Composite Measures for Profiling Hospitals on Surgical Morbidity

Justin B. Dimick, MD, MPH, Douglas O. Staiger, PhD,† Bruce L. Hall, MD, PhD,‡ Clifford Y. Ko, MD, MHS,‡ and John D. Birkmeyer, MD**

Context: Although risk-adjusted morbidity is widely used as a surgical quality indicator, it may not always be a reliable indicator of hospital quality. In this study, we assess the value of a novel composite measure for improving the reliability of hospital morbidity rankings.

Design, Setting, and Patients: Using data from the American College of Surgeons' National Surgical Quality Improvement Program (ACS-NSQIP), we studied all patients undergoing 4 surgical procedures (2008–2009): colectomy, ventral hernia repair, abdominal aortic aneurysm repair, and lower extremity bypass surgery. For these procedures, we created a composite measure by combining quality indicators from several distinct domains of quality: morbidity, reoperation, length of stay, and morbidity with other potentially related procedures. We empirically weighted each measure and adjusted for reliability using empirical Bayes techniques. To validate this approach, we assessed how well composite measures from 1 year (2008) predict morbidity in the next year (2009) compared with the standard ACS-NSQIP approach for assessing hospital rates of risk-adjusted morbidity.

Results: For all 4 operations, the composite measures explained a higher proportion of hospital-level variation in morbidity than the standard approach: ventral hernia repair (58% for the composite vs 8% for the standard approach), colon resection (33% vs 14%), abdominal aortic aneurysm repair (51% vs 38%), and lower extremity bypass surgery (32% vs 3%). When evaluating the ability to discriminate future performance, the composite approach performed best for ventral hernia repair. For this procedure, the bottom 20% of hospitals based on the composite approach had nearly threefold higher (odds ratio: 2.65; 95% confidence interval: 1.83–3.85) morbidity rates than the top 20% of hospitals. However, when using the standard approach, there was only a 1.3-fold difference (odds ratio: 1.30; 95% confidence interval: 0.87–1.96). Although the differences were smaller in magnitude, the composite measure also outperformed the standard approach for the other 3 procedures.

Conclusions: Composite measures better reflect hospital quality than simple rates of risk-adjusted morbidity. In the context of ACS-NSQIP, composite measures would give hospitals a better sense of where they stand and help identify truly exemplary hospitals for benchmarking.

Keywords: hospital, morbidity, outcomes, quality

(*Ann Surg* 2013;257: 67–72)

From the *Center for Healthcare Outcomes and Policy, University of Michigan, Ann Arbor; †Department of Economics, Dartmouth College, and the Dartmouth Institute for Health Policy and Clinical Practice, Hanover, NH; and ‡Division of Research and Optimal Patient Care, American College of Surgeon, Chicago, IL.

Disclosure: This study was supported by a career development award to J.B.D. from the Agency for Healthcare Research and Quality (K08 HS017765). The views expressed herein do not necessarily represent the views of Center for Medicare and Medicaid Services or the United States Government. The authors declare no conflicts of interest.

Presented, in part, at the American College of Surgeons' 97th Annual Clinical Congress, October 2011, Washington, DC.

Reprints: Justin B. Dimick, MD, MPH, Center for Healthcare Outcomes and Policy, 2800 Plymouth Road, Building 16, Room 137E, Ann Arbor, MI 48109. E-mail: jdimick@umich.edu.

ISSN: 0003-4932/12/25701-0067
DOI: 10.1097/SLA.0b013e31827b6be6

Risk-adjusted morbidity is widely used as an indicator of hospital performance with surgery, including by the American College of Surgeons' National Surgical Quality Improvement Program (ACS-NSQIP).[1,2] Although clinically intuitive for surgeons, simple rates of risk-adjusted morbidity may not reliably reflect hospital performance with surgery. Because of low event rates or low hospital caseloads, hospital morbidity rates may be too imprecise (ie, noisy) to correctly identify high and low performing hospitals.[3,4]

Composite measures may be a more effective approach for capturing a hospital's quality with surgical care.[5–7] Compared with rates of risk-adjusted morbidity, composite measures are more effective at addressing problems with statistical "noise." By combining multiple quality indicators for a single operation (eg, morbidity, length of stay, reoperation), this approach strengthens the quality signal and improves reliability. Moreover, composite measures can further improve precision by adding quality information from other related procedures.[8,9] Prior studies demonstrate the superiority of these techniques for profiling hospitals on mortality, but it is unclear whether this approach will also be useful for risk-adjusted morbidity.[5]

In this context, we sought to evaluate whether composite measures could be used to improve the reliability of risk-adjusted morbidity. Using data from the ACS-NSQIP, we developed and evaluated composite measures for several common, high-risk procedures. Each measure was developed by empirically weighting several input measures, including quality indicators for the index operation and other potentially related operations. We then assessed the ability of these measures to explain systematic variation in hospital-level morbidity and predict future risk-adjusted morbidity compared with simple rates of risk-adjusted hospital morbidity.

METHODS

Data Source and Study Population

We used data from the 2008 and 2009 ACS-NSQIP. The ACS-NSQIP is a prospective, multi-institutional, clinical registry created to give feedback on risk-adjusted outcomes to hospitals for quality improvement purposes and includes all participating centers with data for both 2008 and 2009.[1,2] More than 130 pre- and postoperative variables are recorded, including patient demographics, preoperative risk factors, patient laboratory values, intraoperative variables, and postoperative 30-day morbidity and mortality. The data collection process relies on a sampling strategy aimed at collecting a diverse set of operations. Trained surgical clinical nurse reviewers record the data using standardized definitions. The accuracy of the data is ensured through intensive training mechanisms for the surgical clinical nurse reviewers and by conducting interrater reliability audits of participating sites. For this study, we used appropriate Current Procedure Terminology codes to identify all patients undergoing 1 of 4 common, high-risk procedures (ventral hernia repair, abdominal aortic aneurysm repair, colectomy, and pancreatectomy) at participating hospitals. This work was conducted in conjunction with American College of Surgeons' staff for the purpose of innovation and enhancement of the ACS-NSQIP platform; we therefore had access to deidentified hospital-level dummy variables, which were necessary for this project.

Hospital Morbidity Rates

We used standard ACS-NSQIP techniques for calculating risk-adjusted morbidity rates for each hospital.[1] For the purposes of this study, we limited our assessment to serious morbidity, which included organ space infection, wound dehiscence, pneumonia, unplanned reintubation, postoperative bleeding, stroke, acute myocardial infarction, acute renal failure, sepsis, septic shock, deep venous thrombosis, and pulmonary embolism. Data on complications were ascertained by trained nurse clinicians according to standardized definitions. Hospital morbidity rates were risk-adjusted using detailed data on patient characteristics. For this risk-adjustment, stepwise logistic regression was used to create models that included all significant patient-level covariates.[10] The predicted probabilities of each patient were estimated from this model and then summed for each hospital to calculate the "expected" number of deaths. The observed number of deaths was then divided by the expected number to yield an "O/E ratio." This ratio was then multiplied by the overall average to yield a risk-adjusted morbidity rate for each hospital.

Composite Morbidity Measure

We developed a composite measure that incorporates information from multiple quality indicators to optimally predict "true" risk-adjusted morbidity for each operation. In creating these measures, we considered several individual quality measures, including morbidity rates, reoperation, and length of stay. For each operation, we considered morbidity not only for the index operation but also for other related procedures (eg, colectomy morbidity rates were tested as inputs to the composite measure for other general surgery procedures).

Our composite measure is a generalization of the standard shrinkage estimator that places more weight on a hospital's own morbidity rate when it is measured reliably, but shrinks back toward the average morbidity when a hospital's own morbidity is measured with error (eg, for hospitals with small numbers of patients undergoing the procedure).[5,7] Although the simple shrinkage estimator is a weighted average of a single measure of interest and its mean, our composite measure is a weighted average of all available quality indicators—the morbidity rates for all procedures that are thought to be potentially related. The weight on each quality indicator is determined for each hospital to minimize the expected mean squared prediction error, using an empirical Bayes methodology.[7]

Although the statistical methods used to create these measures are described in detail elsewhere,[7] we will provide a brief conceptual overview. The first step in creating the composite measure was to determine the extent to which each individual quality indicator predicts risk-adjusted morbidity for the index operation. To evaluate the importance of each potential input, we first estimated the proportion of systematic (ie, nonrandom) variation in risk-adjusted morbidity explained by each individual quality indicator (Table 1). We included any quality indicator in the composite measure that explained more than 10% of hospital variation in risk-adjusted morbidity during 2008.

Next, we calculated weights for each quality indicator. The weight placed on each quality indicator in our composite measure was based on 2 factors.[7] The first is the hospital-level correlation of each quality indicator with the morbidity rate for the index operation. The strength of these correlations indicates the extent to which other quality indicators can be used to help predict morbidity for the index operation. The second factor affecting the weight placed on each quality indicator is the reliability with which each indicator is measured. Reliability ranges from 0 (no reliability) to 1 (perfect reliability).[11] The reliability of each quality indicator refers to the proportion of the overall variance that is attributable to true hospital-level variation in performance, as opposed to estimation error ("noise"). For example, in smaller hospitals, less weight is placed on mortality and morbidity rates because they are less reliably estimated. We assume that structural characteristics of each hospital (such as hospital volume) are not estimated with error and, therefore, have reliability equal to 1.

Analysis

We determined the value of our composite measure by determining how well it predicted risk-adjusted morbidity in the next year (2009). For each operation, hospitals were ranked based on the composite measure (data from 2008) and assigned 1 of 3 rankings (1-star, 2-star, and 3-star). The "worst" hospitals (bottom 20%) received a 1-star rating, the middle of the distribution (60%) received a 2-star rating, and the "best" hospitals (top 20%) received a 3-star rating. Many hospital rating systems determine tiers of performance by designating high and low outliers by testing for statistically significant differences from the average. Because we used empirical Bayes methods, which adjust each hospital's composite for imprecision (ie, hospital rankings are a valid indicator of relative performance), we used percentile cutoffs. We then calculated the risk-adjusted mortality rates for 1-star, 2-star, and 3-star hospitals during the subsequent 2 years (data from years 2007–2008). We next assessed the ability of our composite measure to predict future performance compared with standard techniques for ranking hospitals on risk-adjusted morbidity. For these analyses, we evaluated the discrimination in future risk-adjusted morbidity, comparing the 1-star hospitals (bottom 20%) to the 3-star hospitals (top 20%) for each of the measures.

We also assessed and compared the ability of the composite measure and standard risk-adjusted morbidity (assessed in 2008) to explain future (2009) hospital-level variation in risk-adjusted morbidity. To avoid problems with "noise variation" in the subsequent time period, we determined the proportion of systemic hospital-level variation explained. We generated hierarchical models with morbidity as the dependent variable (2009) and used them to estimate the hospital-level variance. We first used an "empty model" that contained only patient variables for risk adjustment. We then entered each historical quality measure (assessed in 2008) into the model. We then calculated the degree to which the historical quality measures reduced the hospital-level variance, an approach described in our prior work.[7] All statistical analyses were conducted using STATA 10.0 (StataCorp LP, College Station, TX).

RESULTS

Incidence of Morbidity

The incidence of each individual complication for each of the 4 procedures is shown in Table 1. The proportion of patients with at least 1 serious complication varied across procedures, from 6.8% for ventral hernia repair to 13.7% with abdominal aortic aneurysm repair (Table 1). The most common complications included pneumonia, unplanned reintubation, prolonged mechanical ventilation, sepsis, and acute renal failure (Table 1).

Inputs to the Composite Measure

For each of the 4 procedures, several individual measures explained a significant proportion of hospital-level variation in risk-adjusted morbidity (Table 2). The amount of hospital-level variation explained by each procedure's own morbidity rate varied, ranging from 60% with colon resection to only 17% for abdominal aortic aneurysm repair (Table 2). Morbidity with other related procedures was important in explaining hospital-level variation for all 4 procedures (Table 2). For example, morbidity with colectomy and

TABLE 1. Incidence of Each Complication Included in Our 30-Day Morbidity Measure

Complication	Ventral Hernia Repair	Colon Resection	Abdominal Aortic Aneurysm Repair	Lower Extremity Bypass Surgery
Overall (1 or more)	6.8%	16.3%	13.7%	10.7%
Organ space infection	1.4%	3.9%	0.6%	0.5%
Wound dehiscence	1.1%	1.9%	0.7%	1.8%
Pneumonia	1.6%	3.9%	4.5%	2.2%
Reintubation	1.3%	3.4%	4.2%	2.8%
Prolonged ventilation	1.7%	6.4%	7.9%	3.2%
Acute renal failure	0.4%	1.1%	3.0%	1.2%
Acute renal injury	0.5%	0.9%	1.5%	0.9%
Myocardial infarction	0.1%	0.3%	0.7%	0.7%
Postoperative bleeding	0.2%	0.9%	2.3%	1.0%
Pulmonary embolism	0.7%	0.9%	0.5%	0.2%
Stroke	0.2%	0.5%	0.7%	0.8%
Sepsis	2.6%	5.0%	2.8%	3.2%
Septic shock	0.9%	3.7%	0.7%	1.8%
Coma	0.05%	0.3%	0.2%	0.1%
Cardiac arrest	0.4%	0.9%	1.4%	1.2%
Acute graft failure	0.0%	0.0%	0.8%	4.7%

TABLE 2. Components of the Composite Measure Are Shown, Along With the Proportion of Nonrandom Hospital-Level Morbidity Explained by Each

Procedure	Individual Quality Measures	Proportion of Hospital-Level Variation Explained
Ventral hernia repair	Index operation	
	Morbidity rate	40%
	Length of stay	15%
	Reoperation rate	12%
	Other operations	
	Morbidity with colectomy	26%
	Morbidity with esophagectomy	12%
	Morbidity with liver resection	13%
	Morbidity with pancreatectomy	36%
Colon resection	Index operation	
	Morbidity rate	60%
	Length of stay	16%
	Reoperation rate	21%
	Other operations	
	Morbidity with appendectomy	12%
	Morbidity with cholecystectomy	13%
	Morbidity with liver resection	15%
	Morbidity with pancreatectomy	10%
	Morbidity with proctocolectomy	11%
	Morbidity with ventral hernia repair	15%
Abdominal aortic aneurysm repair	Index operation	
	Morbidity rate	17%
	Length of stay	29%
	Reoperation	5%
	Other operations	
	Morbidity with ventral hernia repair	11%
Lower extremity bypass surgery	Index operation	
	Morbidity rate	41%
	Length of stay	11%
	Reoperation rate	13%
	Other operations	
	Morbidity with gastric bypass	19%

pancreatectomy explains 26% and 36% of the hospital-level variation in risk-adjusted morbidity with ventral hernia repair, respectively.

Hospital length of stay with the index procedure also explained a large proportion of hospital-level variation in morbidity, varying from 29% with abdominal aortic aneurysm repair to 11% with lower extremity bypass surgery (Table 2). Similarly, hospital reoperation rates explained up to 21% of hospital-level morbidity with colon resection, but only 5% for abdominal aortic aneurysm repair.

Ability of the Composite Measure to Explain Hospital-Level Variation

Table 3 shows the characteristics of the patients going to hospitals. The composite measures explained a high proportion of systematic hospital-level variation in subsequent risk-adjusted morbidity (Table 4). For each operation, the composite measure explained a much higher proportion of variation than the standard approach to measuring morbidity: ventral hernia repair (58% vs 8%), colon resection (33% vs 14%), abdominal aortic aneurysm repair (51% vs 38%), and lower extremity bypass surgery (32% vs 3%) (Table 4).

Ability of the Composite Measure to Predict Future Performance

The composite score created by combining these individual measures performed well at predicting future hospital performance (Table 4, Fig. 1). For all 4 procedures, the composite measure based on 2008 data was better at discriminating future performance in 2009 than the standard approach to measuring risk-adjusted morbidity (Table 4, Fig. 1). For example, with ventral hernia repair, historical risk-adjusted morbidity predicted a smaller difference between the best (top 20%) and worst (bottom 20%) hospitals (odds ratio: 1.30; 95% confidence interval: 0.87–1.96) when compared with the composite measure (odds ratio: 2.65; 95% confidence interval: 1.83–3.85) (Table 4). These differences in mortality could not be explained by differences in patient severity of illness, as the differences in patient characteristics shown in Table 3 were adjusted for in all comparisons.

DISCUSSION

In this study, we demonstrate the value of a novel composite measure for profiling hospitals on risk-adjusted rates of surgical morbidity. The Achilles heel of outcomes measurement is unreliability due to small sample size and low event rates. As a result, the standard approach for assessing risk-adjusted morbidity, and other surgical outcomes, is prone to misclassification of surgeons and hospitals. The composite measure described in this article addresses this problem in 2 ways: (1) applying statistical techniques for filtering out noise and (2) borrowing signals wherever they are available, including from other related operations. In this study, we demonstrate that such a composite measure that integrates multiple outcomes, including morbidity with other related procedures, is a better predictor of hospital performance than standard approaches for assessing risk-adjusted morbidity.

Several previous studies demonstrate the potential value of composite measures of surgical quality. In prior work, using Medicare claims data, we have established the value of empirically weighted composite measures for predicting future mortality. This work has shown that composite measures of structure (eg, hospital volume) and multiple outcomes (eg, with the index operation and with other related procedures) are better than existing approaches for assessing quality with surgery.[5,7] In this prior work, we focused exclusively on mortality, which is relatively uncommon for most surgical procedures. Because morbidity rates are much more common than mortality, and therefore less plagued by problems with statistical "noise," it would be reasonable to expect that composite measures would have less value when used to assess nonfatal outcomes. However, in this study, we

TABLE 3. Patient Characteristics for the Best, Middle, and Worst Hospitals in 2008

	Hospitals Ranked on 2008 Composite Measures		
Procedure	Worst 20% of Hospitals	Middle 60% of Hospitals	Best 20% of Hospitals
Ventral hernia repair			
Age, mean	60	58	56
Nonwhite race, %	23	23	23
Male, %	39	37	42
Emergent surgery, %	8.0	12.6	11.6
Expected morbidity rates, %	6.1	7.1	6.7
Colon resection			
Age, mean	63	63	60
Nonwhite race,%	22	21	29
Male, %	50	53	53
Emergent surgery, %	14.7	18.0	17.6
Expected morbidity rates, %	15.2	16.7	16.3
Elective abdominal aortic aneurysm repair			
Age, mean	73	73	73
Non-white race, %	13	15	17
Male, %	81	78	80
Emergent surgery, %	8.2	12.3	13.4
Expected morbidity rates, %	12.1	13.8	15.2
Lower extremity bypass surgery			
Age, mean	68	66	66
Non-white race, %	21	22	38
Male, %	63	63	61
Emergent surgery, %	6.4	6.0	7.4
Expected morbidity rates, %	10.7	10.7	10.6

TABLE 4. Relative Ability of the Composite Measure and Standard Risk-Adjusted Morbidity From 2008 to Forecast Risk-Adjusted Morbidity in 2009

	2009 Risk-Adjusted Morbidity	
2008 Hospital Rankings	**Odds Ratio, "Best" vs "Worst" Hospitals (95% Confidence Interval)**	**% Hospital-Level Variation Explained**
Ventral hernia repair		
Composite measure	2.65 (1.83–3.85)	58%
Risk-adjusted morbidity	1.30 (0.87–1.96)	8%
Colon resection		
Composite measure	1.70 (1.41–2.04)	33%
Risk-adjusted morbidity	1.47 (1.21–1.78)	14%
Abdominal aortic aneurysm repair		
Composite measure	1.72 (1.20–2.45)	51%
Risk-adjusted morbidity	1.35 (0.95–1.92)	38%
Lower extremity bypass surgery		
Composite measure	2.05 (1.42–2.95)	32%
Risk-adjusted morbidity	1.33 (0.91–1.93)	3%

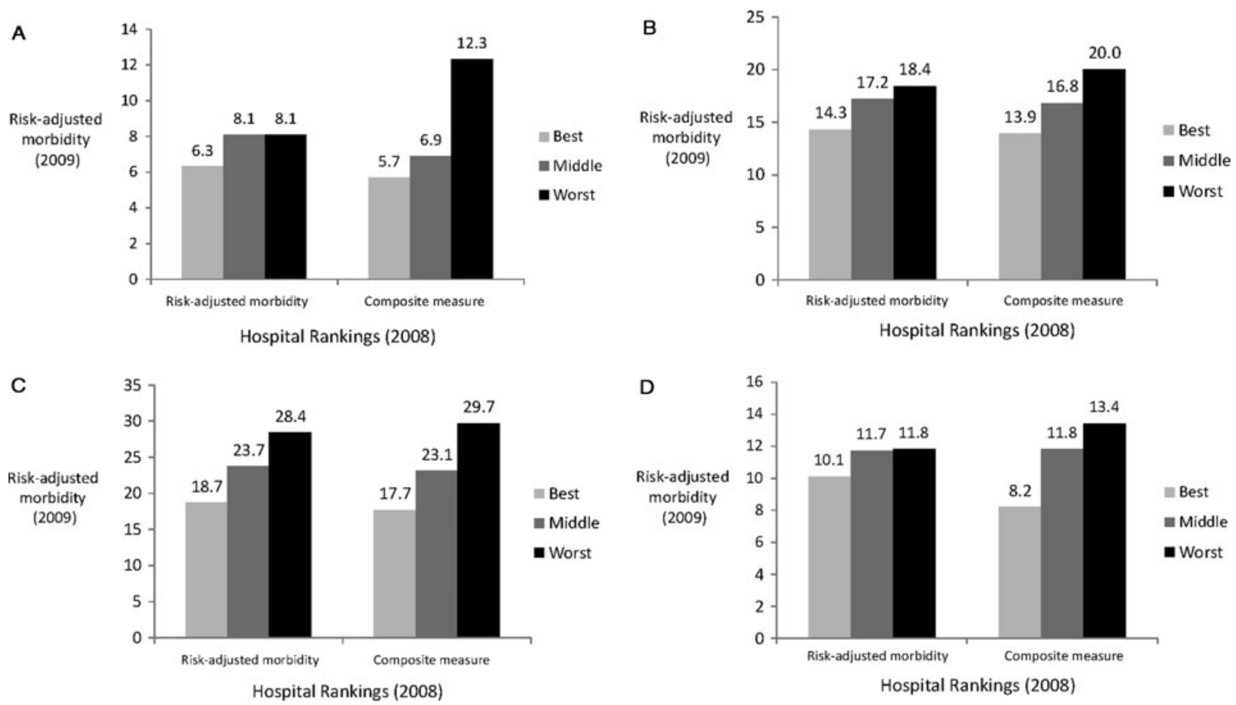

FIGURE 1. Future risk-adjusted mortality rates (2009) for the "best" (top 20%), "middle," and "worst" (bottom 20%) hospitals as assessed using the composite measure and standard ACS-NSQIP techniques in the previous year (2008). A, Ventral hernia repair. B, Colon resection. C, Elective abdominal aortic aneurysm repair. D, Lower extremity bypass surgery.

found that composite measures also improved the predictive ability of morbidity measures. Thus, we believe this technique would be useful to improve the reliability of efforts to profile both hospital-level morbidity and mortality.

The findings of this study also demonstrate the value of incorporating information from other surgical procedures into a composite quality score. For each procedure, we found that adding risk-adjusted morbidity rates with "other" procedures enhanced the reliability of hospital performance assessment. The ability to "borrow" signals from these other operations reflects the presence of shared structure and process that lead to better outcomes for all surgical procedures, including nurse-to-patient ratios, quality improvement to the infrastructure, and adherence to evidence-based perioperative practices. Previous studies showing strong hospital-level correlations in surgical outcomes for different procedures (eg, coronary artery bypass surgery and cardiac valve surgery) are consistent with these findings.[8,9]

The results of this study should be viewed in the context of certain limitations. Because the ACS-NSQIP uses a sampling strategy (ie, the registry does not capture 100% of cases in a hospital), our results may not be applicable to other quality measurement platforms. With 100% of the cases, the standard approach of assessing rates of hospital risk-adjusted morbidity would likely be more reliable. If morbidity rates were more reliable, the additional "signal" gained from other measures may not be as important. However, in our prior work, we have seen the benefits of the composite measure persist in data sources that capture all patients. This study is also limited by the lack of information on structural characteristics, such as hospital volume. In our prior work, hospital volume is one of the most important inputs to the composite measure.[5] If we added structural characteristics, the composite measure would likely be an even better predictor of future morbidity.

Composite measures are increasingly used in several real world quality measurement activities. The Center for Medicare and Medicaid Services/Premier uses composite measures of quality with cardiac and orthopedic surgery for their pay-for-performance program. The Society of Thoracic Surgeons also uses a composite quality measure for evaluating global performance with coronary artery bypass surgery. Both of these composite scores combine various process and outcome measures by assigning equal weights to each quality measure. In contrast, the approach described in this article empirically weights inputs on the basis of their reliability and the strength of their relationship to some gold standard outcome, such as risk-adjusted morbidity. We believe empirical weighting of inputs to a quality measure is superior to expert opinion or equal weighting because it allows systematic decision-making about which measures to include and how much weight they should be afforded. Ultimately, this approach can be used to maximize efficiency by only collecting information on those measures that are most important.

The Leapfrog Group, a large coalition of health care purchasers, currently uses an approach analogous to the one in this article for their evidence-based hospital referral program.[5] This measure combines hospital mortality and provider volume into a single score that reflects the likelihood a patient will survive surgery through 5 complex operations. These "Survival Predictor" scores are publicly reported on the Leapfrog Group Web site. These techniques were vetted and subsequently endorsed by the National Quality Forum for use with 3 high-risk surgical procedures: pancreatic resection, esophageal resection, and abdominal aortic aneurysm repair.

In addition to their value in public reporting, these composite measures could also be useful for quality improvement in the context of the ACS-NSQIP and other reporting platforms. As discussed earlier, standard approaches to surgical outcome measurement are plagued by statistical "noise" and imprecision, which translates into inaccurate assessments of relative hospital (or physician) performance. Such inaccurate assessments of performance can lead to both false positives (ie, hospitals perceive a problem where none exists) and false negatives (ie, hospitals miss a problem when it really does exist). Improving the reliability of outcome measurement is therefore an important goal of any quality measurement platform.

The ACS-NSQIP has been forward-looking and is currently implementing changes to enhance the reliability of their risk-adjusted outcome measures. One of the most important innovations will be the adoption of a "procedure-targeted" approach for data collection. Rather than a sample of all surgical procedures at a hospital, the next generation of ACS-NSQIP will collect up to 100% of the cases for several common, high-risk procedures (eg, colon resection). As noted earlier, this change will increase the sample size and improve the reliability of outcome measures. Moreover, there are plans to change to a hierarchical modeling technique that separates the "signal" from the "noise,"—a so-called reliability adjustment. The technique presented in this article, which brings in all available information to optimally predict procedure-specific morbidity, is the next logical step toward improving the reliability of outcome measurement. These composite measures could further improve the reliability of benchmarking by the ACS-NSQIP and give providers a truer sense of where they stand relative to their peers.

REFERENCES

1. Hall BL, Hamilton BH, Richards K, et al. Does surgical quality improve in the American College of Surgeons National Surgical Quality Improvement Program: an evaluation of all participating hospitals. *Ann Surg.* 2009;250:363–376.
2. Khuri SF, Henderson WG, Daley J, et al. Successful implementation of the Department of Veterans Affairs' National Surgical Quality Improvement Program in the private sector: the Patient Safety in Surgery study. *Ann Surg.* 2008;248:329–336.
3. Kao LS, Ghaferi AA, Ko CY, et al. Reliability of superficial surgical site infections as a hospital quality measure. *J Am Coll Surg*. 2011;213:231–235.
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, Staiger DO, Baser O, et al. Composite measures for predicting surgical mortality in the hospital. *Health Aff (Millwood)*. 2009;28:1189–1198.
6. 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(suppl):S13–S26.
7. Staiger DO, Dimick JB, Baser O, et al. Empirically derived composite measures of surgical performance. *Med Care*. 2009;47:226–233.
8. Dimick JB, Staiger DO, Birkmeyer JD. Are mortality rates for different operations related? Implications for measuring the quality of noncardiac surgery. *Med Care*. 2006;44:774–778.
9. Goodney PP, O'Connor GT, Wennberg DE, et al. Do hospitals with low mortality rates in coronary artery bypass also perform well in valve replacement? *Ann Thorac Surg*. 2003;76:1131–1136; discussion 1136–1137.
10. Dimick JB, Osborne NH, Hall BL, et al. Risk adjustment for comparing hospital quality with surgery: how many variables are needed? *J Am Coll Surg.* 2010;210:503–508.
11. Dimick JB, Staiger DO, Birkmeyer JD. Ranking hospitals on surgical mortality: the importance of reliability adjustment. *Health Serv Res.* 2010;45():1614–1629.