Reliability of Surgical Outcomes for Predicting Future Hospital Performance

Robert W. Krell, MD, Douglas O. Staiger, PhD,† and Justin B. Dimick, MD, MPH**

Background: Because of small sample sizes and low event rates, risk-adjusted surgical outcomes often do not meet reliability benchmarks for distinguishing hospital performance. Nonetheless, it is unclear whether these measures may still be useful for predicting future hospital surgical performance.

Methods: We used national Medicare data to analyze patients undergoing colectomy from 2007 to 2010 (n = 462,959 patients). We first quantified 2007–2008 outcome reliability (ability to differentiate quality differences) and ranked hospitals based on their 2007–2008 risk-adjusted outcome rates. To assess the ability of adjusted outcomes to predict true performance, we evaluated future (2009–2010) outcomes across quintiles of past performance. We then systematically sampled 2007–2008 cases to evaluate performance prediction when hospitals' past performance was measured with progressively lower reliability levels.

Results: Outcomes in 2007–2008 were good predictors of outcomes in the next 2 years (2009–2010), but predictive strength depended upon reliability. With progressive sampling of 2007–2008 caseloads, outcome reliability and predictive strength decreased. With 100% sampling of 2007–2008 caseloads, the worst versus best hospital quintile based on past performance had 1.52 [95% confidence interval (CI), 1.44–1.60] times the odds of mortality and 1.50 (95% CI, 1.44–1.56) times the odds of complications in 2009–2010. With 10% sampling, outcome reliability was well below commonly accepted benchmarks, but the worst quintile of hospitals in 2007–2008 still had 1.12 (95% CI, 1.06–1.19) times the odds of mortality and 1.16 (95% CI, 1.11–1.21) times the odds of complications in 2009–2010 compared with the best quintile of hospitals.

Conclusions: Even at very low reliability levels, risk-adjusted outcome measures may distinguish best and worst hospitals' surgical performance. This study suggests that commonly accepted reliability thresholds may be too high, especially in the context of selective referral.

From the *Department of Surgery, University of Michigan Health System, Ann Arbor, MI; and †Department of Economics, Dartmouth College, Hanover, NH.

R.W.K. is supported by NIH Grant 5T32CA009672-22.

J.B.D. and D.O.S. have a financial interest in ArborMetrix Inc., which had no involvement in the collection or analysis of any data, nor the interpretation of any results presented herein. R.W.K. received payment from Blue Cross/Blue Shield of Michigan for data entry, unrelated to the submitted work.

Reprints: Robert W. Krell, MD, Center for Healthcare Outcomes and Policy, 2800 Plymouth Road, Building 16, Office 016-100 N-13, Ann Arbor, MI 48109. E-mail: rkrell@med.umich.edu.

ISSN: 0025-7079/14/5206-0565

Key Words: reliability, postoperative complications, hierarchical modeling

(*Med Care* 2014;52: 565–571)

Performance measures are increasingly prominent in policy initiatives for identifying high-quality hospitals. Perhaps, because it is easy to measure, hospital volume has been central to many existing programs. For example, to identify top-performing hospitals selective referral initiatives such as the Leapfrog Group, Blue Cross Distinction Centers, and Centers of Excellence programs establish procedure volume thresholds to identify hospitals where patients should undergo high-risk surgery.[1–4] Because of increased recognition of the limitations of volume thresholds for identifying top-performing centers, some authors have suggested using direct outcome measures to benchmark hospital surgical performance.[5–9]

However, it is unclear whether outcomes can be reliably used to assess hospital performance. Analogous to power calculations designed to minimize type II error (failure to detect a difference between groups) in clinical trials, reliability denotes outcomes' ability to distinguish quality differences between providers.[10,11] The existing literature highlights the inability of most hospitals' surgical outcomes to meet established reliability thresholds.[11–14] However, there is little empiric evidence validating these existing reliability thresholds. A major challenge in establishing reliability thresholds is the lack of a method for evaluating them—that is, how do we know when an outcome measure is "reliable enough"? For public reporting and selective referral initiatives, the ability to predict future performance is arguably the best criterion of an outcome's usefulness, because patients decide where to undergo surgery *now* based on historical hospital performance.

In this context, we sought to evaluate the ability of outcome measures of different reliability to predict future performance. We used 4 years of Medicare data to assess the ability of outcomes following colon resections from one time period (2007–2008) to predict future outcomes (2009–2010), when the outcomes were measured using progressively lower sample sizes and reliability levels.

METHODS

Data Source, Study Population, and Outcomes

We used data from the 2007–2010 Medicare Provider Analysis and Review files, which include hospital discharge

information and all fee-for-service acute care hospitalizations for Medicare beneficiaries. Using relevant International Classification of Diseases, 9th Revision, Clinical Modification codes, we identified all patients aged 65–99 years undergoing colorectal resections to form our study cohort.

Hospital outcomes included risk-adjusted mortality (death within 30 d of operation or before hospital discharge), complications, and reoperation for any reason. We identified complications and reoperations from International Classification of Diseases, 9th Revision, Clinical Modification codes using established methods for assessing administrative databases.[15,16] Complications included respiratory failure (518.81, 518.4, 518.5, 518.8), pneumonia (481, 482.0–482.9, 483, 484, 485, 507.0), myocardial infarction (410.00–410.91), venous thromboembolism (415.1, 451.11, 451.19, 451.2, 451.81, 453.8), renal failure (584), postoperative hemorrhage or hematoma (998.1), surgical site infection (958.3, 998.3, 998.5, 998.59, 998.51), or gastrointestinal hemorrhage (530.82, 531.00-531.21, 531.40, 531.41, 531.60, 531.61, 532.00–532.21, 532.40, 532.41, 532.60, 532.61, 533.00-533.21, 533.40, 533.41, 533.60, 533.61, 534.00–534.21, 534.40, 534.41, 534.60, 534.61, 535.01, 535.11, 535.21, 535.31, 535.41, 535.51, 535.61, 578.9). We also assessed serious complications, which we defined as any complication in conjunction with length of hospital stay greater than the 75th percentile for the cohort. Using extended length of stay in conjunction with complication data has been proposed as a means to increase the specificity of the outcome.[5,17] Reoperations included reopening of surgical site or reclosure of dehiscence (5412, 3402-3, 5411, 5471), management of shock/hemorrhage, including splenectomy (3998, 4995, 5793, 60984, 3941, 415), removal of retained foreign body (5492, 9820), management of surgical site infection (540, 5419, 4694), repair of organ injury or wound complications (4461, 4671-6, 4871, 5061, 5581, 5675, 5682, 5686, 5689, 5781, 5783-4, 5841), and management of stoma complications (4640-3).

Analysis

The primary goal of our analysis was to assess the ability for past outcomes to predict hospital performance when measured with decreasing reliability levels. Reliability is a measure of the statistical "power" of an outcome measure and is largely influenced by sample size (ie, caseload).[10,11] In this study, we performed 4 iterations of the same strategy: we ranked hospitals based on their risk-adjusted and reliability-adjusted outcome rates in 2007–2008, and then compared future (2009–2010) risk-adjusted outcomes across quintiles of past hospital performance. To assess the effect of decreasing outcome reliability, we used systematic sampling to lower all hospital caseloads, creating 4 cohorts for analysis: a 100% sampled cohort, a 50% sampled cohort, a 25% sampled cohort, and a 10% sampled cohort.

Calculating Risk-adjusted Outcome Rates

We used multivariable logistic regression models to calculate hospital risk-adjusted outcome rates (mortality, overall complications, serious complications, and reoperations) for 2007–2008. Each model included patient age, sex, race, median ZIP-code income, emergent admission, and comorbidities identified from secondary diagnosis codes using the methods of Elixhauser and colleagues.[18,19] Dividing each hospital's observed events by the sum of its predicted outcomes generated 2007–2008 observed: expected (O/E) outcome ratios, which when multiplied by the overall outcome yielded that hospital's risk-adjusted rate. To account for random outcome variation across hospitals, we further adjusted outcome rates using shrinkage estimators derived from hierarchical modeling and empirical Bayes techniques.[20–22] This practice is also referred to as "reliability adjustment" and is becoming more common in surgical quality reporting platforms.[20,21] Although hierarchical modeling techniques have been shown to improve some outcomes' performance forecasting ability,[22] the reliability levels necessary for performance prediction have not been empirically evaluated. Moreover, the reliability levels at which reliability adjustment fails to provide useful information (fails to predict hospital performance) are unknown.

Calculating Outcome Reliability

As stated previously, reliability is a measure of the statistical "power" of an outcome measure. Mathematically, it is a ratio of quality "signal" (true quality differences) to "noise" (measurement error, which is primarily influenced by sample size and error from risk-adjustment models).[10,22] Reliability estimates range from 0 (no reliability, all provider differences due to measurement error) to 1 (perfect reliability, all differences due to true quality differences). Commonly accepted reliability thresholds for quality reporting are 0.7–0.9, although recently some authors have suggested that 0.5 may be adequate for surgical quality reporting.[10,11,21] We used previously described methods to estimate each outcome's reliability.[13,23] In brief, we used hierarchical logistic regression models assigning the hospital as the higher level in the model. The hospital-level random intercept variance after adjusting for patient factors is used to estimate outcome "signal." We estimated each hospital's "noise" using established methods to estimate the SE of a proportion.[11,22] We then defined hospital outcome reliability as [signal/(signal+hospital "noise")].

We also calculated the mean square root of reliability for each outcome. From classical test theory, the square root of a reliability estimate is the correlation between an observed test result and the "true" result.[24,25] In the present study, mean square root reliability represents the degree of correlation between "true" quality signal (hospital performance based on a measure with perfect reliability) and observed hospital performance (ie, that based on the "noisy" risk-adjusted outcome measure). For example, an outcome measure with 0.70 reliability (a common reliability threshold), would be expected to provide 83.7% correlation between hospital "true" and "observed" quality. As outcome reliability decreases, the degree of correlation between true and observed outcomes decreases as well.

Assessing Predictive Ability

The primary goal of our analysis was to assess the ability for past outcomes to predict hospital's true performance when past outcomes were measured with decreasing reliability. To do this, we grouped hospitals into quintiles based on their risk-adjusted and reliability-adjusted 2007–2008 outcome rates. We

then merged the hospital performance quintiles from 2007 to 2008 with the patient-level data from 2009 to 2010. We used multivariable logistic regression models to calculate risk-adjusted outcome rates across quintiles of past hospital performance. In this portion of the study, the patient was the unit of analysis. Each model adjusted for patient age, sex, race, median ZIP-code income, emergent admission, and comorbidities as above and utilized robust SEs to account for within-hospital outcomes correlation (clustering). Rankings from each 2007–2008 cohort (100% sample, 50% sample, 25% sample, 10% sample) were merged with the full (unsampled) 2009–2010 data to perform the analysis.

Finally, we assessed how caseload sampling influenced both outcome reliability and top-bottom quintile performance discrimination. The mean square root of a reliability estimate should be directly proportional to the difference between predicted outcome rates of the highest and lowest performing hospital quintiles.[24,25] For each outcome, we quantified the proportional change in mean square root of reliability at each level of caseload sampling. We also assessed the proportional change in the difference between 2009 and 2010 adjusted outcome rates of the top and bottom quintiles' at each level of caseload sampling. This analysis quantified the expected differentiation in hospital performance based on a specific outcome's reliability.

We performed all statistical analyses using STATA release 12 (StataCorp, College Station, TX). All statistical tests were 2-sided and P values considered significant if <0.05. The University of Michigan Institutional Review Board approved the study protocol.

TABLE 1. Characteristics of Medicare-eligible Adult Patients 65–99 Years Undergoing Colorectal Resections, 2007–2010

	2007–2008 N = 227,618	2009–2010 N = 235,341
Demographics		
Mean age (SD)*	76.6 (7.4)	76.4 (7.4)
Male (%)	42.0	41.7
Race (%)*		
White	87.8	87.2
Black	8.3	8.7
Other	3.9	4.1
Comorbidities†		
Cardiac (%)*	14.9	12.7
Chronic lung disease (%)*	15.7	13.1
Peripheral vascular disease (%)*	4.4	4.7
Neurological (%)	4.5	4.5
Diabetes (%)*	15.4	15.9
Renal failure (%)*	6.2	5.7
Liver disease (%)	1.0	1.0
Colon cancer diagnosis (%)*	41.2	39.5
Other neoplasm (%)*	20.3	18.6
Depression or psychosis (%)*	4.2	4.5
Obesity (%)*	3.3	4.1
Anemia (%)*	17.5	16.3
Emergent admission (%)*	31.8	32.6
Postoperative outcomes		
30-d or in-hospital mortality (%)*	10.3	9.7
Any complication (%)*	29.4	30.9
Serious complications (%)*	14.8	14.4
Reoperations (%)	1.0	1.0

*Univariate P-value < 0.05.
†As defined by Elixhauser et al.[18]

RESULTS

We identified 462,959 Medicare-eligible adults aged 65–99 who underwent colectomy procedures between 2007 and 2010. Demographic and comorbidity data, as well as unadjusted outcome rates for the 2 study periods are presented in Table 1. Average hospital caseloads in 2007–2008 were 143 cases in the 100% sample cohort, decreasing to 15 cases in the randomly sampled 10% cohort.

Reliability for all outcomes decreased as sample size decreased (Table 2). For example, mean reliability for 30-day overall complications was 0.40 in the 100% sample cohort, whereas in the 10% sample cohort, mean reliability for 30-day overall complications was 0.09. Mean reliability for all outcomes was lower than established benchmarks for all cohorts and decreased in a stepwise manner as caseloads decreased with sampling (Table 2).

Figure 1 and Table 3 demonstrate the ability of 2007–2008 hospital performance rankings to predict future outcomes as caseloads and outcome reliability decreased. Although there was a stepwise change in outcomes based on past performance when hospitals were ranked from a 100% sample, the differences between quintiles became attenuated as more restrictive samples were used to rank hospitals (Fig. 1). The effect of decreasing reliability was most prominent among the hospitals in the middle performance quintiles (Table 3). For example, if a patient underwent surgery in a middle-performing (quintile 3) hospital based on 2007–2008 rankings from the 100% sample cohort, they would have had 1.18 [95% confidence interval (CI), 1.12–1.24] times adjusted odds of experiencing a complication compared with the best performing (quintile 1) hospitals. However, when caseloads were smallest and outcome reliability lowest, that patient would have similar odds of complications compared with the best hospitals (adjusted odds ratio, 1.00; 95% CI, 0.94–1.05) (Table 3). Absolute differences between the top and bottom hospital quintiles decreased with smaller samples as well (Table 3). For example, there was 3.1% mortality difference between the top and bottom quintiles when past performance was measured with 100% sampling, decreasing to a 0.9% absolute difference in quintile performance when 10% sampling was used.

Although the ability to discriminate future performance in the middle quintiles relative to the best performing hospitals was lost at very low caseloads and reliability levels, the ability to discriminate between the best and worst hospitals' future performance was still present (Table 3). For example, based on outcomes from the 10% sampled cohort, if a patient had an operation in the worst versus best performing hospitals based on 2007–2008 performance, they would still have had 1.26 (95% CI, 1.19–1.34) times the odds of experiencing serious complications, 1.16 (95% CI, 1.06–1.19) times the odds of experiencing any complication, and 1.12 (95% CI, 1.06–1.19) times the odds of 30-day mortality (Table 3).

Figure 2 shows the relationship between outcomes' mean square root reliability and their ability to discriminate between the top and bottom hospital quintiles as hospitals

TABLE 2. Risk-adjusted Outcome Rates and Reliability Across Sampled Cohorts, 2007–2008

	100% Sample	50% Sample	25% Sample	10% Sample
Hospitals (N), 2007–2008	3423	3379	3272	2956
2007–2008 hospital caseload [mean (SD)]	143 (105)	72 (53)	36 (27)	15 (11)
30-d mortality				
2007–2008 risk-adjusted outcome rate [mean (SD)] (%)	11.0 (9.8)	11.0 (12.4)	11.0 (14.7)	9.9 (16.9)
2007–2008 outcome reliability [mean (SD)]*	0.29 (0.21)	0.19 (0.15)	0.10 (0.09)	0.06 (0.06)
Outcome correlation with quality "signal" (mean square root reliability)	0.50	0.39	0.28	0.22
Decline in mean square root reliability relative to 100% sampling (%)	Reference	22.0	44.0	56.0
30-d overall complications				
2007–2008 risk-adjusted outcome rate [mean (SD)] (%)	27.7 (13.2)	27.9 (16.1)	27.7 (19.6)	26.7 (23.9)
2007–2008 outcome reliability [mean (SD)]*	0.40 (0.23)	0.25 (0.18)	0.14 (0.11)	0.09 (0.08)
Outcome correlation with quality "signal" (mean square root reliability)	0.60	0.46	0.33	0.28
Decline in mean square root reliability relative to 100% sampling (%)	Reference	22.0	44.1	54.2
30-d serious complications				
2007–2008 risk-adjusted outcome rate [mean (SD)] (%)	12.8 (9.8)	12.9 (11.7)	12.6 (13.8)	12.4 (17.5)
2007–2008 outcome reliability [mean (SD)]*	0.43 (0.25)	0.30 (0.21)	0.20 (0.16)	0.12 (0.10)
Outcome correlation with quality "signal" (mean square root reliability)	0.62	0.50	0.41	0.31
Decline in mean square root reliability relative to 100% sampling (%)	Reference	19.4	33.9	50.0
Reoperation				
2007-08 risk-adjusted outcome rate [mean (SD)] (%)	1.1 (3.1)	1.2 (5.5)	1.1 (6.0)	1.2 (10.2)
2007–2008 outcome reliability [mean (SD)]*	0.05 (0.05)	0.04 (0.04)	0.01 (0.02)	0.01 (0.01)
Outcome correlation with quality "signal" (mean square root reliability)	0.20	0.18	0.10	0.10
Decline in mean square root reliability relative to 100% sampling (%)	Reference	10.0	50.0	50.0

*Commonly quoted reliability benchmarks for quality reporting: 0.50–0.90.[10,21]

were profiled with progressively smaller sample sizes. As sample sizes decreased, mean square root reliability decreased with a corresponding decrease in discrimination between the top and bottom hospital quintiles. With 50% sampling, outcomes' mean square root reliability decreased from 10% to 22% with a corresponding decrease in quintile discrimination of approximately 20%. With 25% sampling, outcome mean square root reliability decreased from 33.9% to 50.0% with a corresponding decrease in top-bottom quintile discrimination from 35% to 50% (Fig. 2). Because

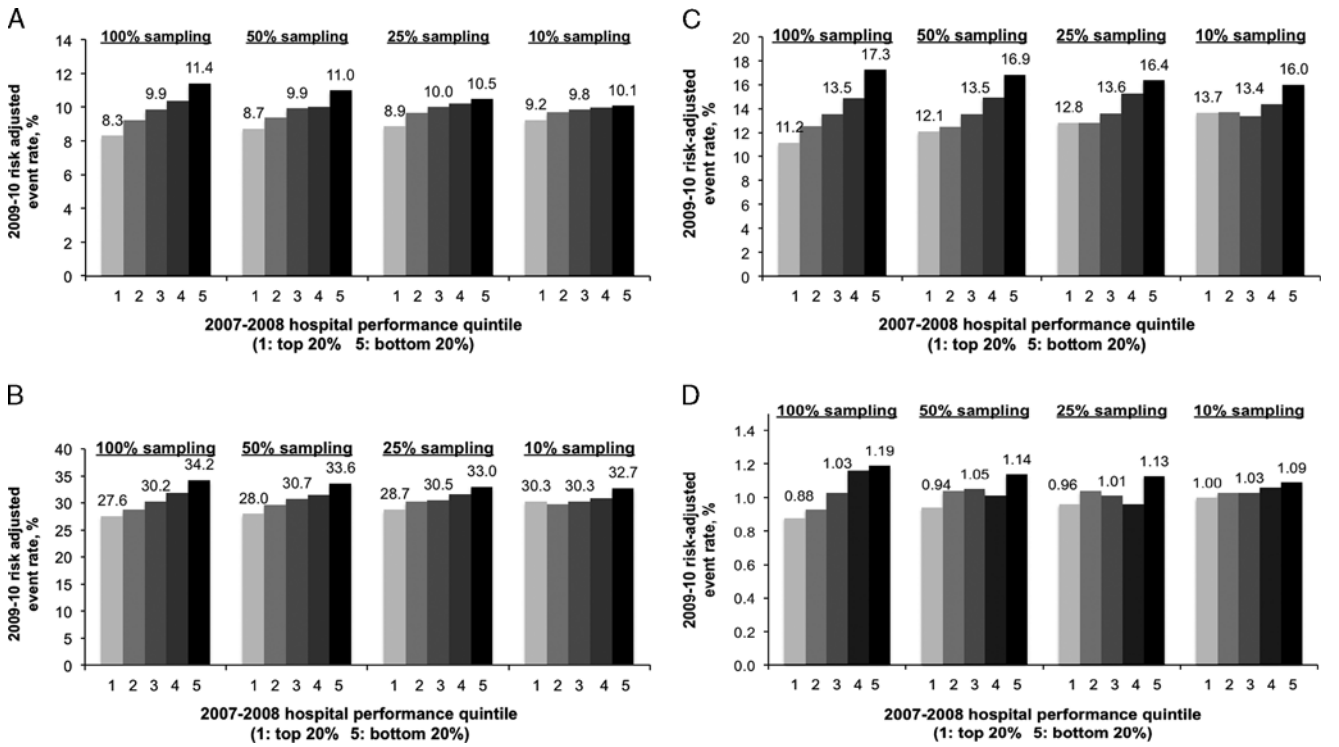

FIGURE 1. Risk-adjusted outcome rates in 2009–2010 across quintiles of past hospital performance derived from different hospital caseload samples. A, Risk-adjusted 30-day mortality. B, Risk-adjusted 30-day overall complications. C, Risk-adjusted serious complications. D, Risk-adjusted reoperations.

TABLE 3. Adjusted Odds of 2009–2010 Adverse Outcomes Based on 2007–2008 Hospital Performance Rankings, and Absolute Differences in Adjusted Outcome Rates Between "Best" and "Worst" Performing Hospitals

	Adjusted Odds Ratio of 2009–2010 Adverse Outcomes (95% CI)			
	100% Sample	**50% Sample**	**25% Sample**	**10% Sample**
30-d mortality				
2007–2008 performance quintile				
1 ("best" performance)	Reference	Reference	Reference	Reference
2	1.15 (1.08–1.21)	1.10 (1.04–1.17)	1.10 (1.04–1.19)	1.07 (1.00–1.14)
3	1.25 (1.17–1.34)	1.19 (1.11–1.27)	1.17 (1.09–1.26)	1.09 (1.01–1.17)
4	1.33 (1.26–1.42)	1.20 (1.13–1.28)	1.20 (1.13–1.27)	1.11 (1.04–1.17)
5 ("worst" performance)	1.52 (1.44–1.60)	1.36 (1.29–1.44)	1.25 (1.18–1.32)	1.12 (1.06–1.19)
Outcome rate difference, first vs. fifth quintiles (% events)	3.1	2.3	1.6	0.9
30-d morbidity				
2007–2008 performance quintile				
1 ("best" performance)	Reference	Reference	Reference	Reference
2	1.08 (1.03–1.13)	1.11 (1.05–1.16)	1.10 (1.04–1.15)	0.97 (0.92–1.02)
3	1.18 (1.12–1.24)	1.18 (1.13–1.24)	1.11 (1.06–1.17)	1.00 (0.94–1.05)
4	1.30 (1.25–1.36)	1.23 (1.18–1.29)	1.19 (1.14–1.25)	1.03 (0.98–1.09)
5 ("worst" performance)	1.50 (1.44–1.56)	1.41 (1.35–1.47)	1.30 (1.25–1.36)	1.16 (1.11–1.21)
Outcome rate difference, first vs. fifth quintiles (% events)	6.6	5.7	4.3	2.4
30-d serious morbidity				
2007–2008 performance quintile				
1 ("best" performance)	Reference	Reference	Reference	Reference
2	1.18 (1.10–1.26)	1.05 (0.97–1.13)	1.03 (0.96–1.10)	1.00 (0.93–1.08)
3	1.31 (1.22–1.40)	1.17 (1.09–1.25)	1.09 (1.00–1.20)	0.97 (0.90–1.05)
4	1.50 (1.42–1.59)	1.35 (1.27–1.43)	1.23 (1.16–1.31)	1.08 (1.01–1.15)
5 ("worst" performance)	1.88 (1.78–1.98)	1.62 (1.54–1.71)	1.42 (1.35–1.51)	1.26 (1.19–1.34)
Outcome rate difference, first vs. fifth quintiles (% events)	6.1	4.8	3.5	2.4
Reoperation				
2007–2008 performance quintile				
1 ("best" performance)	Reference	Reference	Reference	Reference
2	1.07 (0.92–1.23)	1.11 (0.97–1.27)	1.08 (0.96–1.21)	1.03 (0.91–1.16)
3	1.18 (1.00–1.41)	1.12 (0.94–1.33)	1.05 (0.90–1.23)	1.03 (0.91–1.17)
4	1.33 (1.17–1.50)	1.08 (0.93–1.25)	1.00 (0.79–1.27)	1.06 (0.90–1.25)
5 ("worst" performance)	1.37 (1.22–1.53)	1.22 (1.10–1.36)	1.17 (1.05–1.31)	1.08 (0.94–1.25)
Outcome rate difference, first vs. fifth quintiles (% events)	0.31	0.20	0.17	0.09

CI indicates confidence interval.

the reliability of reoperation was very low (0.05) with 100% sampling, the decreases in mean square root reliability and quintile discrimination did not demonstrate a linear relationship as mortality, overall complications, or serious complications did.

DISCUSSION

Despite increased enthusiasm for using outcomes to identify high-performing hospitals, their usefulness for this purpose is not entirely clear. In this study, we hoped to contribute to a deeper understanding of how an outcome measure's reliability relates to its ability to predict future hospital performance. Using 4 years of national Medicare data for a common high-risk procedure, we have demonstrated that hospitals' past surgical performance can predict their future performance, even with small hospital caseload samples and low outcome reliability. This finding has important implications for quality measurement, particularly those efforts aimed at steering patients to the best hospitals. Even measures with very low reliability are still useful for discriminating hospital performance, particularly for the highest and lowest performing hospitals.

Most studies examining outcome reliability have used arbitrary reliability benchmarks to assess outcome measures' utility. Commonly accepted reliability thresholds for quality reporting are 0.7–0.9.[11,13,14] Investigations of the American College of Surgeons National Surgical Quality Improvement Project[13] and Nationwide Inpatient Sample[12] have shown that few hospitals have adequate caseloads to meet reliability benchmarks, especially in quality reporting platforms that measure hospital performance using samples of patients. Rather than applying benchmarks to the data, our aim was to assess the reliability levels at which an outcome lost usefulness. Our study demonstrates that commonly accepted reliability benchmarks may be too high and that surgical outcomes with reliability below those levels still have significant predictive ability. Our findings imply that sampled outcome measures may have significant value for policy efforts aimed at steering patients to higher quality hospitals.

No prior study has sought to empirically evaluate the relationship between reliability and the ability to predict future surgical performance. Most work to date has assessed outcome reliability in the context of quality improvement initiatives, where the goal is to accurately identify outlying performers.[10,11,21] The reliability levels required for an outcome to be useful may be different for the purposes of

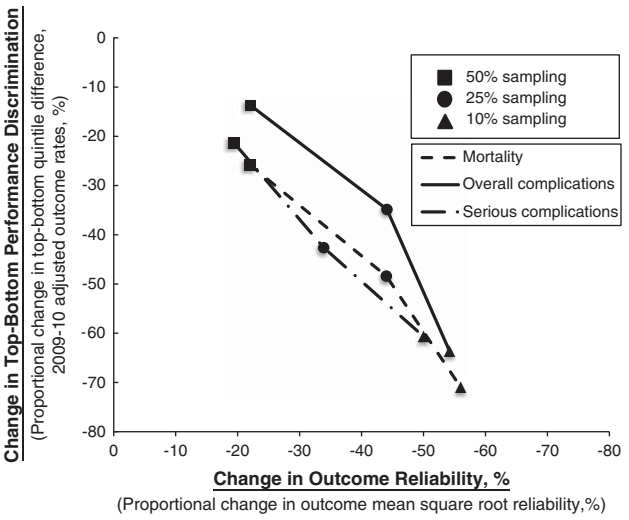

FIGURE 2. Proportional declines in outcome mean square root reliability (*x*-axis) and top-bottom quintile discrimination in adjusted 2009–2010 outcome rates (*y*-axis) at different levels of caseload sampling. Proportional changes depicted to account for different starting reliability levels for each outcome. Each point represents the change in outcome mean square root reliability and quintile discrimination at a given caseload sample size relative to 100% sampling. The connecting lines represent the particular outcomes from which the points were derived. Reoperation not shown due to starting reliability level < 0.1.

steering patients towards best performing hospitals. Using cross-sectional data, others have highlighted the ability of past hospital performance to predict future performance across different procedures.[22,26] Our work goes further by assessing the reliability limits required for future performance forecasting. We showed that even with highly sampled data, a patient treated in a worst versus best hospital based on past performance had 12% increased odds of death, 26% increased odds of serious complications, and 16% increased odds of experiencing any complication. With increasing reliability levels, the difference in odds became even more pronounced.

With lower caseloads and reliability levels, the ability to differentiate performance diminished. With extreme sampling (10%, less than that utilized by many surgical clinical registries), there was minimal difference between the top and bottom quintiles' outcome rates (eg, 0.9% mortality difference) despite statistically significant differences. This highlights the importance of considering the overall prevalence of an outcome. A 25% sampling strategy would be expected to produce 45% decreased discrimination between top-performing and bottom-performing hospitals compared with 100% sampling. For prevalent outcomes such as overall complications, this decreased discrimination may be acceptable for the purposes of identifying best and worst performance. For rare outcomes such as reoperation, 25% sampling would result in an essentially meaningless, although statistically significant, discrimination between top-performing and bottom-performing hospitals.

There are important limitations to this study. First, because we examined Medicare recipients undergoing colectomy, our results may not apply to a broader patient population or different procedure group. A broader assessment of volume thresholds for risk-adjusted outcomes to predict future hospital performance across procedure types would help inform the discussion of which measures to use for selective referral initiatives. Second, specific or rare outcomes (eg, reoperation in the present study or anastomotic leak after bariatric surgery) would be expected to have lower overall reliability levels, and higher caseloads will likely be necessary for initiatives using targeted complications for performance forecasting. Third, although we attempted to adjust for all identifiable comorbidities and demographic differences between patients, there are undoubtedly unmeasured confounders contributing to the variation we observed in future hospital performance. However, our risk-adjustment methodology and results are consistent with others examining the influence of past performance on future performance.[22,26] Moreover, our focus was on assessing the limitations of sample size and risk-adjustment methodology to predict future performance. For the purposes of performance forecasting, there are levels of reliability below which analytic methods cannot ensure accurate performance prediction.

This is the first empiric assessment of outcome reliability for future performance prediction in surgery. We chose to evaluate outcomes in the context of selective referral because it has a gold standard for assessing performance measures' utility: their ability to predict future hospital performance. This is arguably the most important criterion for assessing the usefulness of an outcome measure because patients and payers choose hospitals for referral based on historical data. We have demonstrated that even at very low reliability levels, outcome measures still have usefulness in this context and that common outcome reliability benchmarks may need reevaluation for this purpose.

REFERENCES

1. Blue Cross Blue Shield Association. Blue Distinction Centers for Complex and Rare Cancers Program Selection Criteria for 2008/2009 Designations. 2009. Available at: http://www.bcbs.com/why-bcbs/blue-distinction/blue-distinction-complex-and-rare/CRC_MidLevel-Criteria_101309.pdf. Accessed July 25, 2013.
2. Cevasco M, Ashley SW. Quality measurement and improvement in general surgery. *Perm J*. 2011;15:48–53.
3. Leapfrog Group. The Leapfrog Group Fact Sheet. 2013. Available at: http://www.leapfroggroup.org/about_leapfrog/leapfrog-factsheet. Accessed July 25, 2013.
4. Pratt GM, McLees B, Pories WJ. The ASBS Bariatric Surgery Centers of Excellence program: a blueprint for quality improvement. *Surg Obes Relat Dis*. 2006;2:497–503.
5. Dimick JB, Nicholas LH, Ryan AM, et al. Bariatric surgery complications before vs after implementation of a national policy restricting coverage to centers of excellence. *JAMA*. 2013;309:792–799.
6. Dimick JB, Osborne NH, Nicholas L, et al. Identifying high-quality bariatric surgery centers: hospital volume or risk-adjusted outcomes? *J Am Coll Surg*. 2009;209:702–706.
7. Livingston EH. Bariatric surgery outcomes at designated centers of excellence vs nondesignated programs. *Arch Surg*. 2009;144:319–325.
8. Livingston EH. Bariatric surgery centers of excellence do not improve outcomes. *Arch Surg*. 2010;145:605–606.
9. Massarweh NN, Flum DR, Symons RG, et al. A critical evaluation of the impact of Leapfrog's evidence-based hospital referral. *J Am Coll Surg*. 2011;212:150–159.

10. Adams JL. *The Reliability of Provider Profiling: A Tutorial*. Santa Monica, CA: RAND Corporation; 2009. Available at: http://www.rand.org/pubs/technical_reports/TR653.html. Accessed July 25, 2013.
11. Adams JL, Mehrotra A, Thomas JW, et al. Physician cost profiling—reliability and risk of misclassification. *New Engl J Med*. 2010;362:1014–1021.
12. 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.
13. 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.
14. Scholle SH, Roski J, Adams JL, et al. Benchmarking physician performance: reliability of individual and composite measures. *Am J Manag Care*. 2008;14:833–838.
15. Iezzoni LI, Daley J, Heeren T, et al. Identifying complications of care using administrative data. *Med Care*. 1994;32:700–715.
16. Weingart SN, Iezzoni LI, Davis RB, et al. Use of administrative data to find substandard care: validation of the complications screening program. *Med Care*. 2000;38:796–806.
17. Livingston EH. Procedure incidence and in-hospital complication rates of bariatric surgery in the United States. *Am J Surg*. 2004;188:105–110.
18. Elixhauser A, Steiner C, Harris DR, et al. Comorbidity measures for use with administrative data. *Med Care*. 1998;36:8–27.
19. Southern DA, Quan H, Ghali WA. Comparison of the Elixhauser and Charlson/Deyo methods of comorbidity measurement in administrative data. *Med Care*. 2004;42:355–360.
20. Birkmeyer NJ, Dimick JB, Share D, et al. Hospital complication rates with bariatric surgery in Michigan. *JAMA*. 2010;304:435–442.
21. Cohen ME, Ko CY, Bilimoria KY, et al. Optimizing ACS NSQIP modeling for evaluation of surgical quality and risk: patient risk adjustment, procedure mix adjustment, shrinkage adjustment, and surgical focus. *J Am Coll Surg*. 2013;217:336–346.
22. Dimick JB, Staiger DO, Birkmeyer JD. Ranking hospitals on surgical mortality: the importance of reliability adjustment. *Health Serv Res*. 2010;45:1614–1629.
23. Krell RW, Hozain A, Kao LS, et al. Reliability of risk-adjusted outcomes for profiling hospital surgical quality. *JAMA Surg*. 2014 doi: 10.1001/jamasurg.2013.4249. [Epub ahead of print].
24. Allen MJ, Yen WM. *Reliability. Introduction to Measurement Theory*. Long Grove, IL: Waveland Press; 2002:72–94.
25. Novick MR. The axioms and principal results of classical test theory. *J Math Psychol*. 1966;3:1–18.
26. Glance LG, Dick AW, Mukamel DB, et al. How well do hospital mortality rates reported in the New York State CABG report card predict subsequent hospital performance? *Med Care*. 2010;48:466–471.

