Published in final edited form as:

J Am Coll Surg. 2010 April ; 210(4): 503–508. doi:10.1016/j.jamcollsurg.2010.01.018.

Risk-Adjustment for Comparing Hospital Quality with Surgery: How Many Variables Are Needed?

Justin B Dimick, MD, MPH, FACS, Nicholas H Osborne, MD, MS, Bruce L Hall, MD, PhD, FACS*, Clifford Y Ko, MD, MS, FACS†, and John D Birkmeyer, MD, FACS

Michigan Surgical Collaborative for Outcomes Research and Evaluation (M-SCORE), Department of Surgery, University of Michigan, Ann Arbor, MI

* Department of Surgery, Washington University in Saint Louis

† American College of Surgeons Division of Research and Optimal Patient Care, Chicago, IL

Abstract

Background—The American College of Surgeon's National Surgical Quality Improvement Program (ACS NSQIP) will soon be reporting procedure-specific outcomes, and hopes to reduce the burden of data collection by collecting fewer variables. We sought to determine whether these changes threaten the robustness of the risk-adjustment of hospital quality comparisons.

Study Design—We used prospective, clinical data from the ACS National Surgical Quality Improvement Program (NSQIP) from 2005–07 (n= 184 hospitals, n=74,887 patients). For the five general surgery operations in the procedure-specific NSQIP, we compared the ability of the full model (21 variables), an intermediate model (12 variables) and a limited model (5 variables) to predict patient outcomes and to risk-adjust hospital outcomes.

Results—The intermediate and limited models were comparable to the full model in all analyses. In the assessment of patient risk, the limited and full models had very similar discrimination at the patient-level (C-indices for all 5 procedures combined of 0.93 vs. 0.91 for mortality and 0.78 vs. 0.76 for morbidity) and showed good calibration across strata of patient risk. In assessing hospital-specific outcomes, results from the limited and full risk models were highly correlated for both mortality (ranged 0.94 to 0.99 across the 5 operations) and morbidity (ranged from 0.96 to 0.99).

Conclusions—Procedure-specific hospital quality measures can be adequately risk-adjusted with a limited number of variables. In the context of the ACS NSQIP, moving to a more limited model will dramatically reduce the burden of data collection for participating hospitals.

Introduction

The National Surgical Quality Improvement Program (NSQIP) is on the verge of significant transformation (1). Among the most important changes is the evolution from sampling all

Correspondence address: Justin B. Dimick, MD, MPH, M-SCORE offices, 211 N. Fourth Avenue, Suite 301, Ann Arbor, MI 48104, Phone: (734) 998-7470, Fax: (734) 99807473, jdimick@umich.edu.

Disclosure Information: Nothing to disclose.

The American College of Surgeons National Surgical Quality Improvement Program and the hospitals participating in the ACS NSQIP are the source of the data used herein; they have not verified and are not responsible for the statistical validity of the data analysis or the conclusions derived by the authors.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

procedures in a specialty to 100% sampling of a small number of targeted procedures. The goal of this change is to better engage surgeons and accelerate quality improvement by focusing on procedure-level processes of care and outcomes (1).

Collecting data on a finite number of procedures may also have benefits with regards to the number of variables needed for adequate risk-adjustment. In the present version of NSQIP, which collects data on a random mix of procedures, a large number of variables are needed to adequately adjust for differences in patient risk (2). In contrast, when focusing on a single procedure, there is much less heterogeneity, and it will likely be possible to provide adequate risk adjustment with fewer variables. Empiric data from other clinical populations, such as cardiac surgery, reveal that most risk prediction comes from a relatively small number of the most important variables (3,4).

As the ACS-NSQIP moves towards procedure-specific sampling, a reduction in the number of covariates used for risk-adjustment would decrease the burden of data collection and lower the costs to participating hospitals. In this context, we sought to determine whether reducing the number of covariates in the procedure-targeted NSQIP would impact patient risk prediction or compromise hospital quality comparisons. We focused our analysis on the 5 core general surgery operations from the procedure-targeted NSQIP (1).

Methods

Data Source and Study Population

We used data from the 2005–2007 American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP). The ACS-NSQIP is a prospective, multi-institutional, clinical registry created to feedback risk-adjusted outcomes to hospitals for quality improvement purposes and now includes 241 participating centers. Over 130 patient and operative 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 reliability of the data is ensured through intensive training mechanisms for the surgical clinical nurse reviewers and by conducting inter-rater reliability audits of participating sites (5).

Using the appropriate Current Procedure Terminology (CPT) codes, we identified all patients undergoing one of 5 general surgery procedures selected for inclusion in the procedure-targeted version of ACS-NSQIP: colectomy, ventral hernia repair, bariatric surgery, cholecystectomy, and pancreatectomy (6). These operations were chosen for the general surgery core procedures because they contribute substantially to the overall morbidity and mortality (1).

Creating Risk-Adjustment Models

Full Model—We created our full risk-adjustment model using standard ACS NSQIP techniques (7,8). We entered all patient-level variables into a stepwise regression model, where all variables with a $P < 0.1$ are retained (up to 70 variables). Separate analyses were conducted for mortality (death within 30 days) and morbidity (one or more complications). Because these are both binary outcomes, all models were performed using logistic regression. In the present analysis, this full model included 21 variables.

Intermediate and Limited Models—Our aim was to identify a reduced set of core variables that could be collected on all patients and provide robust risk-adjustment for all 5 core general surgery operations. We created a separate logistic regression model for each of these 5

operations. To create these models, we first ran stepwise logistic regression models with all potential risk-adjustment variables included ($P < 0.10$ for entry and exit into the model). The output of these stepwise models includes a rank ordering of their importance (i.e., order of entry into the model), which reflects how strongly they are associated with the outcome variable.

We then created more parsimonious models that included smaller subsets of variables. We created an intermediate variable set by combining the 5 most important variables, in terms of order of entry in the stepwise regression model, for each procedure. Due to substantial overlap in the variables included for each procedure, there were only 12 unique variables for mortality and 11 unique variables for morbidity (Table 1). To determine the impact of even more limited risk-adjustment models on risk-adjustment, we created a limited model using the two most important variables from each procedure-specific model, which included only 5 variables (ASA class, functional status, congestive heart failure, dialysis, and bleeding disorder).

Statistical Analysis

The main purpose of this analysis was to establish whether intermediate and limited models adequately adjust for differences in patient severity across hospitals. We evaluated the models at both the patient- and hospital-level by comparing them to the full model.

To evaluate patient-level risk-adjustment, we first assessed the discrimination of the model by calculating the C-index for the limited, intermediate, and full model for each of the five operations, as well as all operations combined. The C-index, or area under the Receiver Operating Characteristic (ROC) curve, reflects how well the model discriminates between those who live and those who die. The C-index ranges from 0.5 (no ability to discriminate, i.e., a coin-flip) to 1.0 (perfect discrimination). We assessed calibration using the Hosmer Lemeshow statistic, which compares observed and predicted outcomes across deciles of increasing risk. Finally, we assessed the Spearman correlation coefficient of patient's expected mortality rates (predicted probabilities from the regression model) from the limited, intermediate, and full models.

To evaluate hospital-level risk-adjustment, we compared NSQIP outcome measures created using the limited, intermediate, and full models. Consistent with NSQIP methods, we calculated the ratio of observed to expected outcomes ("O/E ratio") at each hospital. The O/E ratio is calculated using logistic regression to predict a probability of the outcome (i.e., expected outcome) for each patient using standard post-estimation techniques. These probabilities are then summed for every hospital. The observed number of events is then divided by the expected number, which yields a risk-adjusted estimate of the outcome of interest; an O/E ratio of 1.0 is "as expected" given that hospital's patient severity, less than 1.0 is better than expected and greater than 1.0 is worse than expected. We calculated O/E ratios for each outcome (mortality and morbidity) using the limited, intermediate, and full models to estimate the expected mortality.

We then assessed how well the O/E ratios created with the limited and intermediate models agreed with the O/E ratios created using the full model. To assess these correlations, we compared the O/E ratios from each model using Spearman correlation coefficients. In these analyses, a correlation coefficient of 1.0 implies perfect agreement in O/E ratios.

Results

The most important variables in the risk-adjustment models overlapped significantly for the 5 general surgery procedures (Table 1). Our intermediate variable set, which included the top 5 variables for each operation, summed to only 12 variables for mortality and 11 variables for

morbidity (Table 1). Our limited variables set, which included only the two most important variables from each procedure, included only 4 variables for mortality (ASA class, functional status, dialysis, and congestive heart failure) and only 3 variables for morbidity (ASA class, functional status, and bleeding disorders) (Table 1). There was also extensive overlap in the most important variables for the two different outcomes, morbidity and mortality. With the intermediate model, 8 of 12 mortality variables were also in the morbidity model, leaving only 15 unique variables. With the limited model, two of the variables overlapped, leaving only 5 unique variables.

The intermediate and limited models provided similar patient-level discrimination when compared to the full risk-adjustment model (Table 2). For both morbidity and mortality, patient-level discrimination based on the C-index was similar for the limited, intermediate, and full models for all 5 procedures (Table 2). For example, with colectomy the C-indices for morbidity ranged from 0.71 (full model) to 0.68 (limited model) and the C-indices for mortality ranged from 0.92 (full model) to 0.89 (limited model). Patient level predicted-probabilities of morbidity and mortality from the limited and intermediate models were also highly correlated for all 5 procedures (Table 2). Calibration of the limited and full models was also similar with very little differences in the Hosmer Lemeshow statistics (data not shown). In the assessment of a composite outcome measure of all 5 procedures combined, the limited, intermediate, and full models also had very similar discrimination for both morbidity and mortality (Table 2).

When comparing hospital-level quality using O/E ratios, rankings created with the limited, intermediate, and full risk models were highly correlated for all 5 procedures (Table 3). For all 5 of the general surgery procedures, the correlations at the hospital-level were even higher than the correlations of predicted probabilities at the patient-level. For morbidity O/E ratios, the hospital-level correlations between the limited and full models ranged from 0.96 for colectomy to 0.99 for pancreatectomy (Table 3). For mortality O/E ratios, hospital-level correlations for the limited and full models ranged from 0.94 for colectomy to 0.99 for pancreatectomy (Table 3). In the assessment of all 5 procedures combined, the limited and intermediate models were highly correlated with the full models for both morbidity and mortality (Table 3, Figures 1 and 2).

Discussion

When using outcome measures to profile hospital quality, it is important to ensure that differences are fully adjusted for patient risk (9). The National Surgical Quality Improvement Program (NSQIP) collects a comprehensive set of detailed clinical data for this purpose (7,8). However, the present study provides empiric data demonstrating that a much smaller set of variables can be used without compromising risk-adjustment. Hospital O/E ratios based on a limited variable set were almost identical to the O/E ratios created using the comprehensive set of variables. This finding held true for both outcomes (morbidity and mortality) and was consistent across all 5 general surgery procedures included in the new procedure-specific NSQIP.

Although this study is the first to describe the adequacy of a limited model for general surgery, there are previous studies in other surgical populations. In a large cohort of patients having coronary artery bypass surgery in Ontario, Tu and colleagues compared a 6 variable risk-adjustment model to a more comprehensive 12 variable model (4). With the additional 6 variables, the C-index increased minimally (0.77 to 0.79) and there were no clinically significant changes in hospital risk-adjusted mortality (4). The authors concluded that risk-adjustment models could be simplified to make data collection more efficient. Our findings are consistent with this work, and extend the findings to general surgery.

There are two possible reasons why the limited and full models provide equivalent risk-adjustment. First, there may be a finite number of important risk domains that are adequately captured in the limited variable set. The additional variables in the “full” model may be redundant, adding little to the predictive ability of the model. Such redundant variables may represent an important risk domain, but the domain may be captured “better” by one of the variables already in the limited set. The most important variables in all models include functional status and ASA score. These variables have strong face validity as they represent important domains of risk: Frailty (functional status) and severity of comorbid disease (ASA score). These two variables also may be the strongest predictors of outcome because they are multidimensional constructs that actually represent many risk-domains. For example, impaired functional status may simply be the downstream consequences of having severe cardiac and pulmonary disease.

Another potential reason the limited and full risk models are equivalent is that some variables may not vary across hospitals. To confound hospital quality comparisons, a variable must satisfy two criteria. First, the variable must be associated with patient risk. A variable that does not relate to the outcome of interest cannot confound hospital comparisons. Second, the variable must vary across hospitals. A variable that is present in the same fraction of patients at every hospital—even if strongly associated with patient risk—cannot act as a confounding variable. For example, consider a variable that is strongly associated with mortality, such as preoperative sepsis. If one hospital has 1% of patients with preoperative sepsis and another has 20% this variable would need to be included in risk adjustment models. However, if all hospitals have 1% of patients with preoperative sepsis, this variable cannot confound comparisons and does not need to be included in risk-adjustment models. There are some empiric data supporting the idea that patients undergoing the same surgical procedure are relatively homogeneous. For example, we have previously shown that patient severity, as measured by the expected mortality rate, varies very little across hospitals performing cardiac surgery in both Pennsylvania and New York (3).

The present study does have certain limitations. Although the NSQIP is rich in clinical detail, the present iteration does not collect some procedure-specific variables that could be important for optimal risk-adjustment (e.g., diverticulitis vs. colon cancer for colon resection). In the next iteration of NSQIP, there will be procedure-specific variables for each of the general surgery operations. These were selected through collaboration with surgeon experts to identify the most important variables for each procedure. The addition of these variables (approximately 3–5 per procedure) will likely dramatically improve the performance of risk-adjustment models. A second limitation of this study is a potential lack of generalizability. At present, the NSQIP disproportionately represents larger, teaching hospitals. As it expands to other hospitals, the variables needed in the “core” set may not be the same. Both of these limitations highlight the necessity of ongoing research to ensure the accuracy of more limited models over time.

This study highlights the need to consider the potential trade-off between the accuracy of risk-adjustment and the efficiency of data collection. There is no doubt collecting all possible patient characteristics would provide the most accurate data for risk-adjustment. Unfortunately, extracting these variables from the medical record is time consuming and expensive. Rather than treating this as a trade-off, we sought to develop a strategy that aims to trim the waste (e.g., eliminate variables that do not add predictive value) without compromising hospital quality comparisons. This study demonstrates that most if not all predictive power is captured by the most important variables. Thus, we can improve the efficiency of data collection without making significant trade-offs in the accuracy of risk-adjustment.

This study has important implications for the ACS NSQIP and other quality measurement platforms. Our findings imply that collection of risk factors can be dramatically reduced.

Limiting collection of patient risk factors will decrease the work of those who collect data and reduce costs to hospitals. Since data collection is the single most expensive item for hospitals participating in NSQIP, this reduction in costs would make NSQIP more affordable. This reduction also creates an opportunity to make NSQIP more useful. With fewer risk factors to collect, hospitals can expand data collection to include other data elements that would help them improve quality. For example, the next iteration of NSQIP will also collect processes of care and outcomes specific to each operation. These will likely be more informative for quality improvement than the summary O/E ratios currently reported. Making NSQIP more affordable and more useful will help the program disseminate to a broader group of hospitals and help achieve the program's goal of becoming the default quality improvement program for all United States hospitals.

Acknowledgments

This study was supported by a career development award to Dr Dimick from the Agency for Healthcare Research and Quality (K08 HS017765).

References

1. Birkmeyer JD, Shahian DM, Dimick JB, et al. Blueprint for a new American College of Surgeons: National Surgical Quality Improvement Program. *J Am Coll Surg* 2008;207:777–782. [PubMed: 18954793]
2. Khuri SF, Daley J, Henderson W, et al. The Department of Veterans Affairs' NSQIP: the first national, validated, outcome-based, risk-adjusted, and peer-controlled program for the measurement and enhancement of the quality of surgical care. National VA Surgical Quality Improvement Program. *Ann Surg* 1998;228:491–507. [PubMed: 9790339]
3. Dimick JB, Birkmeyer JD. Ranking hospitals on surgical quality: does risk-adjustment always matter? *J Am Coll Surg* 2008;207:347–351. [PubMed: 18722939]
4. Tu JV, Sykora K, Naylor CD. Assessing the outcomes of coronary artery bypass graft surgery: how many risk factors are enough? Steering Committee of the Cardiac Care Network of Ontario. *J Am Coll Cardiol* 1997;30:1317–1323. [PubMed: 9350934]
5. ACS NSQIP. Data Collection Overview. Available from: https://acsnsqip.org/main/program_data_collection.asp
6. CPT 2006. Current Procedural Terminology. Chicago: American Medical Association; 2005.
7. Daley J, Khuri SF, Henderson W, et al. Risk adjustment of the postoperative morbidity rate for the comparative assessment of the quality of surgical care: results of the National Veterans Affairs Surgical Risk Study. *J Am Coll Surg* 1997;185:328–340. [PubMed: 9328381]
8. Khuri SF, Daley J, Henderson W, et al. Risk adjustment of the postoperative mortality rate for the comparative assessment of the quality of surgical care: results of the National Veterans Affairs Surgical Risk Study. *J Am Coll Surg* 1997;185:315–327. [PubMed: 9328380]
9. Iezzoni LI. The risks of risk adjustment. *JAMA* 1997;278:1600–1607. [PubMed: 9370507]

Figure 1.

Comparison of hospital ratio of observed to expected outcomes (O/E) ratios created using the intermediate versus full model for (A) mortality and (B) morbidity for all 5 procedures combined.

Figure 2.

Comparison of hospital observed to expected outcomes (O/E) ratios created using the limited versus full model for (A) mortality and (B) morbidity for all 5 procedures combined.

Table 1
Importance of Risk-Adjustment Variables in the Stepwise Logistic Regression Models

Variable	Order of importance				
	1	2	3	4	5
Mortality models					
Cholecystectomy	Functional status	ASA class	Ascites	Weight loss	Albumin
Ventral hernia repair	Functional status	Dialysis	ASA class	Wound class	Dyspnea
Gastric bypass	Functional status	Congestive heart failure	Hypertension	Gender	Dyspnea
Pancreatotomy	Functional status	Congestive heart failure	ASA class	Dyspnea	Albumin
Colectomy	ASA class	Functional status	Emergency	Albumin	Dyspnea
Morbidity models					
Cholecystectomy	ASA class	Functional status	Wound class	Ascites	Albumin
Ventral hernia repair	ASA class	Functional status	Wound class	Dyspnea	Body mass index
Gastric bypass	Functional status	Bleeding disorder	ASA class	Diabetes	Congestive heart failure
Pancreatotomy	Functional status	ASA class	Bleeding disorder	Dyspnea	Body mass index
Colectomy	ASA class	Functional status	Emergency	Albumin	Body mass index

Table 2
Ability of the Different Risk-Adjustment Models to Discriminate Patient-Level Risk of Morbidity and Mortality

Procedure	Model	Morbidity		Mortality	
		C-Index	Correlation of predicted probabilities	C-Index	Correlation of predicted probabilities
Cholecystectomy	Full	0.77	1.0	0.93	1.0
	Intermediate	0.75	0.92	0.93	0.98
	Limited	0.73	0.86	0.92	0.96
Ventral hernia repair	Full	0.70	1.0	0.91	1.0
	Intermediate	0.70	0.85	0.91	0.90
	Limited	0.66	0.82	0.88	0.94
Gastric bypass	Full	0.62	1.0	0.84	1.0
	Intermediate	0.62	0.82	0.83	0.88
	Limited	0.60	0.82	0.79	0.61
Pancreatectomy	Full	0.62	1.0	0.77	1.0
	Intermediate	0.61	0.80	0.75	0.85
	Limited	0.59	0.82	0.73	0.82
Colectomy	Full	0.71	1.0	0.92	1.0
	Intermediate	0.70	0.93	0.90	0.95
	Limited	0.68	0.82	0.89	0.90
All Procedures	Full	0.78	1.0	0.93	1.0
	Intermediate	0.77	0.98	0.92	0.96
	Limited	0.76	0.95	0.91	0.93

Table 3

Comparison of O/E Ratios Created Using the Different Risk-Adjustment Models

		Morbidity	Mortality
Procedure	Model	Correlation of O/E ratios	Correlation of O/E ratios
Cholecystectomy	Full	1.0	1.0
	Intermediate	0.98	0.99
	Limited	0.96	0.97
	None	0.89	0.90
Ventral hernia repair	Full	1.0	1.0
	Intermediate	0.98	0.99
	Limited	0.98	0.99
	None	0.95	0.99
Gastric bypass	Full	1.0	1.0
	Intermediate	0.99	1.0
	Limited	0.99	0.99
	None	0.98	0.99
Pancreatectomy	Full	1.0	1.0
	Intermediate	0.99	0.99
	Limited	0.99	0.99
	None	0.97	0.98
Colectomy	Full	1.0	1.0
	Intermediate	0.99	0.97
	Limited	0.96	0.94
	None	0.89	0.75
All procedures	Full	1.0	1.0
	Intermediate	0.99	0.97
	Limited	0.97	0.93
	None	0.93	0.74