

Agency for Healthcare Research and Quality (AHRQ). 2008 "Inpatient Quality Indicators (IQI) Composite Measure Workgroup Final Report, March 2008" [accessed on March 5, 2010]. Available at http://www.qualityindicators.ahrq.gov/downloads/iqi/AHRQ_IQI_Workgroup_Final.pdf

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.

Center for Medicare and Medicaid Services (CMS). 2010 "Hospital Compare Website" [accessed on March 5, 2010]. Available at http://www.hospitalcompare.hhs.gov

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

Dimick, J. B., and H. G. Welch. 2008. "The Zero Mortality Paradox in Surgery." *Journal of the American College of Surgeons* 206 (1): 13–6.

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.

Glance, L. G., A. Dick, T. M. Osler, Y. Li, and D. B. Mukamel. 2006. "Impact of Changing the Statistical Methodology on Hospital and Surgeon Ranking: The Case of the New York State Cardiac Surgery Report Card." *Medical Care* 44 (4): 311–9.

Hofer, T. P., R. A. Hayward, S. Greenfield, E. H. Wagner, S. H. Kaplan, and W. G. Manning. 1999. "The Unreliability of Individual Physician "Report Cards" for Assessing the Costs and Quality of Care of a Chronic Disease." *Journal of the American Medical Association* 281 (22): 2098–105.

Morris, C. N. 1983. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association* 78 (22): 47–55.

Normand, S. T., M. E. Glickman, and C. A. Gatsonis. 1997. "Statistical Methods for Profiling Providers of Medical Care: Issues and Applications." *Journal of the American Statistical Association* 92 (439): 803–14.

O'Brien, S. M., D. M. Shahian, E. R. DeLong, S. L. Normand, F. H. Edwards, V. A. Ferraris, C. K. Haan, J. B. Rich, C. M. Shewan, R. S. Dokholyan, R. P. Anderson, and E. D. Peterson. 2007. "Quality Measurement in Adult Cardiac Surgery: Part 2–Statistical Considerations in Composite Measure Scoring and Provider Rating." *Annals of Thoracic Surgery* 83 (4 suppl): S13–26.

Parker, J. P., Z. Li, C. L. Damberg, B. Danielsen, and D. M. Carlisle. 2006. "Administrative versus Clinical Data for Coronary Artery Bypass Graft Surgery Report Cards: The View from California." *Medical Care* 44 (7): 687–95.

Shahian, D. M., D. F. Torchiana, R. J. Shemin, J. D. Rawn, and S. L. Normand. 2005. "Massachusetts Cardiac Surgery Report Card: Implications of Statistical Methodology." *Annals of Thoracic Surgery* 80 (6): 2106–13.

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.

Steinbrook, R. 2006. "Public Report Cards–Cardiac Surgery and Beyond." *New England Journal of Medicine* 355 (18): 1847–9.

The Leapfrog Group's website. 2010 [accessed on March 5, 2010]. Available at http://www.leapfroggroup.org/news/leapfrog_news/4729468

The National Quality Forum. 2006. "National Quality Forum's National Voluntary Consensus Standards for Hospital Care: Additional Priority Areas 2005–2006 Report" [accessed on March 5, 2010]. Available at http://www.qualityforum.org

Zaslavsky, A. M., N. D. Beaulieu, B. E. Landon, and P. D. Cleary. 2000. "Dimensions of Consumer-Assessed Quality of Medicare Managed-Care Health Plans." *Medical Care* 38 (2): 162–74.

SUPPORTING INFORMATION

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

Appendix SA1: Author Matrix.
Appendix SA2. Reliability Adjustment of Hospital Mortality Rates.

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.